From Seed Discovery to Deep Reconstruction: Predicting Saliency in Crowd via Deep Networks

Yanhao Zhang[1],Lei Qin[2],Qingming Huang[1,2],Kuiyuan Yang[3],Jun Zhang[4],Hongxun Yao[1]
[1]Harbin Institute of Technology, Harbin, China
[2]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, China
[3]Microsoft Research Asia, Beijing, China, [4]Hefei University of Technology, Hefei, China
{yhzhang,h.yao}@hit.edu.cn,kuyang@microsoft.com,{lqin, qmhuang}@jdl.ac.cn

ABSTRACT

Although saliency prediction in crowd has been recently recognized as an essential task for video analysis, it is not comprehensively explored yet. The challenges lie in that eye fixations in crowded scenes are inherently "distinct" and "multi-modal", which differs from those in regular scenes. To this end, the existing saliency prediction schemes typically rely on hand designed features with shallow learning paradigm, which neglect the underlying characteristics of crowded scenes. In this paper, we propose a saliency prediction model dedicated for crowd videos with two novelties: 1) Distinct units are discovered using deep representation learned by a Stacked Denoising Auto-Encoder (SDAE), considering perceptual properties of crowd saliency; 2) Contrast-based saliency is measured through deep reconstruction errors in the second SDAE trained on all units excluding distinct units. A unified model is integrated for online processing crowd saliency. Extensive evaluations on two crowd video benchmark datasets demonstrate that our approach can effectively explore crowd saliency mechanism in two-stage SDAEs and achieve significantly better results than state-of-the-art methods, with robustness to parameters.

Keywords

Crowd Saliency; Deep AutoEncoders; Reconstruction Errors

1. INTRODUCTION

Crowd saliency essentially targets at identifying the potential regions in crowded scenes that most attract visual attention. Crowded scenes are prevalent in surveillance videos [20, 27, 28, 29], which are however more risky especially in salient regions [18, 3]. It is therefore emerging to design crowd saliency model for video surveillance.

The straightforward solutions directly apply conventional saliency approaches, which focus on detecting saliency regions in regular scenes. In principle, it is assumed that saliency regions can stand out from their neighbors with high contrast, which serves as the basis for various saliency detection models proposed in the image and video domain [8, 5, 1, 21, 2]. Recent advances in saliency detection mainly focus on detecting salient objects via bottom-up measurements, such as rareness [26, 19] and contrast variants [25, 11]. Besides, it is also shown that the context-dependent, top-down mechanisms incorporated with object/background prior are comparably accurate and effective [13]. Nevertheless, these approaches are inappropriate for saliency prediction in crowded scenes as they are with different underlying characteristics. As revealed by the psychological evidence on crowds [24], two important perceptual cues should be highlighted in crowd saliency: The "seed", i.e., an individual or small group attempting to engage with the crowd; The crowd engages with a seed to occur its flow motion. Hence, the salient crowd motion typically appears from the crowd seed where the crowd modifies the existing states to influent its neighbors. This inspires us to model the crowd seeds that are *distinct* to separate the common patterns from the crowds, as well as propagate the influence in *multi-modal* context. To this end, it is nontrivial to design handcraft features to differentiate the salient regions. Thus, we resort to learning such features via deep networks that encodes the hierarchical mechanisms underlying crowd saliency.

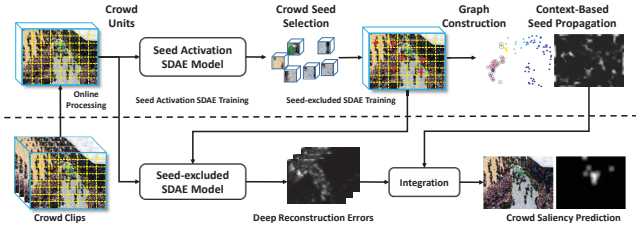


Figure 1: Overview of the framework for crowd saliency prediction based on deep networks.

Following this intuition, in this work, we propose a unified framework for crowd saliency prediction as shown in Figure 1. Firstly, we select crowd seeds from crowd units with a seed activation Stacked Denoising Auto-Encoder (SDAE). We further calculate deep reconstruction errors using a seed-excluded SDAE, which measure global contrast between the salient patterns and common patterns. Crowd seeds are subsequently leveraged to learn the propagation errors in a ranking function. The score of crowd saliency is finally refined by context-based seed propagation. The main contributions of this work are summarized as following three aspects: **(1)** We explore the characteristics of crowd units by training a seed activation SDAE, which introduces

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2967185

activation metrics of deep representation to discover crowd seeds. **(2)** We make use of the reconstruction error by training a seed-excluded SDAE, which effectively measures the contrast-based saliency in crowds. **(3)** A unified model is assembled to predict crowd saliency, and verified on two crowd video datasets with significant performance gains.

2. CROWD SALIENCY DEEP NETWORKS

In this section, we present the proposed model for crowd saliency prediction in details.

2.1 Stacked Denoising Auto-Encoders (SDAE)

SDAE essentially builds a deep architecture by stacking multiple layers of Denoising Auto-Encoders (DAE) [22]. DAE basically consists of two components, i.e., encoder and decoder [22]. Both components attempt to learn two mapping functions, termed as $\mathbf{h}(\mathbf{W}, \mathbf{b})$ and $\mathbf{g}(\hat{\mathbf{W}}, \hat{\mathbf{b}})$, where $\mathbf{W}$, $\mathbf{b}$ denote the weights and bias parameters of the encoder and $\hat{\mathbf{W}}$, $\hat{\mathbf{b}}$ correspond to the terms of decoder. Before encoding, $\tilde{\mathbf{x}}$ is a noise-corrupted version of the clean input data $\mathbf{x}$ by stochastic mapping $\tilde{\mathbf{x}} = \mathbf{D}(\tilde{\mathbf{x}}|\mathbf{x})$. For a corrupted input $\tilde{\mathbf{x}}_i$, the hidden layer representation $\mathbf{z}_i$ can be obtained through $\mathbf{z}_i = \mathbf{h}(\tilde{\mathbf{x}}_i|\mathbf{W}, \mathbf{b}) = s(\mathbf{W}\tilde{\mathbf{x}}_i + \mathbf{b})$, where $s(.)$ is the sigmoid activation function. The decoder tries to map the hidden representation $\mathbf{z}_i$ back to input $\mathbf{x}_i$ by computing $\hat{\mathbf{x}}_i = \mathbf{g}(\mathbf{z}_i|\hat{\mathbf{W}}, \hat{\mathbf{b}}) = s(\hat{\mathbf{W}}\mathbf{z}_i + \hat{\mathbf{b}})$. Given the training set $\mathcal{C} = \{\mathbf{x}_i\}_{i=1}^N$, the parameters $\mathbf{W}, \mathbf{b}, \hat{\mathbf{W}}, \hat{\mathbf{b}}$ can be optimized as a regularized least square optimization problem, i.e.,

$$\min_{\mathbf{W},\mathbf{b},\hat{\mathbf{W}},\hat{\mathbf{b}}} \sum_{i=1}^{N} ||\hat{\mathbf{x}}_i - \mathbf{x}_i||_2^2 + \gamma(||\mathbf{W}||_F^2 + ||\hat{\mathbf{W}}||_F^2), \quad (1)$$

where γ balances the reconstruction error and regularization. The non-linearity of activation function allows SDAE to learn complex mapping and capture the latent patterns that reflect the correlation shared among training data.

2.2 Two-stage SDAEs based Deep Structure

Our goal is to assign a saliency value $\text{Sal}(\mathbf{x}_i, \mathcal{C}) \in [0, 1]$ to each crowd unit $\mathbf{x}_i$ which best fits the given crowd unit set $\mathcal{C} = \{\mathbf{x}_i\}_{i=1}^N$ in a time window t. Let $\mathcal{C}_s$ denote the set containing distinct crowd units as "crowd seeds". $\mathcal{C}_r = \mathcal{C} \backslash \mathcal{C}_s$ denotes the rest. The selection of crowd seeds will separate the distinct units from the common units. The contrast-based saliency are measured by deep reconstruction error between a given crowd unit $\mathbf{x}_i$ and the common units $\mathcal{C}_r$.

A two-stage SDAE model is presented to achieve this goal as shown in Figure 1, which contains two phases, i.e.,

- To distinguish crowd seeds, we design deep representation based metrics using neuron activations of the first SDAE trained on all crowd units.
- To calculate the global contrast, we leverage deep reconstruction errors of the second SDAE trained on crowd units without seed units.

Both of the SDAEs are with the same structure, including one input layer and three hidden layers, as shown in Figure 2. Inspired by [23], an over-complete set of filters is used to better capture the image/video structure in the first hidden layer, followed by two equal-sized hidden layers. In the training phases, SDAEs are layer-wise trained by optimizing Eq. 1 using gradient descent, and then fine-tuned globally. Therefore, the contrast-based saliency can be indicated by the reconstruction cost of the seed-excluded SDAE.

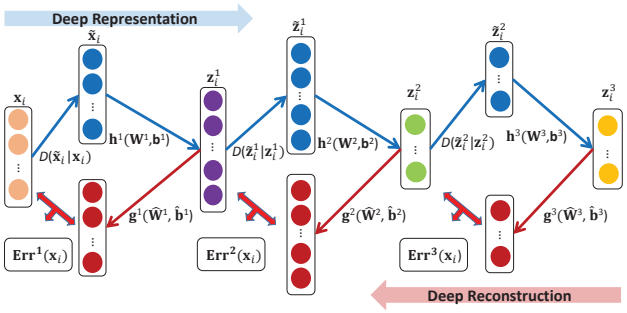


Figure 2: Schematic illustration of two-stage SDAE.

2.2.1 Seed Prediction via Deep Representation

The deep structured SDAE provides an effective means to characterize the crowd seed set $\mathcal{C}_s$, especially in virtue of its ability to learn latent patterns in an unsupervised manner.

Training Seed Activation SDAE: The first SDAE is trained on all crowd units $\mathcal{C}$ to select crowd seeds. The metrics of the output (hidden) representation should favor distinct crowd units as the seed set. In order to distinguish the crowd seeds $\mathcal{C}_s$, the units should be outstanding from the whole set $\mathcal{C}$ according to the activation value.

Crowd Firing Rate: In neural science, neuron's firing rate increases or decreases when certain patterns appear in receptive field [12]. The importance of input signal is reflected as how much energy the activity consumes. It is therefore reasonable to assume that the larger energy of crowd unit consumes, the more distinct it is. Accordingly, the crowd seed can be measured by energy consumption of all hidden units. Thus, we define Crowd Firing Rate (CFR) to describe the average activity of the j-th neuron in the last hidden layer:

$$\text{CFR}_j = \sum_{i=1}^{N} \mathcal{H}_j^3(\mathbf{x}_i)/N, \quad (2)$$

where $\mathcal{H}_j^3(\mathbf{x}_i)$ refers to the activation value of the j-th neuron of the last hidden layer $\mathbf{z}_i^3$ for the i-th input unit $\mathbf{x}_i$, N is the number of crowd units.

Seed Activation Rate: Next, we define Seed Activation Rate (SAR) to calculate how much energy will be consumed for a given crowd unit:

$$\text{SAR}_i = \sum_{j=1}^{r} ||\mathcal{H}_j^3(\mathbf{x}_i) - \text{CFR}_j||_2^2, \quad (3)$$

For each unit $\mathbf{x}_i$, SAR_i denotes the accumulation of energy cost according to CFR of all the neurons. We sort the SAR of all the units in a descending order and select the top n as crowd seeds $\mathcal{C}_s$. Given $\mathcal{C}_s$, we make a binary indicator as $\delta(\mathbf{x}_i) = 1$, if $\mathbf{x}_i$ is $s \in \mathcal{C}_s$, otherwise $\delta(\mathbf{x}_i) = 0$.

2.2.2 Contrast Inference via Deep Reconstruction

By hierarchically capturing latent patterns, SDAE helps to measure the contrast between the salient and the common units through deep reconstruction. We model the common crowd units with SDAE and indicate the contrast-based saliency of $\mathbf{x}_i$ with the deep reconstruction error.

Training Seed-excluded SDAE: We establish the second SDAE trained on $C_r = C \backslash C_s$ to model common crowd units. The principle is to measure the contrast via the differences between the input and the reconstructed patterns by SDAE. The deep reconstruction error is the cost that converts the input to hidden representations of common patterns through non-linear transformations. As in Figure 2, seed-excluded

Methods	*AUC*	*SE*	*NSS*	*SE*	*CC*	*SE*
AWS [4]	0.790	0.0010	1.019	0.0057	0.343	0.0017
IS [8]	0.790	0.0010	0.999	0.0056	0.339	0.0017
SR [9]	0.788	0.0010	1.003	0.0058	0.340	0.0018
FT [1]	0.784	0.0009	1.018	0.0054	0.344	0.0016
GBVS [7]	0.791	0.0009	1.005	0.0051	0.341	0.0015
HBS [15]	0.794	0.0009	1.029	0.0055	0.346	0.0016
OB [15]	0.796	0.0009	1.039	0.0049	0.356	0.0014
Ours-NS	0.797	0.0011	1.042	0.0050	0.357	0.0017
Ours-NR	0.807	0.0011	1.052	0.0050	0.360	0.0017
Ours	**0.815**	0.0010	**1.064**	0.0048	**0.368**	0.0017

Table 1: Comparisons with image saliency methods on ASCMN-CROWD dataset.

Methods	*AUC*	*SE*	*NSS*	*SE*	*CC*	*SE*
RARE [19]	0.703	0.0060	1.253	0.0672	0.216	0.0098
SUN [26]	0.653	0.0089	1.160	0.0654	0.196	0.0109
SDSR [21]	0.673	0.0070	1.013	0.0734	0.166	0.0110
PQFT [6]	0.688	0.0065	1.234	0.0645	0.181	0.0089
CE [14]	0.638	0.0088	0.508	0.1228	0.125	0.0068
DC [30]	0.680	0.0025	0.721	0.0741	0.179	0.0035
SP [17]	0.739	0.0052	0.945	0.1787	0.238	0.0099
Ours-NS	0.772	0.0067	1.375	0.0890	0.271	0.0039
Ours-NR	0.782	0.0065	1.390	0.1120	0.302	0.0050
Ours	**0.785**	0.0022	**1.505**	0.1190	**0.306**	0.0037

Table 2: Comparisons with video saliency methods on ASCMN-CROWD dataset.

SDAE is comprised of 3 encoders and 3 decoders.

Deep Reconstruction Error: We calculate deep reconstruction error in Eq. 4 for $\mathbf{x}_i \in \mathcal{C}$ by seed-excluded SDAE, which globally reflects the contrast with the common crowd units. The reason is that common patterns are fitted better in the deeper layer, which is further verified in the experiment.

$$\mathrm{Err}^d(\mathbf{x}_i) = ||\mathbf{g}^3(\mathbf{h}^3(\mathbf{D}(\tilde{\mathbf{z}}_i^2|\mathbf{z}_i^2))) - \mathbf{z}_i^2||_2^2, \quad (4)$$

The errors $\mathrm{Err}^d(\mathbf{x}_i)$ captures the difference between the unit $\mathbf{x}_i$ and common patterns, which generally leads to promising results in contrast-based saliency detection.

Context-based Seed Propagation: Considering that crowd seeds strongly affect the neighborhood in crowd flow, we intend to purify the salient foreground by the propagation of the seeds. Therefore, crowd seeds are set as queries and processed sequentially by propagating the relevances in a ranking manifold as [16]. We establish the graph $G=(V,E)$ with all units encoded in the affinity, where V represents crowd units, E denotes the edge weighted by $\mathbf{W}=[W_{ij}]_{N\times N}$ with W_{ij} is the difference between crowd units. The degree matrix is $\mathbf{S} = \mathrm{diag}(\mathbf{s}_{11}, \ldots, \mathbf{s}_{NN})$, where $\mathbf{s}_{ii} = \Sigma_j \mathbf{W}_{ij}$. This ranking function assigns a ranking value f_i as to each unit $\mathbf{x}_i$, and $\mathbf{f}$ can be viewed as a vector $\mathbf{f} = [f_i|_{i=1}^N]$. Let $\mathbf{y}=[\delta(\mathbf{x}_i)|_{i=1}^N]$ denote a binary vector indicating crowd seed queries. The normalized Laplacian matrix is denoted as $\tilde{\mathbf{L}} = \mathbf{S}^{-1/2}\mathbf{W}\mathbf{S}^{-1/2}$. We get the close form solution as $\mathbf{f} = (\mathbf{I} - \alpha\tilde{\mathbf{L}})^{-1}\mathbf{y}$, where $\mathbf{I}$ is an identity matrix, $\alpha = 1/(1+\mu)$. For all queries indicated by $\mathbf{y}$, the rank score f_i gives the propagation errors of the seeds.

2.3 Online Processing Crowd Saliency

When crowd unit set $\mathcal{C}$ arrives, we first train the two SDAEs based on $\mathcal{C}$ and $\mathcal{C}_r$. We make a simple forward pass on $\mathbf{x}_i \in \mathcal{C}$ through seed-excluded SDAE to get the deep reconstruction error. The saliency score $\mathrm{Sal}(\mathbf{x}_i, \mathcal{C})$ is then determined based on Eq. 5 by integrating the refined propagation error and center bias [8] as,

$$\mathrm{Sal}(\mathbf{x}_i, \mathcal{C}) = (\mathrm{Err}^d(\mathbf{x}_i) + f_i) \cdot \exp\left\{-||\mathrm{Ctr} - \mathrm{Pos}(\mathbf{x}_i)||^2\right\} \quad (5)$$

where $\mathrm{Pos}(\mathbf{x}_i)$ defines the coordinates of the unit, Ctr is the center of the frame. We assign the saliency score to corresponding $\mathrm{Pos}(\mathbf{x}_i)$ in the frame to generate the saliency map with Gaussian smoothing for robustness.

3. EXPERIMENTS

We evaluate our method on two dynamic video datasets, i.e., ASCMN-CROWD dataset [20] and CRCNS-CROWD dataset [10]. We present results of applying our model in yielding saliency maps for predicting eye fixations in crowds.

Parameter Settings. For input videos, each frame is resized to 120×160 and uniformly partitioned into non-overlapping spatio-temporal cubes of size $5\times5\times5$. We compute 3D gradient of the cube as raw motion signals and 2-dim coordinates of the cube center to describe each cube as a *crowd unit*. Two SDAEs are both fixed as 3 hidden layers with 250, 100 and 100 hidden nodes, respectively. For each time window t containing 5 frames, we select top $n = 20\%$ as $\mathcal{C}_s$ according to Eq. 3. We set $\gamma = 0.002$, $\alpha = 0.99$ and the mini-batch size to 10. The corruption size in DAE is set to 0.2. These parameters are empirically chosen and fixed through all the following experiments. Three metrics are used to quantify the performance of saliency models for predicting eye fixations [17], including AUC, NSS and CC with corresponding Standard Error (SE).

3.1 Evaluations on ASCMN-CROWD Dataset

In order to evaluate the performance on the eye fixation prediction in the crowded scenes, we conduct experiments on CROWD subset of ASCMN dataset [20] containing 14 videos in 3 categories that are crowd-specific: 5 videos for Abnormal Crowd (Abn-Crd), 4 for Surveillance Crowd (Sur-Crd), 5 for Crowd Activity (Act-Crd).

(1) Image saliency approaches. We compare our model with the widely compared methods and accomplish an objective comparison with original fixation data as groundtruth. Table 1 shows the AUC, NSS and CC with SE for different models. Note that the metrics of these compared models are obviously surpassed by our model, which is attributed to meaningful deep representation. Our model explores the high-level correlation of crowd units for modeling crowd saliency, which achieves better performance than others with improvements 0.02-0.08 in all metrics.

(2) Video saliency approaches. To further validate our model on capturing dynamic crowded scenes, we compare our approach to video saliency methods. In this case, the evaluation measures are designed to use fixation heat maps as the groundtruth. The average measures are reported in Table 2, which shows Ours is superior to other studied methods. There is a consistent improvement in the performance of all metrics. Overall, our approach outperforms the other models under all metrics, which indicates that semantics of crowd perception encoded in deep networks is beneficial in cluttered foreground.

Validation of Model Components. We further evaluate the effectiveness of our approach with respect to model components. From Table 1 and 2, we note that, 1) Without the seed prediction (Ours-NS), the AUC score slightly drops, illustrating the significance of distinct crowd seed. 2) Without reconstruction component (Ours-NR), also performs worse than ours, which indicates that the effectiveness of the reconstruction stage for the contrast measurement. Besides, crowd seeds are helpful to eliminate the distraction and enhance the salient regions in crowd context.

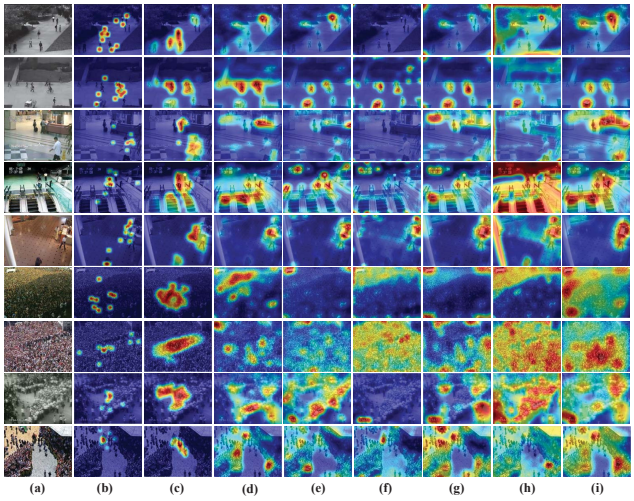
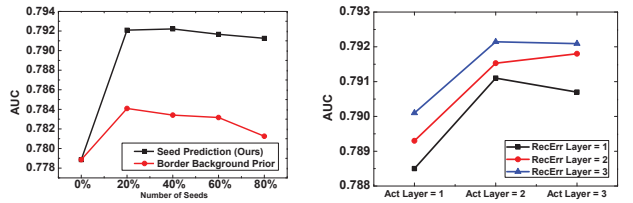


Figure 3: Visual comparisons of state-of-the-art saliency detection models. (a) Video frame, (b) Eye fixation, (c) Ours, (d) RARE, (e) PQFT, (f) SR, (g) AWS, (h) FT and (i) IS. Best viewed in color.



(a) (b)

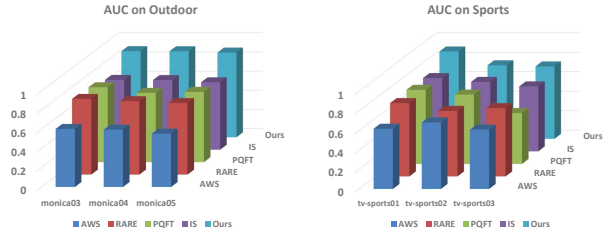
Figure 4: (a) The effect of the number of seeds. (b) The effect of layers in SDAEs.

3) Exploring contrast (Ours) by fusing seed prediction and reconstruction yields the best results, demonstrating both components are complimentary to each other. Ours-NS and Ours-NR still outperform all existing methods show the advancement of the deep representation. Qualitative comparisons of the selected image and video saliency methods are shown in Figure 3. It is obvious that most approaches look like to be influenced significantly when the crowd is cluttered and similar appearances occupy the scenes. On the contrary, ours tends to be less distracted by the common foreground as well as the cluttered background, which still predicts desirable fixations in such crowded scenes.

Robustness to Parameters. Figure 4(a) shows how the main metric AUC score for Act-Crd Category is affected by the number of seeds ranging from 0% to 80%. We also compare with the alternative method with the seed selected from the border set using border background prior [13]. As we can see, AUC starts to increase dramatically and gradually decrease when the seed number is larger. It is not very sensitive within a range of values. It is superior to border background prior, indicating common patterns are falsely detected in border background set while ours not. Keeping seed number as 20%, we tune the layers of the two SDAEs (Figure 4(b)). Note that AUC generally increase with deeper layers of reconstruction (RecErr Layer) when crowd seeds are predicted from different activation layers (Act Layer = 1, 2, 3). Deeper activation and reconstruction layers help more to identify salient regions.

Methods	*AUC*	*SE*	*NSS*	*SE*	*CC*	*SE*
AWS [4]	0.596	0.0024	0.340	0.0386	0.028	0.0002
RARE [19]	0.727	0.0018	0.923	0.0554	0.078	0.0004
PQFT [6]	0.678	0.0082	0.772	0.1502	0.065	0.0011
IS [8]	0.710	0.0009	0.756	0.0157	0.062	0.0001
Ours	**0.844**	0.0036	**1.662**	0.0439	**0.141**	0.0005

Table 3: Comparisons of saliency models on crowd videos in CRCNS-CROWD dataset.



(a) (b)

Figure 5: AUC on (a) Outdoor, (b) Sports scenarios.

3.2 Evaluations on CRCNS-CROWD Dataset

We further evaluate our model on a more challenging CRCNS-CROWD dataset [10]. We select totally 10 crowd videos, including "Outdoor", "Sports" and "TV news", etc, which is provided together with eye traces of 8 subjects.

Together, we quantitatively compare our model with four saliency models that show superior performances on ASCMN-CROWD dataset, i.e., AWS, RARE, PQFT and IS. The AUC, NSS and CC values with SE on CRCNS-CROWD dataset are provided in Table 3. Consistent with the results on ASCMN-CROWD dataset, our model performs significant improvements compared with other models. To gain a deeper insight, AUC values for videos in "Outdoor" and 'Sports" are shown in Figure 5, suggesting our model's ability to adapt to diversified scenarios of crowds. Objective evaluation shows our model yields more correct and robust eye fixation prediction on various scenes. Our model tends to learn the scene-specific deep representation that is less distracted by the common crowd patterns than existing models. Moreover, it is capable of highlighting salient regions in crowd context through deep networks. The great improvement shows the scene adaptiveness and robustness of our model as well.

4. CONCLUSION

In this paper, we have proposed a deep-structured saliency model for predicting eye fixations in crowd videos, which benefits from the rich and discriminative properties of S-DAE to discover crowd seeds and measure the contrast-based saliency. Our proposed model takes advantages of deep representation and reconstruction capability of deep networks, which simultaneously explores perceptual properties of both crowd and saliency. Comprehensive experiments are conducted on crowd videos of the publicly available benchmarks, which demonstrates superior results with comparisons to the state-of-the-art methods.

5. ACKNOWLEDGMENTS

This work was supported in part by National Basic Research Program of China (973 Program):2015CB351802 and 2012CB316400, in part by National Natural Science Foundation of China:61133003, 61332016, 61390510, 61572465. The authors also thank the suggestion and help from Rongrong Ji and Pengfei Xu.

6. REFERENCES

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *IEEE CVPR*, pages 1597–1604, 2009.

[2] D. Culibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic, and D. Kukolj. Salient motion features for video quality assessment. *IEEE TIP*, 20(4):948–958, 2011.

[3] J. K. Dutta and B. Banerjee. Online detection of abnormal events using incremental coding length. In *AAAI*, 2015.

[4] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Decorrelation and distinctiveness provide with human-like saliency. In *Advanced concepts for intelligent vision systems*, pages 343–354. Springer, 2009.

[5] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE TPAMI*, 34(10):1915–1926, 2012.

[6] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *IEEE CVPR*, pages 1–8, 2008.

[7] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006.

[8] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE TPAMI*, 34(1):194–201, 2012.

[9] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *IEEE CVPR*, pages 1–8, 2007.

[10] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *IEEE CVPR*, volume 1, pages 631–637, 2005.

[11] R. M. Jiang and D. Crookes. Visual saliency estimation through manifold learning. In *AAAI*, 2012.

[12] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. Mack, and J. Dodd. Principles of neural science. *McGraw-Hill.*, 50(6):823–839, 2000.

[13] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang. Saliency detection via dense and sparse reconstruction. In *IEEE ICCV*, pages 2976–2983, 2013.

[14] Y. Li, Y. Zhou, J. Yan, Z. Niu, and J. Yang. Visual saliency based on conditional entropy. In *ACCV 2009*, pages 246–257. Springer, 2010.

[15] M. Liang and X. Hu. Predicting eye fixations with higher-level visual features. *IEEE TIP*, 24(3):1178–1189, 2015.

[16] M. K. Lim, V. J. Kok, C. C. Loy, and C. S. Chan. Crowd saliency detection via global similarity structure. In *IEEE ICPR*, pages 3957–3962, 2014.

[17] Z. Liu, X. Zhang, S. Luo, and O. Le Meur. Superpixel-based spatiotemporal saliency detection. *IEEE TCSVT*, 24(9):1522–1540, 2014.

[18] C. C. Loy, T. Xiang, and S. Gong. Salient motion detection in crowded scenes. In *ISCCSP*, pages 1 – 4, 2012.

[19] M. Mancas, N. Riche, J. Leroy, and B. Gosselin. Abnormal motion selection in crowds using bottom-up saliency. In *IEEE ICIP*, pages 229–232, 2011.

[20] N. Riche, M. Mancas, D. Culibrk, V. Crnojevic, B. Gosselin, and T. Dutoit. Dynamic saliency models and human attention: a comparative study on videos. In *ACCV*, pages 586–598. Springer, 2012.

[21] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12):15, 2009.

[22] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JLMR*, 11:3371–3408, 2010.

[23] N. Wang and D.-Y. Yeung. Learning a deep compact image representation for visual tracking. In *NIPS*, pages 809–817, 2013.

[24] K. Zeitz, H. Tan, M. Grief, and C. Zeitz. Crowd behavior at mass gatherings: a literature review. *Prehospital and disaster medicine*, vol.24(1):pp.32, 2010.

[25] J. Zhang, M. Wang, J. Gao, Y. Wang, X. Zhang, and X. Wu. Saliency detection with a deeper investigation of light field. In *IJCAI*, 2015.

[26] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32, 2008.

[27] Y. Zhang, L. Qin, R. Ji, H. Yao, and Q. Huang. Social attribute-aware force model: Exploiting richness of interaction for abnormal crowd detection. *IEEE TCSVT*, 25(7):1231–1245, 2015.

[28] Y. Zhang, L. Qin, S. Zhang, H. Yao, and Q. Huang. Formation period matters: Towards socially consistent group detection via dense subgraph seeking. In *ICMR*, 2015.

[29] Y. Zhang, L. Qin, S. Zhao, R. Ji, X. Lu, H. Yao, and Q. Huang. Crowd video retrieval via deep attribute-embedding graph ranking. In *ICME*, 2016.

[30] S.-h. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren. Video saliency detection via dynamic consistent spatio-temporal attention modelling. In *AAAI*, 2013.