A Computational Model of Multi-scale Spatiotemporal Attention in Video Data

Roman Palenichka[1](✉), Rafael Falcon[1,2], Rami Abielmona[1,2], and Emil Petriu[1]

[1] University of Ottawa, Ottawa, Canada
rpalenyc@uottawa.ca
[2] Larus Technologies, Ottawa, Canada

Abstract. This paper describes a spatiotemporal saliency-based attention model in applications for the rapid and robust detection of objects of interest in video data. It is based on the analysis of feature-point areas, which correspond to the object-relevant focus-of-attention (FoA) points extracted by the proposed multi-scale spatiotemporal operator. The operator design is inspired by three cognitive properties of the human visual system: detection of spatial saliency, perceptual feature grouping, and motion detection. The model includes attentive learning mechanisms for object representation in the form of feature-point descriptor sets. The preliminary test results of attention focusing for the detection of feature-point areas have confirmed the advantage of the proposed computational model in terms of its robustness and localization accuracy over similar existing detectors.

Keywords: Spatiotemporal attention · Video data · Attention operator
Local scale · Object detection · Spatial saliency · Property coherence
Temporal change

1 Introduction

The study of biologically-inspired vision systems can be considered as a two-way street. On the one hand, biological systems can provide a source of inspiration for new computationally-efficient and robust models while, on the other hand, computer vision approaches reveal new insights for understanding biological sensing systems such as the Human Visual System (HVS) [1]. This paper is dedicated to a saliency-based multi-scale spatiotemporal attention in video data mostly for computer vision objectives. We consider applications for object detection in video data of dynamic scenes, which usually proceeds by initially extracting some object-relevant spatiotemporal features [2]. Their extraction is based on the detection of spatiotemporal attention points. The bio-inspired approach has certain advantages over the conventional methods using the object-background segmentation, because video segmentation of dynamic scenes is a computationally complex and error-prone process [3, 4].

Most of the existing computational attention models deal with still images [5–8] and few actually tackle video data. One of such spatiotemporal detector models was proposed in [9] as a generalization of the spatial saliency-based attention [6]. This method was

A. Campilho et al. (Eds.): ICIAR 2018, LNCS 10882, pp. 125–135, 2018.
https://doi.org/10.1007/978-3-319-93000-8_15

further extended to detect objects in dynamic scenes by capitalizing on coherent motion characteristics in video frames [10]. Another general spatiotemporal attention model was described in [11]. A spatiotemporal isotropic attention (STIA) operator was proposed to detect attention points, which explicitly combines spatial saliency and temporal change [12]. It is a multi-scale area-based operator, in which the spatial saliency is defined as the area isotropic spatial contrast relative to the homogeneity of an image area.

An entire class of attention models is represented by detectors of spatiotemporal feature (interest) points in video data. Detection of feature points is based on various local image properties such as area saliency, temporal change, motion, shape, local area homogeneity, etc. [13]. The feature point extraction is a computationally simpler and more reliable procedure than the spatiotemporal image segmentation. The first spatiotemporal feature-point detector, called the Harris 3D detector, was proposed in [14] as a space-time extension of the Harris detector [15]. It is based on the computation of a spatiotemporal second-moment matrix at each video point using scale selection, Gaussian smoothing functions, and space-time gradients. Another detector of feature points is the generalization of the Laplacian-of-Gaussian (LoG) operator to the space-time domain with the selection of spatial and temporal scales [16]. The Hessian detector was proposed in [17] as a spatiotemporal version of the Hessian saliency measure used in [18] for blob detection in images. The Cuboid detector of feature points is mostly based on the spatiotemporal Gabor filters [19].

The main drawback of the abovementioned detectors is that simple extensions of spatial detectors to the temporal domain, through the introduction of the time variable, will result in poor detection of still and moving objects at the same time. Another weakness of the existing models is their inability to computationally represent the perceptional grouping of local low-level features to avoid getting distracted by irrelevant low-level features. This is an important element of the so-called *gestalt* model for the HVS's attention focusing [20]. Neural networks and the Deep Learning models are biologically inspired approaches too that can be utilized for feature extraction and object detection [21, 22]. They are not considered in our current study as they do not provide attention focusing mechanisms.

The computational model of visual attention proposed in this paper is primarily aimed at eliminating or diminishing shortcomings of the existing models in computer vision applications. It is achieved through the introduction of a new spatiotemporal attention operator, which combines basic cognitive hypothesises of the HVS such as the multi-scale attention through spatial saliency, temporal change detection and perceptual feature grouping. Another contribution is the tuneable attention operator via machine learning algorithms to make it relevant to the objects of interest. In the literature, except for the deep learning approach, no attention focusing models related to learning processes are considered.

The rest of this paper is concentrated on the extraction of spatiotemporal attention points based on the proposed computational model. Section 2 gives an overview of the adopted approach in the context of object detection tasks. The cornerstone of the model is our proposed spatiotemporal attention operator (Sect. 3). The model's application capability for object detection is shortly discussed in Sect. 4 since the detailed handling is out of scope of the current paper. The experimental results (Sect. 5) and the algorithm's advantageous characteristics (Sect. 6) for the attention-point detection confirm the viability of this approach.

2 Spatiotemporal Visual Attention

The computational model of spatiotemporal visual attention consists in attention-guided image sequence analysis by first extracting multi-scale spatiotemporal attention regions called feature-point areas (FPAs) and sequentially analyzing in detail the neighborhoods of the FPAs for object detection and classification. The attention-point area, which is currently analyzed in detail, is called the focus-of-attention (FoA) area. The FoA points are determined by the local maxima locations of the proposed multi-scale spatiotemporal attention operator. It considers three spatiotemporal image area characteristics for the determination of FPAs: spatial saliency, area properties' coherence (e.g., area homogeneity by an image local property) and temporal change (e.g., motion).

The flowchart of the proposed computational model is shown in Fig. 1. In the latter, all the attention points are first determined at a single spatiotemporal resolution of the video data. Initially, some simple image properties are time-efficiently extracted in a dense mode, i.e., pixel-by-pixel. The current FoA point is determined as the new local maximum of the attention operator excluding previously analyzed FoA-point areas. Object detection, tracking and classification is based on spatiotemporal descriptor sets estimated in the corresponding FPAs. No image pre-segmentation into object and background regions is required. Figure 1 describes a single-stage spatiotemporal attention model for multi-scale image analysis; however, a multi-stage hierarchical computational model is a straightforward generalization [2]. During the first stage, attention focusing at a single largest scale or narrow scale range is performed, while in subsequent stages, the FoA areas are analyzed at lower scale ranges (or higher image resolutions) by the same computational model in Fig. 1.

The proposed model is a multi-scale approach to image sequence analysis, which involves the concept of local spatial scale [12]. It is the diameter of a circular area, which is homogeneous according to one or more image properties. A local temporal scale can also be introduced in the video data analysis similarly to the definition proposed in [23]. It characterizes how fast the temporal change or motion occurs at a given location and for the determined local spatial scale. The model in Fig. 1 includes a machine learning stage powered by the attentive learning mechanisms. The goal is to effectively store objects' spatiotemporal descriptors in the form of reference sets of video descriptors to perform matching of descriptor sets. Another objective of the attentive learning is to automatically tune the parameter values of the computational attention model such as the saliency coefficients in the attention operator (Sect. 3). The attentive machine learning uses FPA extraction to obtain reference sets of FPA descriptors from the training samples of short-duration videos as the centers of descriptor-set clusters [24].

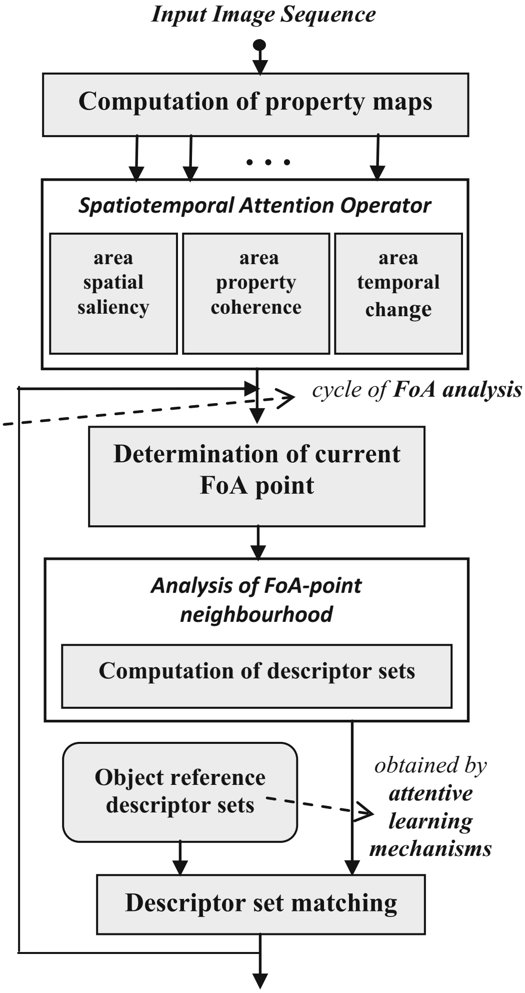

Input Image Sequence
Computation of property maps
Spatiotemporal Attention Operator
area spatial saliency
area property coherence
area temporal change
cycle of FoA analysis
Determination of current FoA point
Analysis of FoA-point neighbourhood
Computation of descriptor sets
Object reference descriptor sets
obtained by attentive learning mechanisms
Descriptor set matching

Fig. 1. The proposed visual attention model.

3 Multicomponent Spatiotemporal Attention Operator

The cornerstone of the proposed model is the multi-scale spatiotemporal attention operator designed to fulfil the extraction of attention points in *spatiotemporally salient* (i.e., detectable), *locally unique* (i.e., unambiguous), and *object-relevant* (i.e., object-positioning) locations of video data. To achieve that, the operator is applied to multiple components of the input video data and contains three response factors for each component: spatial saliency, temporal change, and area homogeneity as a measure of coherence. The proposed multi-scale spatiotemporal attention operator $F[\{g(i, j, t)\}, \rho]$ as applied to the intensity image sequence $\{g(i,j,t)\}$ is composed of three terms, which are aggregated into a single attention function of pixel coordinates (i,j), time t, and scale ρ,

$$F[i,j,t,\rho] = c(i,j,t,\rho) + \alpha \cdot e(i,j,t,\rho) - \gamma \cdot h(i,j,t,\rho), \tag{1}$$

where $c(i,j,t,\rho)$ is the area spatial saliency, $e(i,j,t,\rho)$ is the area temporal change, $h(i,j,t,\rho)$ is the area inhomogeneity measure, and $\alpha > 0$ and $\gamma > 0$ are the change and coherence coefficients, respectively. The values of α and γ can be determined by the maximum likelihood rule using a representative training sample of FoA areas to determine the conditional distribution parameters, which provide the coefficients' optimal values. It is based on the probabilistic formulation of the attention-focusing mechanisms and derivation of the attention function in Eq. (1) by the maximum likelihood rule [25]. The spatial saliency $c(i,j,t,\rho)$ is defined as an area isotropic contrast [12], which is the mean value of squared intensity deviations in the background ring $Q_\rho(i,j)$ with respect to the mean intensity in the disk $S_\rho(i,j)$ for the feature area $W_\rho = S_\rho \cup Q_\rho$ (Fig. 2). The area inhomogeneity $h(i,j,t,\rho)$ is the intensity mean deviation inside the disk $S_\rho(i,j)$ [12]. The computational scheme for the temporal change $e(i,j,t,\rho)$ in Eq. (1) consists of the accumulated temporal differentiation using consecutive video frames and the isotropic contrast computation over the accumulated differentiation result.

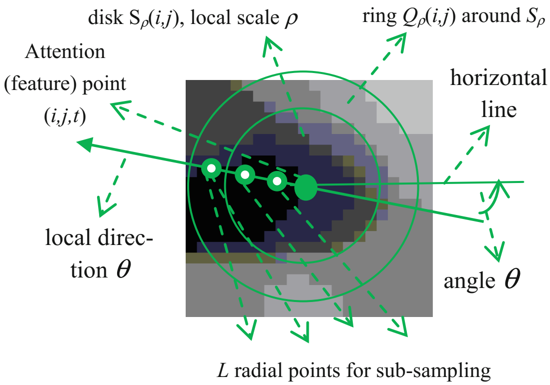

disk $S_\rho(i,j)$, local scale ρ
ring $Q_\rho(i,j)$ around S_ρ
Attention (feature) point (i,j,t)
horizontal line
local direction θ
angle θ
L radial points for sub-sampling

Fig. 2. Attention point detection and estimation of area direction: example of a corner area.

The operator in Eq. (1) is generalized as acting on multiple video (image) property maps. The property maps include sensors' raw components such as the three-color components, infrared and multispectral components if available. Additionally, computationally simple property maps such as the spatial and temporal derivatives of different orders, smoothing filters as well as other local object-relevant functions can be included as the property maps. The multi-component spatiotemporal isotropic attention (MSIA) function $F[i,j,t,\rho]$ becomes:

$$F[i,j,t,\rho] = \sum_{n=1}^{N} \beta_n \cdot s_n(i,j,t,\rho), \tag{2}$$

where $s_n(i,j,t,\rho)$ is the nth spatiotemporal saliency component corresponding to nth property map and β_n is the nth saliency coefficient. The derivation of the optimized values for $\{\beta_n\}$ in the MSIA operator is based on the Fisher's Linear Discriminant Analysis (LDA) method [25]. The LDA-based transformation of N saliency components onto a single axis of $F[i,j,t,\rho]$ values is aimed at maximizing the difference between the $F[i,j,t,\rho]$ values for object and background points respectively while minimizing differences between $F[i,j,t,\rho]$ values corresponding to the object points only.

The sequential detection and analysis of the FoA points $\left\{(u,v,\tau)_k\right\}$ and their associated local scale values $\{r_k\}$ as the FPAs' diameters proceeds as follows:

$$(u,v,\tau,r)_k = \underset{(i,j,t)\in A,\rho\in\Omega}{\arg\max}\ \{F[i,j,t,\rho],(i,j,t)\bar{\in}Z_{k-1}\}, \tag{3}$$

where $Z_{k\text{-}1}$ is the set of previously detected FoA points, Ω is the scale range of the attention operator, and A is a subset of video data under current analysis. A *fast-recursive implementation* of the MSIA operator alike the recursive STIA algorithm described in [12] makes the computation independent of the window size and becomes O(N) per pixel and per scale value, where N is the total number of property maps (Eq. 2). It is based on the recursive implementation of 2D filters with the square window shape, which approximate the circular (isotropic) window $W_\rho = S_\rho \cup Q_\rho$.

4 Application to Object Detection in Video Data

One of the applications of the proposed attention model is the fast and robust detection of objects of interest in video data. Multiple-object detection proceeds by sequentially detecting object-relevant subsets of FPAs in the input image sequence. This is implemented through the clustering of feature-point areas around the current FoA-point. A descriptor set is extracted for each subset of FPAs. Three different types of descriptors are extracted to form a single descriptor set: (1) area pose; (2) local appearance; and (3) temporal change. To achieve rotation invariance, local appearance descriptors are extracted through the generalization of the Radial Descriptor Pattern (RDP) algorithm, which was originally used to rotation-invariantly extract planar shape descriptors [26]. The RDP algorithm consists of two basic steps: (a) determination of a dominant direction; and (b) estimation of angular-radial descriptor components (Fig. 2). The dominant direction is the *local direction* angle θ included in the pose descriptors [26]. The second step is the angular-radial sub-sampling of image intensity within the window $W_\rho = S_\rho \cup Q_\rho$, for M angle values and L radial points per angular position (Fig. 2). The values of L and M are determined by considering a tradeoff between the accuracy of description and computational costs. To detect and classify objects of interest, the computational model in Fig. 1 proceeds by matching the observed FPA descriptors with the reference ones obtained during the machine learning stage. The Euclidian distance between descriptor sets can be used as the dissimilarity measure of matching.

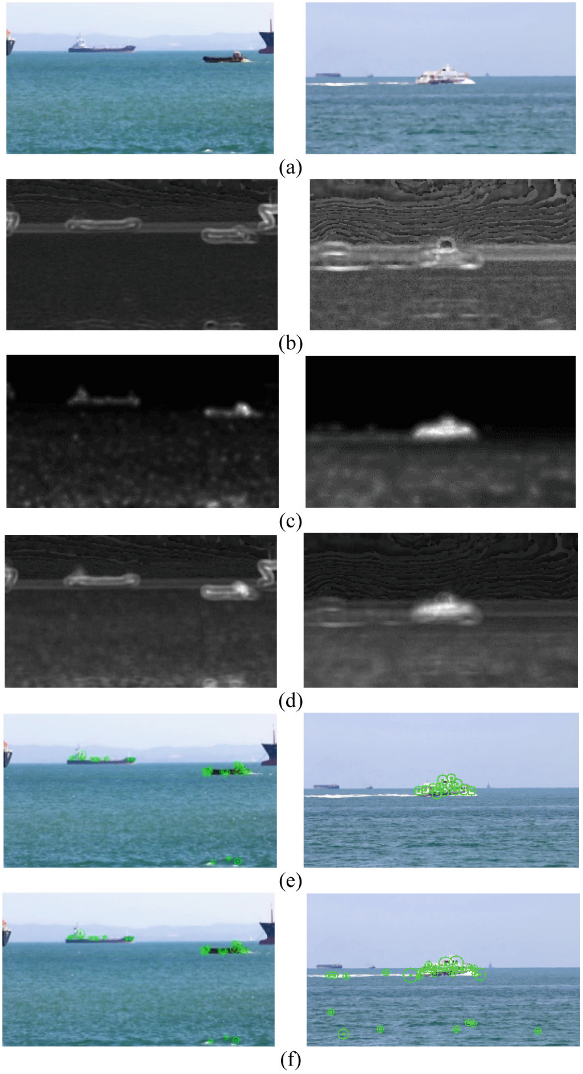
5 Experimental Results

We conducted numerous experiments to investigate several basic performance characteristics of the proposed attention model. The latter was tested on the operator adequate response (detection capability) to high-contrast areas, high property-coherence areas, and areas with object rigid motion. Another type of experiments was the accuracy of the FoA point localization and spatial scale determination. We used the Singapore maritime datasets (https://sites.google.com/site/dilipprasad/home/singapore-maritime-dataset). The application is vessel detection in maritime scenes.

The testing was comparative with respect to existing computational models. We compared the model's performance with the following algorithms: attention model-based STIA algorithm [12] and feature-point detector in the HSIP method (Harris' Spatiotemporal Interest Points) [14]. Figure 3 illustrates the computation of the MSIA operator and the process of attention-point extraction for the color maritime videos. The MSIA operator uses 8 spatial scales and 5 spatiotemporal saliency components in Eq. (2): three normalized RGB components, mean intensity component, and temporal change. Coefficients $\{\beta_1 = 0.15, \beta_2 = 0.08, \beta_3 = 0.05, \beta_4 = 0.19, \beta_5 = 0.35\}$ in Eq. (2) were learned by the LDA approach (Sect. 3).

The performance of attention focusing in terms of correct response to such video stimuli as spatial saliency, coherence and motion presence is reported in Table 1 as the F-measure, which combines the standard precision and recall rates [24]. Table 1 provides the comparative performance of attention-point detection in general. The MSIA operator provides adequate detection of FoA points due to the effective combination of saliency components in Eq. (2) and the introduction of a new temporal change filter. Many false attention points were detected by the HSIP method in water ripples, specular reflections and ship wakes. A drawback of the HSIP detector as well as other feature-point detectors is the instability of feature point extraction due to their sensitivity to irrelevant spatiotemporal changes (Fig. 3f). The MSIA operator locates the FPAs with the priority to be inscribed into object regions (Fig. 3e). Estimated errors for the FoA-point localization and local scale estimation are given in Table 2. The normalized root mean-squares error (RMSE) was used to characterize the localization accuracy.

The normalized run-time for the proposed MSIA operator is summarized in Table 3. The normalization consists in dividing the current run-time by that of the minimal window size (3 × 3 pixels) to get rid of a particular CPU speed figure. For comparison purposes, we estimated the run-time of direct (non-recursive) implementation for the MSIA-based attention focusing as well as the HSIP-based attention model using the fast (iterative) Gaussian smoothing algorithm.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 3. Attention point extraction in maritime videos: (a) initial video images; (b) spatial saliency map; (c) temporal change map; (d) MSIA operator map; (e) MSIA points; (f) HSIP points;

Table 1. *F*-measure of attention focusing models by the stimuli response.

Response of attention focusing to	HSIP detector	Attention operator STIA	Attention operator MSIA
Area saliency	0.81	0.80	**0.86**
Area coherence	0.71	0.78	**0.92**
Area motion	0.75	0.79	**0.91**
Attention point	0.72	0.77	**0.87**

Table 2. RMSE-based accuracy of FoA-point localization and local scale estimation.

Algorithm	FoA localization		Local scale	
	Low scale range	High scale range	Low scale range	High scale range
MSIA operator	0.03	0.18	0.11	0.09
HSIP detector	0.31	0.34	0.3	0.18
STIA operator	0.23	0.2	0.05	0.13

Table 3. Normalized run-time per pixel of the attention-focusing process.

Window (scale) size:	HSIP-based attention	Direct computation MSIA	Fast recursive computation MSIA
3 x 3	1	1	**1**
5 x 5	1.7	2.6	**1.2**
7 x 7	2.6	3.8	**1.6**
15 x 15	4.8	9.1	**1.6**
31 x 31	9.2	24.5	**1.6**

6 Conclusions

A computational model for spatiotemporal attention-guided analysis of videos is put forth. Our technique is based on the sequential detection of FoA points, identification of FPAs and matching of the FPA's descriptor sets with the reference ones. The proposed approach showed the following advantageous characteristics confirmed by the preliminary experiments reported herein: (1) localization of FoAs points inside object-relevant and homogeneous areas by selected properties; (2) local uniqueness of feature-point areas achieved through the isotropic definition of the multi-scale local contrast; (3) enhanced sensitivity in motion detection due to the temporal change determination using the area-time accumulated differentiation.

Acknowledgement. We gratefully acknowledge the financial support of the Ontario Centers of Excellence (OCE) and the National Sciences and Engineering Research Council of Canada (NSERC) towards the project “Big Data Analytics for the Maritime Internet of Things”.

References

1. Cristóbal, G., Perrinet, L., Keil, M.S. (eds.): Biologically Inspired Computer Vision: Fundamentals and Applications, 458 p. (2015)
2. Frintrop, S., et al.: Computational visual attention systems and their cognitive foundation: a survey. ACM Trans. Appl. Percept. **7**(1), 1–46 (2010)
3. Feichtenhofer, C., Pinz, A., Wildes, R.: Dynamic scene recognition with complementary spatiotemporal features. IEEE Trans. PAMI **38**(12), 2389–2401 (2016)
4. Bregonzio, M., Gong, S., Xiang, T.: Recognizing action as clouds of space-time interest points. In: Proceedings of the CVPR, pp. 1948–1955 (2009)
5. Felzenszwalb, P.F., et al.: Object detection with discriminatively trained part-based models. IEEE Trans. PAMI **32**(9), 1627–1645 (2010)
6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. PAMI **20**(11), 1254–1259 (1998)
7. Kadir, T., Brady, M.: Saliency, scale and image description. Int. J. Comput. Vis. **45**(2), 83–105 (2001)
8. Bruce, N.B., Tsotsos, J.K.: Saliency, attention, and visual search: an information theoretic approach. J. Vis. **9**(3), 1–24 (2009)
9. Itti, L., Baldi, P.: A principled approach to detecting surprising events in video. In: IEEE Conference Computer Vision and Pattern Recognition, pp. 631–637 (2005)
10. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in highly dynamic scenes. IEEE Trans. PAMI **32**(1), 171–177 (2010)
11. Adelson, E.H., Bergen, J.R.: Spatiotemporal energy models for the perception of motion. Opt. Soc. Am. **2**, 284–299 (1985)
12. Palenichka, R., et al.: A spatiotemporal attention operator using isotropic contrast and regional homogeneity. J. Electron. Imaging **20**(2), 1–15 (2011)
13. Shabani, A., Clausi, D., Zelek, J.S.: Evaluation of local spatiotemporal salient feature detectors for human action recognition. In: Proceedings of the CRV 2012, pp. 468–475 (2012)
14. Laptev, I.: On space-time interest points. Int. J. Comp. Vis. **64**(2/3), 107–123 (2005)
15. Harris, C., Stephens, M.J.: A combined corner and edge detector. In: Alvey Vision Conference, pp. 147–151 (1988)
16. Lindeberg, T.: Generalized Gaussian scale-space axiomatics comprising linear scale-space, affine scale-space and spatiotemporal scale-space. J. Math. Imaging Vis. **40**(1), 36–81 (2011)
17. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88688-4_48
18. Lindeberg, T.: Feature detection with automatic scale selection. IJCV **30**(2), 79–116 (1998)
19. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatiotemporal features. In: Proceedings of the VS-PETS, pp. 65–72 (2005)
20. Treisman, A., Gelade, G.: A feature integration theory of attention. Cogn. Psychol. **12**, 97–136 (1980)
21. Erhan, D., et al.: Scalable object detection using deep neural networks. In: Proceedings of the CVPR, pp. 2147–2154 (2014)
22. Curtis, P., Harb, M., Abielmona, R., Petriu, E.: Feature selection and neural network architecture evaluation for real-time video object classification. In: IEEE CEC, pp. 1038–1045 (2016)

23. Lindeberg, T.: Spatio-temporal scale selection in video data. J. Math. Imaging Vis., 1–38 (2017)
24. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs (1988)
25. McLachlan, G.J.: Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience, Hoboken (2004)
26. Palenichka, R., et al.: Model-based extraction of image area descriptors using a multi-scale attention operator. In: ICPR, Tokyo, pp. 853–856 (2012)