Fixation Prediction in Videos using Unsupervised Hierarchical Features

Julius Wang, Hamed R. Tavakoli, Jorma Laaksonen
Department of Computer Science
Aalto University, Finland
{tzu-jui.wang, hamed.r-tavakoli, jorma.laaksonen}@aalto.fi

Abstract

This paper presents a framework for saliency estimation and fixation prediction in videos. The proposed framework is based on a hierarchical feature representation obtained by stacking convolutional layers of independent subspace analysis (ISA) filters. The feature learning is thus unsupervised and independent of the task. To compute the saliency, we then employ a multiresolution saliency architecture that exploits both local and global saliency. That is, for a given image, an image pyramid is initially built. After that, for each resolution, both local and global saliency measures are computed to obtain a saliency map. The integration of saliency maps over the image pyramid provides the final video saliency. We first show that combining local and global saliency improves the results. We then compare the proposed model with several video saliency models and demonstrate that the proposed framework is capable of predicting video saliency effectively, outperforming all the other models.

1. Introduction

Visual saliency is the quality of an item, e.g., an object, region or superpixel, standing out relative to its neighbors. Visual saliency is often considered a fast bottom-up, feed-forward process, which sometimes can be overridden by top-down and task-driven factors. In computer vision, saliency detection techniques are favored as a preprocessing step to curtail the immense amount of visual information and to speedup algorithms. Saliency techniques have hence been employed in applications such as signal compression [12], object detection [16], object recognition [11], object tracking [2, 26], unsupervised background subtraction [25], and video summarization [20], where efficiency is a concern.

Motivated by the wide range of applications (see [22] for a review), there has been an increasing interest in the topic of saliency prediction. In computer vision community, the saliency research has followed two main tracks: (1) fixation prediction, and (2) salient object detection [1]. The first one deals with predicting the locations a human observer will be looking at in images, while the second deals with segmentation of the most salient object. Although the two problems are highly related, the formulation and evaluation criteria are different. In this paper, we address the saliency estimation problem as fixation prediction in natural videos.

2. Related Work

The saliency prediction problem is widely studied for still images [3]. The video saliency has been, however, overlooked in comparison to image saliency. In this section, for brevity, we briefly review some of the video saliency methods and refer the readers to existing surveys [9].

An early method of video saliency prediction is [8], which extends [15] by incorporating a foveation mechanism, motion and flicker features. Itti and Baldi [14] estimate video saliency based on Bayesian surprise, which quantifies the saliency by measuring the differences between posterior and prior beliefs of the observers. The fast and efficient saliency model [23] is extended to video by employing a spherical structure, where a Bayesian center-surround is used [24]. A Bayesian approximation is also derived by [28] using self-resemblance features in order to predict saliency in videos.

Inspired by the information maximization model of attention [4], a technique based on temporal and spatial decorrelation is proposed in [31]. It extracts maximum change over time by using principal component analysis (PCA) to compute temporal saliency while non-negative matrix factorization (NMF) is used to extract the spatial saliency. The two cues are merged in order to infer the final saliency. Mancas et al. [7] compute saliency in terms of the rarity of features, where motion amplitude and direction are used as a temporal feature in the video domain.

There exists some recent works using deep learning techniques. Chaabouni et al. [5] employed transfer learning to adapt a previously trained deep network to the task of saliency prediction in natural videos. Their motivation for transfer learning is the relatively small corpus of video

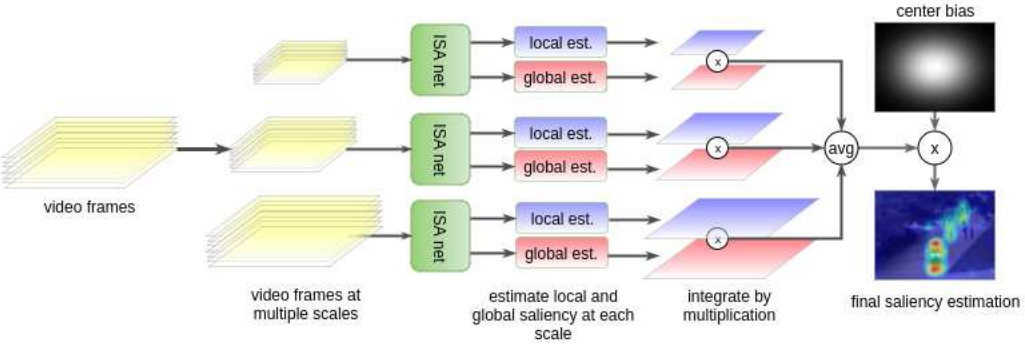

center bias
ISA net
local est.
global est.
x
ISA net
local est.
global est.
x
avg
x
ISA net
local est.
global est.
x
video frames
video frames at multiple scales
estimate local and global saliency at each scale
integrate by multiplication
final saliency estimation

Figure 1: General overview of the saliency prediction

databases with human fixations for the training of a deep network. Alternatively, to get around the limited amount of training data in end-to-end learning pipelines, we employ an approach based on unsupervised greedy layer-wise learning of features from natural image statistics, which approximates deep learning techniques [18]. While such a representation may not be optimal, it has been proven appropriate and efficient for action recognition in videos [17].

The most relevant prior work to the current framework is our previous unsupervised hierarchical model (UHM) [30] for saliency prediction in still images. In essence, we are here extending UHM for video stimuli. To achieve this end, we learn a spatio-temporal feature representation. In learning features from image statistics, contrary to the hand-crafted features, temporal properties are explicitly learned from data in an unsupervised process.

To summarize, our contributions are: (1) extending the UHM model for video, (2) reformulating the model to include spatial prior information, and (3) assessing the performance of the proposed model using a standard saliency benchmark [27] for video stimuli.

3. Method

Figure 1 illustrates an overview of the proposed fixation prediction framework. For a series of video frames, we first form a pyramid representation. On each scale, the video frames are processed using the proposed saliency prediction pipeline, based on stacked ISA networks. The saliency pipeline computes both global and local saliency, which are fused to obtain a unique conspicuity map at each scale. The intermediate maps are eventually averaged to derive the final saliency map, which can be further enhanced using a spatial prior (e.g. a center bias).

3.1. Saliency Estimation

We define the saliency, **Sal** as

$$\mathbf{Sal} = P(S|F,X) = P(S|F)P(S|X), \tag{1}$$

where S is a random variable indicating saliency, F represents an RGB image, X is the spatial locations, $P(S|X)$ determines the spatial prior, and

$$P(S|F) = \frac{1}{m}\sum_{\gamma=1}^{m}\mathcal{S}_l(F,\gamma)\mathcal{S}_g(F,\gamma), \tag{2}$$

which defines the saliency from the image, γ indicates scale, $\mathcal{S}_l(\cdot,\cdot)$ and $\mathcal{S}_g(\cdot,\cdot)$ are the functions estimating local and global saliency at a given scale, respectively.

Local saliency. To measure local saliency, we employ Shannon's self-information measure in small image patches via a sliding window procedure. That is, a window sweeps the image and $\mathcal{S}_l(F,\gamma)$ is evaluated for each patch and mapped to the central pixel of the patch. This process produces the local saliency for all the pixels in the image. We thus define local saliency of an image patch, $\mathbf{p}$, as:

$$\mathcal{S}_l(F=\mathbf{p},\gamma) = -\log(P(\mathbf{f}^\gamma)) = -\sum_{i=1}^{n}\log(P(f_i^\gamma)) \tag{3}$$
$$\mathbf{f}^\gamma = \mathrm{ISAnet}(\mathbf{p},\gamma),$$

where f_i is the i-th feature of the n-dimensional feature vector $\mathbf{f}$, obtained by employing the ISA network, denoted by the function $\mathrm{ISAnet}(\cdot,\cdot)$. We will discuss the ISA network in full details later. Under the normality assumption, $f_i^\gamma \sim \mathcal{N}(0,1)$, the local saliency is obtained by

$$\mathcal{S}_l(F,\gamma) = \alpha + \frac{1}{2}\sum_{i=1}^{n}(f_i^\gamma)^2, \tag{4}$$

where α is a constant. It is worth noting that the local saliency as formulated above corresponds to the amount of self-information in the center of the fovea given a small neighbourhood.

Global saliency. In the proposed method, the global saliency is modeled in terms of the equilibrium distribution of a random walk on an ergodic Markov chain, a chain in which it is possible to go from every state to every state. Thus, a graph-based representation is adopted where each pixel in the image is a node in the graph and a state in the random walk. That is, given an image consisting of several patches, denoted $F = \{\mathbf{p}_i, i = 1, \cdots, q\}$, the global saliency of the image consists of the saliency of the individual patches as

$$\begin{aligned} \mathcal{S}_g(F = \{\mathbf{p}_i, i = 1, \cdots, q\}, \gamma) = \{e^{\frac{-1}{k\pi_i}}; i = 1, \cdots, q\} \\ \boldsymbol{\pi}\mathbf{A} = \boldsymbol{\pi} \\ \mathbf{A} = \psi(F, \gamma), \end{aligned} \tag{5}$$

where π_i is the i-th element of the eigenvector $\boldsymbol{\pi}$ corresponding to the equilibrium distribution, k is a smoothing factor set to the maximum dimension of the image, and $\mathbf{A}$ is a stochastic transition matrix computed by a transition function $\psi(\cdot, \cdot)$ as follows:

$$\begin{aligned} \psi(F = \{\mathbf{p}_i, i = 1, \cdots, q\}, \gamma) = \begin{cases} \frac{e^{-\mathcal{D}(\mathbf{f}_i^\gamma, \mathbf{f}_j^\gamma)}}{\sum_z e^{-\mathcal{D}(\mathbf{f}_i^\gamma, \mathbf{f}_z^\gamma)}}, & i \neq j \\ 0, & i = j \end{cases} \\ \mathbf{f}_i^\gamma = \text{ISAnet}(\mathbf{p}_i, \gamma), \end{aligned} \tag{6}$$

where $\mathcal{D}(\cdot, \cdot)$ is a distance function. We are employing $1-\rho$ as the distance, where ρ is the Spearman's rank correlation coefficient between two given vectors. Given image patches and their corresponding feature scale γ, the transition function $\psi(\cdot, \cdot)$ computes the probabality of transition between all the patches as a transition matrix.

Spatial prior. Spatial priors can be used to boost the performance, particularly in a top-down attention model where a task-specific model is considered [21], e.g., in the task of detecting cars. While top-down models can exploit a prior based on the locations of specific items, bottom-up models follow a simpler prior based on the fact that many interesting items are located in the center of images. In other words, they exploit center-bias phenomenon [29]. Thus, we define $P(S|X)$ as a two-dimensional Gaussian distribution, i.e.,

$$P(S|X) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{7}$$

where $\boldsymbol{\mu} = (h/2, w/2)$, with h and w denoting the height and width of the image plane, and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \sigma_2)$ is a diagonal covariance matrix. We choose σ_1, σ_2 through cross-validation.

3.2. ISAnet: Network of Independent Subspace Analysis

In this section, we explain the feature representation using stacked layers of independent subspace analysis (ISA) [30]. Figure 2 illustrates the inner architecture of an ISA network (ISAnet) unit. Each ISAnet unit receives a video patch and produces a neural response, i.e., a feature representation of the patch. As depicted, the vectorized video patch is passed through two stacked layers of ISA, where the concatenation of layer responses produces the final feature representation. The ISAnet is a fast feed-forward network that is easy to employ in saliency estimation after learning the weights of the network. This network is trained in a greedy layer-wise unsupervised manner. To continue with the explanation of the training procedure, we first review the basic principles of ISA.

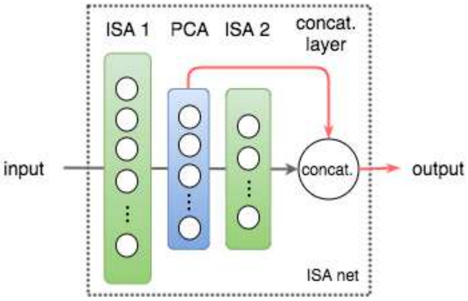


Figure 2: ISAnet architecture.

Independent Subspace Analysis (ISA). Independent subspace analysis is an unsupervised learning algorithm mimicking a neural network consisting of a hidden and an output layer [13]. The neural interpretation of ISA is illustrated in Figure 3. As depicted, ISA consists of a fully connected layer and a pooling layer and each output of the network is a non-linear function of the inputs and the weights of the layers expressed as

$$o_i(\mathbf{x}; \mathbf{W}, \mathbf{V}) = \sqrt{\sum_{z=1}^{s} V_{iz} (\sum_{j=1}^{n} W_{zj} x_j)^2}, \tag{8}$$

where $\mathbf{x}$ is an n-dimensional input vector, $\mathbf{W} \in \mathbb{R}^{s \times n}$ is the matrix of input weights, and $\mathbf{V} \in \mathbb{R}^{l \times s}$ indicates the matrix of pooling weights, and n, s, l, are the input dimensionality, and the number of simple and pooling units, respectively.

In ISA, the pooling weights $\mathbf{V}$ are typically fixed. Thus, the training procedure boils down to estimation of $\mathbf{W}$ via the following minimization for a set of q' unlabelled and whitened samples, $\{\mathbf{x}_t, t = 1, ..., q'\}$:

$$\begin{aligned} \underset{\mathbf{W}}{\text{minimize}} \sum_{t=1}^{q'} \sum_{i=1}^{l} o_i(\mathbf{x}_t; \mathbf{W}, \mathbf{V}), \\ \text{s.t.} \quad \mathbf{W}\mathbf{W}^T = \mathbf{I} \end{aligned} \tag{9}$$

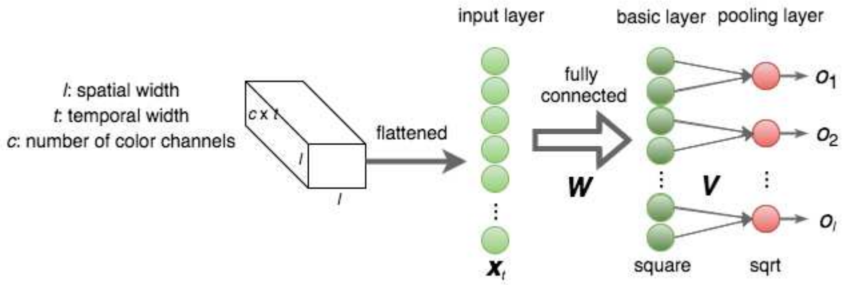


Figure 3: Independent subspace analysis as a neural architecture with a subspace of size 2, each pooling neuron looks at two simple neurons in the previous layer.

Training the ISAnet. The training of an ISAnet is done layer-wise, where one ISA neural unit is learned in each layer. The procedure is depicted in Figure 4 and has two phases. That is, batches of video patches of size $s_1 \times s_1 \times c \times t_1$, where s_1 is the spatial resolution, c number of image channels, and t_1 the temporal length, are flattened and the weights of the first ISA unit, denoted as $\mathbf{W}_1$, are estimated using (9). To train the second ISA in an efficient way, akin to [17], slightly bigger video patches are sampled and the response of the first ISA unit is calculated since the $\mathbf{W}_1$ is known and fixed. The responses of the first ISA layer are flattened to a vector and dimensionally reduced by principal component analysis (PCA) and used as input to the second ISA unit. Thus, the second ISA unit weights in $\mathbf{W}_2$ are obtained using (9), where the inputs are based on the responses of the first ISA. The first ISA unit filters extract the low-level features and the second ISA filters are representing richer contextual information such as long-term frame dependencies in the temporal domain [17].

Figure 5 visualizes some exemplar filters learned for each layer. While visualizing the $\mathbf{W}_1$ filters is straightforward, to visualize $\mathbf{W}_2$, we employed activation maximization method [10] which finds the input pattern that can maximally activate each output of a neuron in the second ISA unit.

Feature extraction using ISAnet. The feature extraction is performed using the ISAnet, depicted in Figure 2. Having the weights, $\mathbf{W}_1$ and $\mathbf{W}_2$ determined, akin to [30], we employ a sliding window approach. For each video patch, the ISAnet response is computed and used to obtain the saliency for the central pixel of the middle frame in the saliency pipeline. It is worth noting that we use the same ISAnet with different video resolutions.

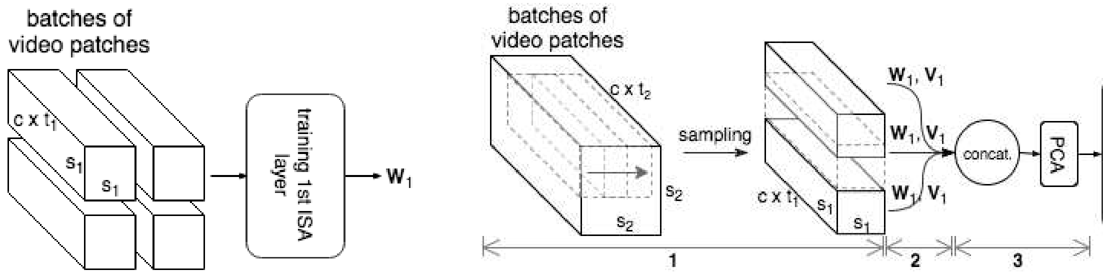
4. Experiments

We evaluated the proposed model using the ASCMN [27] dataset. ASCMN consists of five video categories, including, *abnormal*, *surveillance*, *crowd*, *moving*, and *noise*, where human fixations are provided as the ground-truth. The diversity of video types helps us to obtain a better understanding of the performance of the saliency estimation methods. Furthermore, ASCMN tries to establish a benchmark, which makes comparison to various methods easy. We thus compare the proposed model to Mancas [19], Culibrk [6], Seo [28], SUN [32], and vFES [24].

The comparison is carried out using the area under the curve (AUC) of receiver operating characteristic (ROC) curve as the performance measure as recommended by [27]. That is, the values of the estimated saliency map are treated as a classification score where the classification task is to determine if a pixel is a fixation or not. The positive samples are fixation locations and the negative samples are picked randomly from non-fixation points. To compute the metric, we use the toolbox provided in ASCMN, which makes comparison with other methods easy.

4.1. Model Parameters

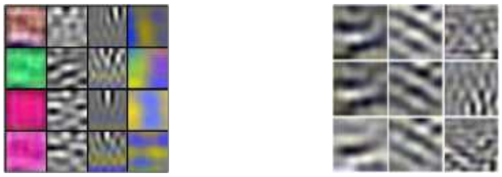
The training of ISAnet was done by random sampling of approximately one million video patches from training videos, which were randomly picked from YouTube. The resolution of each video was adjusted to 640 × 360. Afterwards, $\mathbf{W}_1$ and $\mathbf{W}_2$ were learned as described in the previous section. $\mathbf{V}_1$ and $\mathbf{V}_2$ were set to form subspaces of size 4, i.e., $l = 4$. The parameters, including, the temporal width of patches in both layers (t_1, t_2), the multi-resolution scales (γ), and (σ_1, σ_2) for the spatial prior, were chosen through repeated 5-fold cross-validation, where validation and test sets come from the ASCMN database. We ensured each fold of validations contained at least one video of each category and no overlap between videos of the folds. The parameter values are $t_1 = t_2 \in \{5, 7, 9, 11\}$, $\sigma_1 = \sigma_2 \in \{0.01, 0.05, 0.1\}$, and the scale sets for constructing videos at multiple resolutions are $\gamma \in \{\{0.5, 0.7, 1\}, \{0.7, 1, 1.4\}\}$. The spatial size of the fovea was fixed at 24×24. Nonetheless, using the fast implementation trick



(a) Training ISA.

(b) Training the second ISA unit in ISAnet.

Figure 4: Training phases of stacked convolutional ISA for videos in ISAnet. Two ISA units are trained, where first ISA unit learns $\mathbf{W}_1$ and the second one learns $\mathbf{W}_2$ meanwhile the $\mathbf{W}_1$ is fixed fixed. (a) Learning of $\mathbf{W}_1$ is similar to any ISA by sampling video patches. (b) learning of $\mathbf{W}_2$ has three steps, (1) some bigger patches are sampled, (2) the responses of $\mathbf{W}_1$ to interior patches are calculated and linearized, (3) the $\mathbf{W}_2$ are learned from the patches of linearized responses.



(a) First ISA unit filters ($\mathbf{W}_1$) (b) Second ISA unit filters ($\mathbf{W}_2$)

Figure 5: Exemplar filters of the first and second ISA units in an ISAnet. Visualization of $\mathbf{W}_2$ is carried out by the activation maximization method [10].

of [17], we sampled 32 × 32 patches for the second ISA in the training phase of the ISAnet architecture.

4.2. Performance Analysis and Discussion

The performance of the proposed model is summarized in Table 1. We report the individual performance of the local and global pipelines. While on the average the global pipeline outperforms the local one, it falls short in terms of *noise* and *crowd* categories in comparison to the local pipeline. This indicates that for an input stimuli where the elements of interest are relatively numerous and small (e.g. people in *crowd*), the local self-information is a better measure of saliency. Nonetheless, computing the saliency using the product of the global and local saliencies, i.e. the full pipeline, improves the overall performance on all the categories, except *crowd*, achieving average performance of 0.66. It is worth noting that *crowd* category seems the most difficult stimuli for the proposed framework. The global pipeline and local pipeline achieve the average performance of 0.64 and 0.62, respectively.

Spatial prior. Incorporating a spatial prior improves the performance of the proposed model significantly. On average an improvement of 21% is obtained by employing the center-bias. The largest improvement is 41% for the *crowd* category. This can be an indicator that salient motions are present in the center of *crowd* videos. Figure 6 depicts examples of the *crowd* videos and the effect of center bias on the prediction of saliency in them. Further examples from each category are provided in Figure 7. As depicted, the output is more similar to the ground-truth once the spatial prior is employed, i.e., the responses are re-weighted.

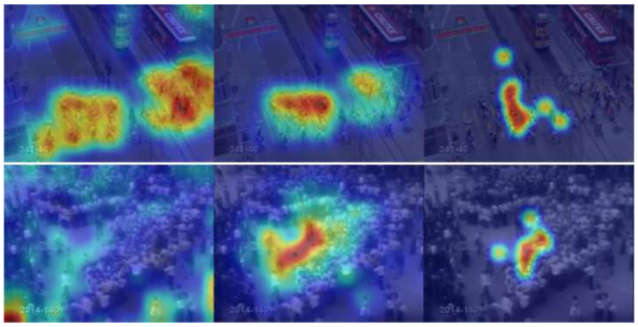


Figure 6: The effect of center bias prior on the *crowd* videos. In each row, from left to right: estimated saliency with no center prior, estimated saliency with center prior, and human fixations overlaid on the images.

Comparison. Comparing the proposed model and other saliency methods, we learn that the proposed model outperforms all the other methods. It is worth noting that this is achievable with the help of the spatial prior. Otherwise, the proposed model without such a prior outperforms the other models only in the category of *abnormal* videos. This suggests that the proposed model detects the saliency far from the center more effectively than the saliency in the center, examples of such responses can be seen in Figure 6. By reweighing the saliency map to emphasize the central area, this deficiency is, however, addressed. It is worth noting

Table 1: Performance of the proposed model, local pipeline, global pipeline, and full pipeline with and without spatial prior in comparison with other models. The scores of other models are extracted from the relevant publications. The AUC score is reported.

Model	*abnormal*	*surveillance*	*crowd*	*moving*	*noise*	mean
Proposed (local)	0.71	0.69	0.56	0.55	0.65	0.62
Proposed (global)	0.73	0.69	0.54	0.66	0.60	0.64
Proposed (full, no prior)	0.76	0.71	0.55	0.68	0.64	0.66
Proposed (full)	**0.82**	**0.80**	**0.79**	**0.83**	**0.76**	**0.80**
vFES [24]	0.75	0.77	0.74	0.79	0.75	0.76
Mancas [19]	0.74	**0.80**	0.65	0.60	0.67	0.68
Culibrk [6]	0.73	0.72	0.68	0.63	0.65	0.67
SUN [32]	0.68	0.68	0.60	0.59	0.62	0.65
Seo [28]	0.71	0.65	0.61	0.63	0.60	0.63

that the role of spatial prior is multiplicative and re-weights the detected saliency. This necessitates correct detection of salient regions compared to additive spatial prior which boosts both salient and non-salient areas.

5. Conclusion

This paper presented a method based on unsupervised hierarchical feature learning for predicting fixations in videos. The features are extracted using a hierarchy of independent subspace analysis networks, consisting of two stacked ISAs, in multiple resolutions. To predict fixations, global and local saliency are computed and combined in each resolution. The final saliency is obtained by averaging the saliency over multiple resolutions. Our evaluations revealed that the proposed model significantly improves over the other methods. Nonetheless, the spatial prior plays an important role in the results. This indicates that: (1) while the proposed model detects the salient areas, it fails to weight them correctly, and (2) there is potentially a strong center bias in the database. The second phenomenon is less investigated in videos and it needs to be studied further. Future studies will investigate deeper models and different techniques for feature learning from natural image and video statistics.

Acknowledgements. The authors would like to acknowledge the support of the Finnish Center of Excellence in Computational Inference Research (COIN) and the computational resources of the Science-IT project, Aalto University.

References

[1] A. Borji, M. Cheng, H. Jiang, and J. Li. Salient object detection: A survey. *CoRR*, abs/1411.5878, 2014. 1

[2] A. Borji, S. Frintrop, D. Sihite, and L. Itti. Adaptive object tracking by learning background context. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 23–30, 2012. 1

[3] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 99, 2012. 1

[4] N. D. B. Bruce and J. K. Tsotsos. Saliency based on information maximization. In *Advances in Neural Information Processing Systems*, NIPS'05, pages 155–162, MIT Press, 2006. MIT Press. 1

[5] S. Chaabouni, J. Benois-Pineau, and C. B. Amar. Transfer learning with deep networks for saliency prediction in natural video. In *IEEE International Conference on Image Processing*, pages 1604–1608, Sept 2016. 1

[6] D. Culibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic, and D. Kukolj. Salient motion features for video quality assessment. *IEEE Trans. Image Process.*, 20(4):948–958, 2011. 4, 6

[7] M. Decombas, N. Riche, F. Dufaux, B. Pesquet-Popescu, M. Mancas, B. Gosselin, and T. Dutoit. Spatio-temporal saliency based on rare model. In *IEEE International Conference on Image Processing*, pages 3451–3455, Sept 2013. 1

[8] N. Dhavale and L. Itti. Saliency-based multifoveated mpeg compression. In *Seventh International Symposium on Signal Processing and Its Applications*, volume 1, pages 229–232, july 2003. 1

[9] K. Duncan and S. Sarkar. Saliency in images and video: a brief survey. *IET Computer Vision*, 6(9):514–523, November 2012. 1

[10] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. Technical Report 1341, University of Montreal, June 2009. 4, 5

[11] D. Gao and N. Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In *International Conference on Neural Information Processing Systems*, NIPS'04, pages 481–488, Cambridge, MA, USA, 2004. MIT Press. 1

[12] H. Hadizadeh and I. V. Bajic. Saliency-aware video compression. *Trans. Img. Proc.*, 23(1):19–33, Jan. 2014. 1

[13] A. Hyvärinen, J. Hurri, and P. O. Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.* Springer Publishing Company, Incorporated, 1st edition, 2009. 3

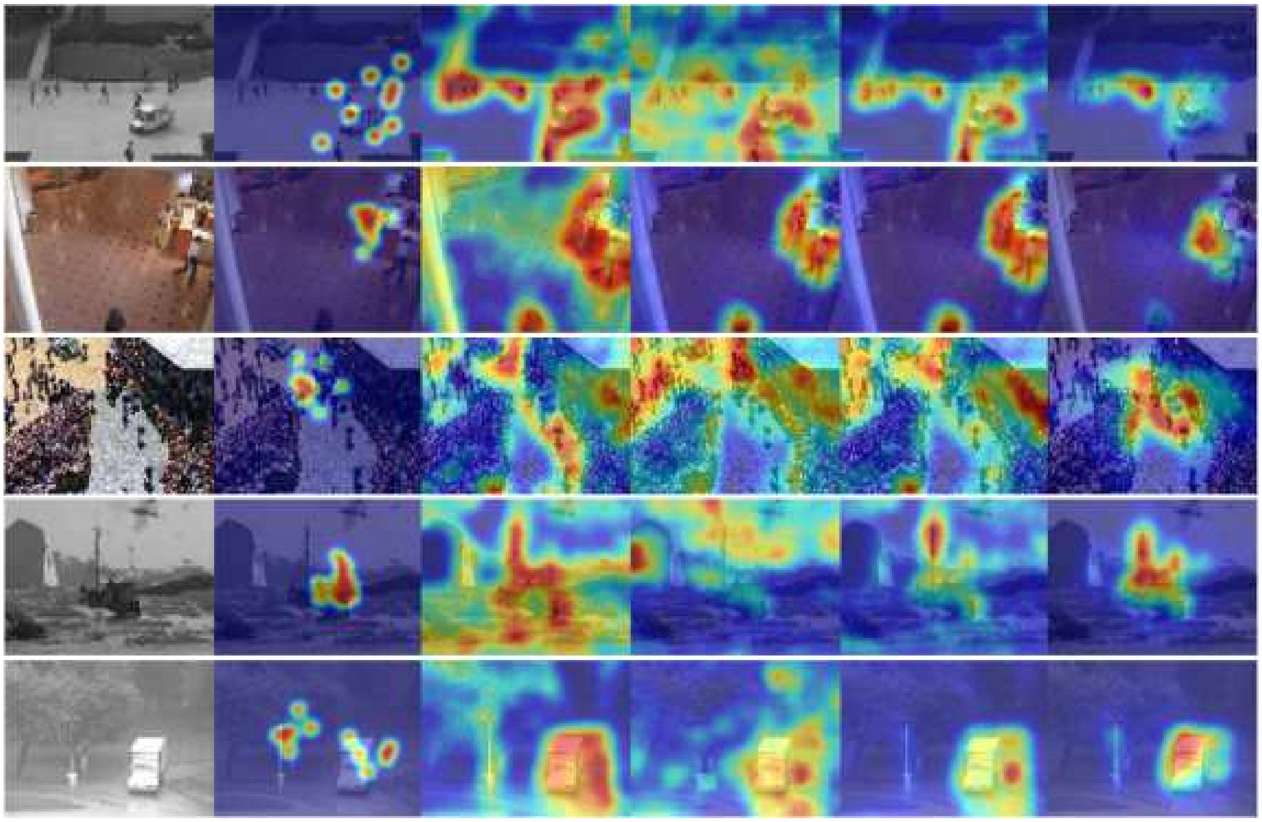


Figure 7: Example saliency maps. From top to bottom: abnormal, surveillance, crowd, moving, and noise videos, From left to right: original image, ground-truth, local saliency, global saliency, full saliency pipeline without spatial prior, and full saliency pipeline with spatial prior.

[14] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 631–637, June 2005. 1

[15] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, Nov. 1998. 1

[16] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 2083–2090, Washington, DC, USA, 2013. IEEE Computer Society. 1

[17] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 3361–3368, Washington, DC, USA, 2011. IEEE Computer Society. 2, 4, 5

[18] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Annual International Conference on Machine Learning*, ICML '09, pages 609–616, New York, NY, USA, 2009. ACM. 2

[19] M. Mancas, N. Riche, J. Leroy, and B. Gosselin. Abnormal motion selection in crowds using bottom-up saliency. In *IEEE International Conference on Image Processing*, pages 229–232. IEEE, 2011. 4, 6

[20] S. Marat, M. Guironnet, and D. Pellerin. Video summarization using a visual attention model. In *European Signal Processing Conference*. 1

[21] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson. Top-down control of visual attention in object detection. In *International Conference of Image Processing*, pages 253–256, 2003. 3

[22] H. Rezazadegan Tavakoli. *Visual Saliency and Eye Movement: Modeling and Applications*. PhD thesis, University of Oulu, 2014. 1

[23] H. Rezazadegan Tavakoli, E. Rahtu, and J. Heikkilä. Fast and efficient saliency detection using sparse sampling and kernel density estimation. *Image Analysis*, 6688:666–675, 2011. 1

[24] H. Rezazadegan Tavakoli, E. Rahtu, and J. Heikkilä. Spherical center-surround for video saliency detection using sparse sampling. In *Advanced Concepts for Intelligent Vision Systems. Lecture Notes in Computer Science, vol 8192*, volume 8192 of *Lecture Notes in Computer Science*, pages 695–704, 2013. 1, 4, 6

[25] H. Rezazadegan Tavakoli, E. Rahtu, and J. Heikkilä. Temporal saliency for fast motion detection. In J.-I. Park and

J. Kim, editors, *Computer Vision - ACCV 2012 Workshops*, volume 7728 of *Lecture Notes in Computer Science*, pages 321–326. Springer Berlin Heidelberg, 2013. 1

[26] H. Rezazadegan Tavakoli, M. Shahram Moin, and J. Heikkilä. Local similarity number and its application to object tracking. *International Journal of Advanced Robotic Systems*, 10(184), 2013. 1

[27] 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 *11th Asian Conference on Computer Vision*, ACCV'12, pages 586–598, Berlin, Heidelberg, 2012. Springer-Verlag. 2, 4

[28] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *J. Vis.*, 9(12):15–15, 2009. 1, 4, 6

[29] B. Tatler. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor bases and image feature distributions. *J. Vis.*, 14(7), 2007. 3

[30] H. R. Tavakoli and J. Laaksonen. Bottom-up fixation prediction using unsupervised hierarchical models. In *Assistive Vision - ACCV 2016 Workshops*, 2016. 2, 3, 4

[31] H. R. Tavakoli, E. Rahtu, and J. Heikkilä. Stochastic bottom-up fixation prediction and saccade generation. *Image and Vision Computing*, 31(9):686–693, 2013. 1

[32] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *J. Vis.*, 8(7):32–32, 2008. 4, 6