

Chapter 1

Bottom-up saliency models for videos: a practical review

1.1 Background

Research on visual saliency initially focused on still images rather than on video content. However, in the recent years, an increasing demand of video saliency appeared for some applications like gaming, editing, video retargeting, smart TV, robot navigation, surveillance, etc. Therefore, remarkable progress has been made first in the understanding on eye tracking data with dynamical stimuli and in a second time, in the modeling process.

There are fundamental differences between videos and still images. For example, each video frame is only observed during a fraction of a second, while a still image can be viewed much longer. Some videos can feature varying camera motion such as tilting, panning, zooming, etc. For this reason, videos are probably viewed differently by human observers than still images and some comprehensive comparative studies have emerged. In [1] for example, the authors study the influence of tasks on gaze behavior in static and dynamic scenes. In [2], the gaze on static and dynamic scene is compared; it is also shown that the center bias decreases with dynamic stimuli.

In terms of modeling, static models have first been extended to video. This is the case for GBVS, SDSR, NMPT or SSOI where authors added dynamic features to their models. Though these existing models are major contributions, video saliency estimation methods should then differ substantially from image saliency methods. Indeed, camera motions have great impacts on the

saliency estimation and models need to be specifically designed to manage the temporal aspect. This is the case for STVSM or SMQVA.

In this section, the video saliency models which will be used in the next chapters for saliency validation are described and discussed. In order to compare salient models for videos, four characteristics have been chosen and added into the descriptive sheets, following the color convention introduced by the colored keywords describing each characteristic below, for reader's convenience.

- The first characteristic such as for still images divides models based on their **approaches**. Some models have a **global** approach which is applied to the entire image while others compute a saliency map with a **local** approach which is applied to a picture area.
- The second one classifies models which use or not **prior** information. As an improvement, some models practice some top-down factors (**TD**) a 2D centered Gaussian bias, a face recognition algorithm or a segmentation at the end of the process.
- Third, the kind of **features** used to compute the saliency map classified the models. Indeed, some only use **static** features (colors, texture, etc.) while others compute **dynamical** features (motion, flicker, etc.). Some models can use **both** features.
- Finally, the last characteristic is similar to the last one for still images and shows if the **stimuli** are exploited either with all their channels (**color** images) or with just the **grayscale** information.

The eight saliency models for videos which are represented by their acronyms in Fig. 1.1 will be describe in the following of this section and use in the validation framework.

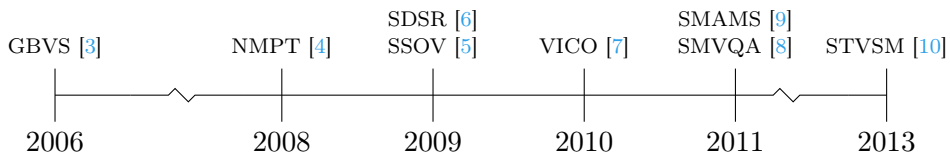


Figure 1.1. Chronological overview of salient models for videos.

GBVS: Graph-Based Visual Saliency (2006)

Characteristics: **local** | **HL** | **static** | color

Authors: J. Harel, C. Koch and P. Penora [3].



Figure 1.2. Illustration of the GBVS method: on the first row: 4 frames of a video sequence, on the second row: the corresponding saliency maps. Inspired by [3].

Description: This model uses an approach similar to the model having the same name [3] for static scenes to create feature maps at multiple spatial scales and propose a Graph-Based Visual Saliency model (GBVS). There are again three main steps but during the feature extraction step, motion and flicker channels can be added to compute the saliency maps of some video sequences. The algorithm then builds a fully connected graph over all grid locations of each feature map (intensity, orientation, color such as RGB or Derrington Krauskopf Lennie color space (DKL), motion and flicker). Weights are assigned between nodes that are inversely proportional to the similarity of feature values and their spatial distance. A center Gaussian is used to take advantage of the center bias and to improve the results.

NMPT: Nick's Machine Perception Toolbox (2008)

Characteristics: **local** | / | **static** | **color**

Authors: N. Butko, L. Zhang, G. Cottrell and J. Movellan [4].

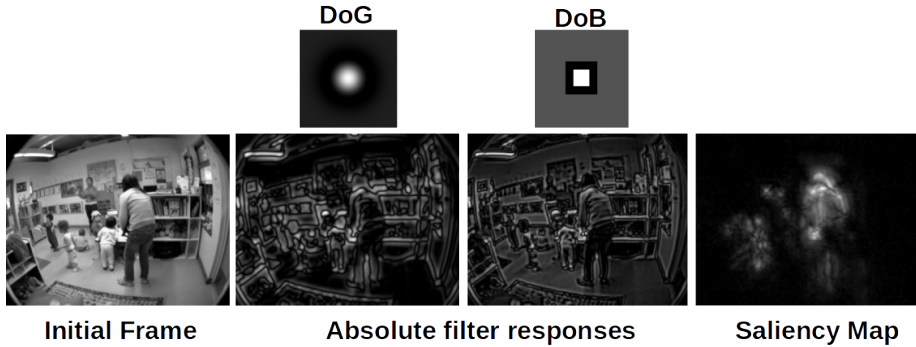


Figure 1.3. The NMPT model computes saliency map using spatio-temporal filters on grayscale frame (left). The filters and their outputs are shown for difference of Gaussians filter (second and third columns) and difference of Boxes approximation (fourth and fifth columns). Adapted from [4].

Description: This algorithm proposes a fast approximation to dynamic scenes of the visual saliency model for still images proposed in [11] and called SUN. It introduces spatio-temporal filters and fits a generalized Gaussian distribution to the estimated distribution for each filter response. Spatio-temporal filters can be tuned with different settings to use only spatial, use only temporal, use color contrast to be efficient and similar to the Human Visual System (HVS). The probability distributions of these spatio-temporal features were learned from a set of videos from natural environments. This model calculates its features and estimates the bottom-up saliency for each point.

SSOV: Segmenting Salient Objects for Videos (2009)

Characteristics: **local** | / | **static** | **color**

Authors: E. Rahtu, J. Kannala, M. Salo and J. Heikkilä [5].

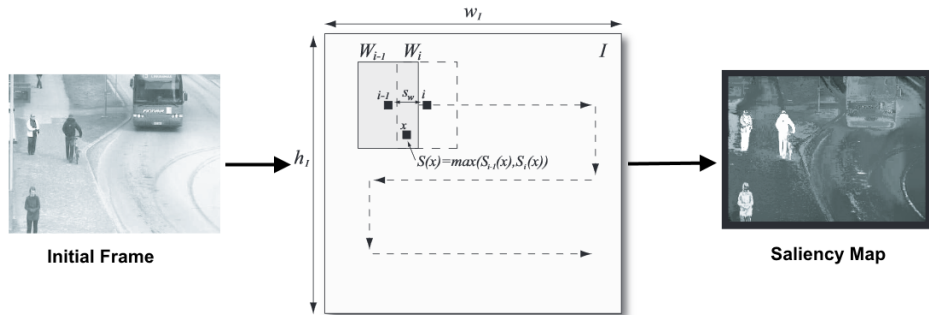


Figure 1.4. Illustration of the SSOV method: from left to right: initial frame, an example of the sliding window applied to compute the saliency values and saliency map. Adapted from [12].

Description: In order to adapt SSOI [5] for static scenes to video sequences, the CIE LAB perceptual color information of each frame is combined with the magnitude of the optical flow as input features at several scales. The optical flow was computed using an available online implementation [13]. The proposed saliency measure is formulated using a statistical framework and local feature contrast in motion, illumination and color information. The final salient segments were computed using the energy function in the Conditional Random Field (CRF) segmentation model for videos. The model is multiscale, does not require training but the weight between the color space and motion intensity components have to be defined manually.

SDSR: Saliency Detection by Self-Resemblance (2009)

Characteristics: **local** | / | **static** | grayscale

Authors: H. J. Seo and P. Milanfar [6].

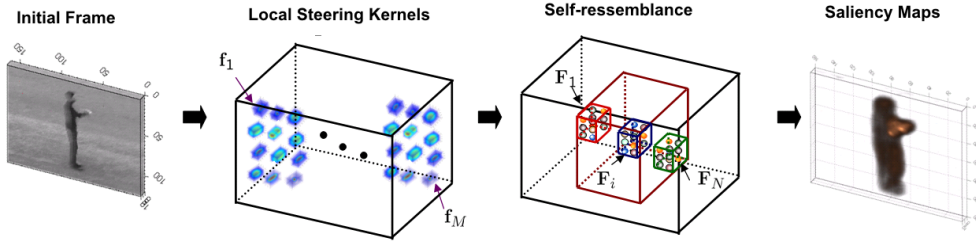


Figure 1.5. Illustration of the SDSR method: from left to right: the grayscale video, space-time local steering kernels to compute feature maps from a space-time neighborhood, the self resemblance algorithm and the final space-time saliency map. Adapted from [6].

Description: The SDSR model is an approach similar to the model having the same name [14]. It uses local regression kernels as features. Kernel density estimation that estimates the distribution of the features in a patch is then applied. In statistics, the kernel density estimation is a non-parametric way to estimate the probability density function of a random variable. The time dimension is added to the static model to obtain a 3D local steering kernel to manage the case of video sequences. This model has the advantage to be robust to noise and other systemic perturbation.

VICO: Visual Competitive attention model (2010)

Characteristics: **local** | / | **static** | **color**

Authors: M. Da Silva, V. Courboulay, A. Prigent and P. Estrailier [7] .

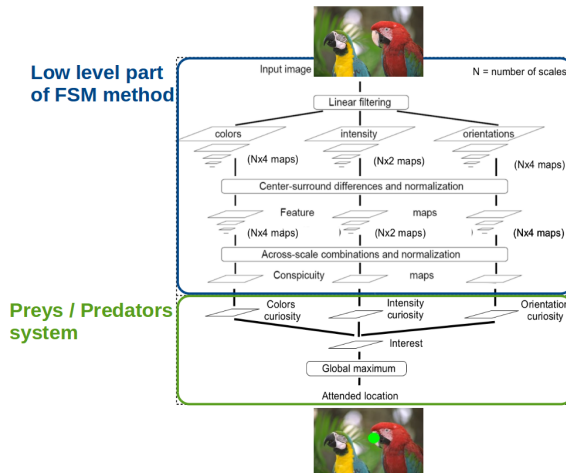


Figure 1.6. Illustration of the VICO model: from top to bottom: input image, low-level of the FSM method, preys-predators system and attention location. Adapted from [7].

Description: This approach proposes a new version of the FSM model [15] for static scenes. The second part of FSM classical fusion is replaced by using preys-predators systems to merge conspicuity maps. The results reveal that preys-predators systems can help modeling visual attention and allows fast maps generation while improving saliency maps accuracy. VICO simulated the scan-path of an observer across the frames of a video. Therefore, to obtain a density map at each frame, the model need to be run multiple times (corresponding of the number of viewers by database) on the same video.

SMVQA: Salient Motion for Video Quality Assessment (2011)

Characteristics: **global** | / | **dynamic** | grayscale

Authors: D. Culibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic and D. Kukolj [8].

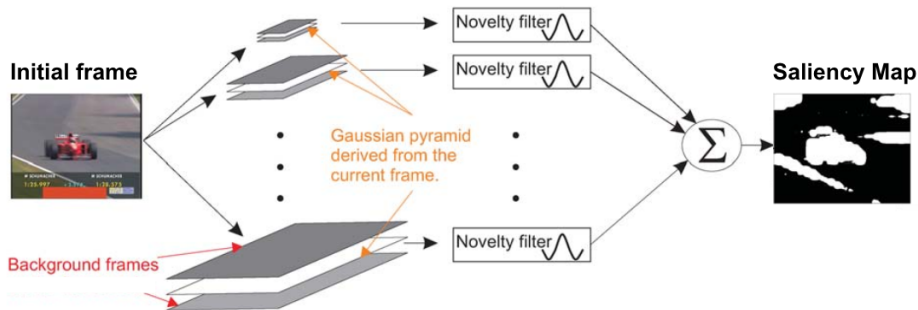


Figure 1.7. Illustration of the SMVQA model: from left to right: initial frame, Gaussian pyramids derived from the current frame, novelty filters, sum and saliency map. Adapted from [8].

Description: The SMVQA motion-based salient model has three main steps: first, it uses a multiscale Gaussian pyramid derived from the current frame and two background frames as described in [16]. Novelty temporal filters are then performed on each pyramid level to indicate the extent to which the current frame differs from the background frames. Finally, the single saliency map is obtained by summing the score of the pixels from the filter outputs at different scales and a modified Z-score test is used to detect the outliers in the frame. By efficiently managing the temporal information, this model detects cross-scale motion consistency, outlier and temporal coherence on each frame handles also videos with camera motion.

SMAMS: Saliency Models for Abnormal Motion Selection (2011)

Characteristics: **global** | / | **dynamic** | grayscale

Authors: M. Mancas, N. Riche, J. Leroy and B. Gosselin [9].

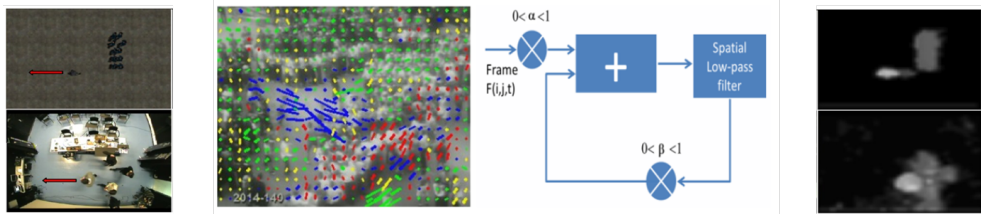


Figure 1.8. Illustration of the SMAMS model: from left to right: synthetic and real video frames, optical flow applied on a frame, schematization of the 3D low-pass filtering and the saliency maps for the corresponding input video frames. Adapted from [9].

Description: This algorithm proposes a model that detects abnormal motion. The SMAMS architecture has four main steps: first, motion features are extracted with an optical flow and output velocity and direction feature maps. Those two features are then spatio-temporally averaged with a 3D low-pass filter. The spatio-temporal averages separate each feature map into five bins at two different scales. Third, a self-information algorithm is computed for each map to highlight rare motion as salient. Indeed, the motion which is the most different in terms of speed and direction will have a higher saliency value as it is considered as abnormal. Finally, a fusion mechanism merge channels to give a single saliency map. As illustrated in the Fig. 1.8, some movements can be more salient than others. The model is effective for complex videos or dense crowds. Nevertheless, the model does not include any static cues as colors for example.

STVSM: Spatio-Temporal Visual Saliency Model (2013)

Characteristics: **local** | **HL** | **both** | grayscale

Authors: S. Marat, A. Rahman, D. Pellerin, N. Guyader and D. Houzet [10].

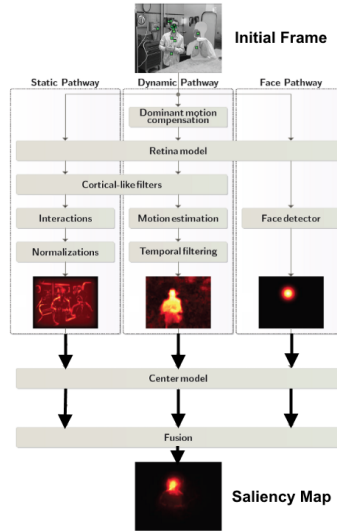


Figure 1.9. Illustration of the STVSM model: three pathways are computed from the grayscale input frame: from left to right: the static, the dynamic and the face ones. A 2D center gaussian is then applied on each one before merging them to build the final saliency map. Adapted from [10].

Description: The STVSM model [10] is inspired by the biology of the visual system, and breaks down each frame of a video into three maps: a *static* saliency map emphasizes regions that differ from their context in terms of luminance, orientation and spatial frequency. A *dynamic* saliency map emphasizes moving regions with values proportional to motion amplitude. A *face* saliency map emphasizes areas where a face is detected with a value. Finally, a 2D center gaussian is applied on each map and fuse all of them into a single saliency map.

ST-RARE: Spatio-Temporal multiscale rarity mechanism (2013)

Characteristics: **global** | / | **both** | **color**

Authors: M. Decombas, N. Riche, F. Dufaux, B. Pesquet-Popescu, M. Mancas, B. Gosselin and T. Dutoit *et al.* [17].

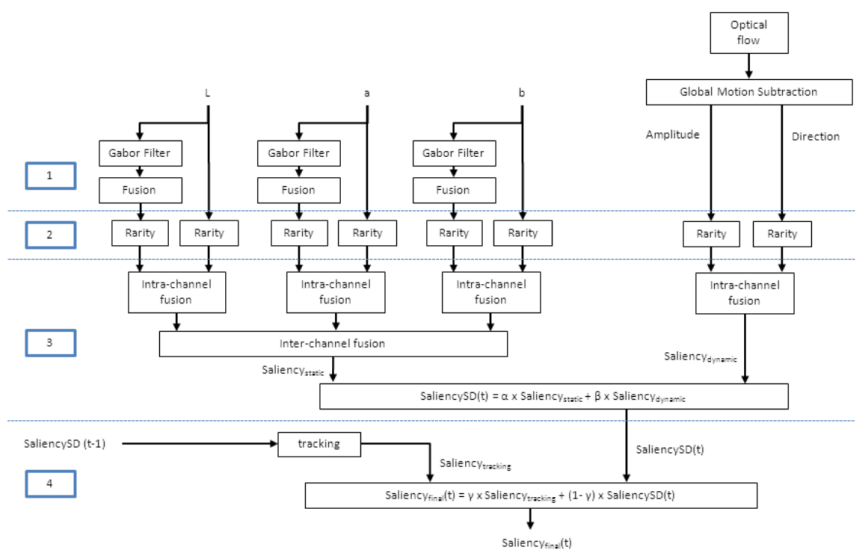


Figure 1.10. Overview of the ST-RARE saliency model. From top to down: (1) feature extraction, (2) multi-scale rarity mechanism, (3) fusion steps, and (4) tracking and temporal filtering (the static features are on the left while the dynamic features are on the right). Adapted from [17].

Description: The ST-RARE model combines spatial and temporal information to provide the map Saliency. First, six spatial feature maps: three low-level (which are the colours from the first path) and three medium-level (the orientation and texture information coming from the Gabor filters) and two temporal features maps: motion amplitude and direction are extracted from video frame. Then, a multiscale is used on each feature map and a fusion algorithm provides the saliency map. The last step is the temporal tracking framework in order to improve temporal coherence and robustness.

STRAP: Spatio-Temporal multiscale Rarity Algorithm with Priors (2013)

Characteristics: **global** | **HL** | **both** | **color**

Authors: M. Decombas, N. Riche, F. Dufaux, B. Pesquet-Popescu, M. Mancas, B. Gosselin and T. Dutoit [17].

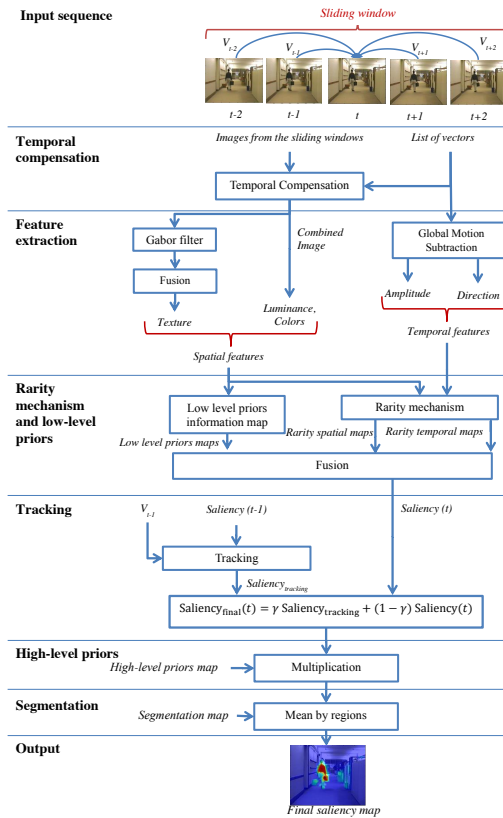


Figure 1.11. Overview of the STRAP saliency model. From top to down: (1) temporal compensation, (2) feature extraction, (3) multi-scale rarity mechanism and priors, (4) tracking, (5) high-level priors, (6) segmentation. Adapted from [17].

Description: STRAP is a new saliency model based on a spatio-temporal rarity mechanism and integrating priors information. It builds upon ST-RARE and includes several novel contributions: 1) a temporal motion compensation over a sliding window. In this way, neighboring frames can be jointly analyzed to increase temporal robustness, 2) color and frequency-based low-level priors are used together with the rarity mechanism and the fusion algorithm is optimized to this new architecture, 3) high-level priors like a centered Gaussian or face detection are then combined with the saliency results and 4) a spatio-temporal segmentation is finally used to improve the accuracy of the results and better detect the objects of interest.

1.2 Conclusion: a taxonomy of the algorithms

Saliency models for videos which will be used in the validations in the next chapters have been presented with descriptive sheets to provide readers with a global view of each model. However, as seen for still images in the previous chapter, this is not sufficient to classify dynamic models according to their structure. This is why some characteristics have been added into the descriptive sheets. Tab. 1.1 summarizes these 4 characteristics which have been chosen to compare the ten state-of-the-art saliency models for videos. It shows which of the 4 characteristics each model owns.

In order to provide an idea of pros and cons of each characteristic, some observations have to be conducted. The first characteristic for still images compares the local approach which detects clearly contrast in images against the global approach which highlights features which are different but not necessarily highly contrasted. The second characteristic classifies models which use or not top-down information. Saliency models can add some modules at the end of the process considered as top-down factors such as a 2D centered Gaussian, a face detector or a segmentation algorithm. The purpose is to better detect the salient areas and therefore to improve the scores. However, if these modules are inappropriately used, they will do just the opposite. It is important to correctly weight the 2D centered Gaussian, to adjust the parameters of the segmentation algorithm or to choose a face detectors with a lower false positive rate.

Table 1.1. Comparison of eight saliency models for videos on 7 characteristics.

	Approach	Prior	Feature	Stimuli
GBVS	local	TD	static	color
NMPT	local		static	color
SSOV	local		both	color
SDSR	local		static	gray
VICO	local		static	color
SMVQA	global		dynamic	gray
SMAMS	global		dynamic	gray
STVSM	local	TD	both	gray
ST-RARE	global		both	color
STRAP	global	TD	both	color

The third characteristic shows which kind of features are extracted to compute the saliency maps. Some models only use static features while others compute only dynamical features. Finally, some models can combine both kind of features. This last class of models is able to predict salient areas when there is or not motion in the videos while models with only static features can not detect motion and models with only dynamic features can not detection salient areas when there is no motion in the videos. Finally, the last characteristic shows how the stimuli are exploited: with color or grayscale informations. Although, most of psychophysical theories show the importance of color during the visual attentive process and the color information is used in many saliency models, its contribution for saliency modeling in videos was less clear. However, some studies such as [18] show the importance of color information which help to better predict fixations distribution in videos than models which only exploit the grayscale information.

To complete this comparison, the classical multidimensional scaling (MDS) technique similar to the one exposed in Chapter ?? has been chosen. The distances of this MDS are computed from the characteristics of Tab. 1.1. The

purpose is to have a better visualization of the level of similarity between saliency models for videos.

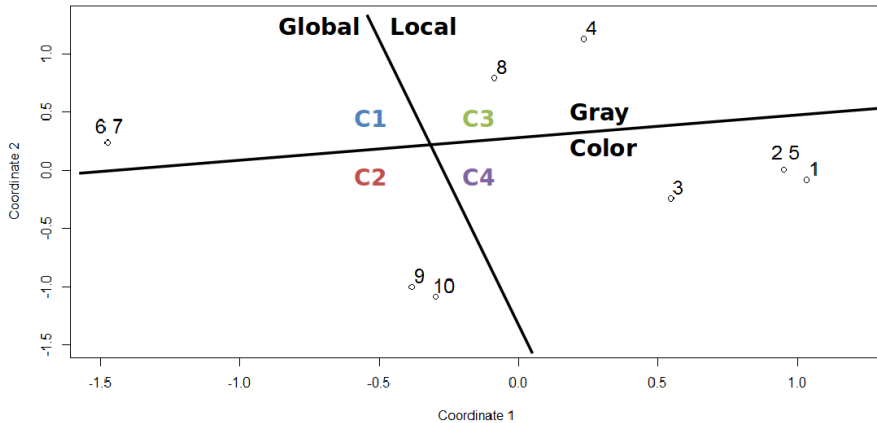


Figure 1.12. Multidimensional scaling of ten saliency models for videos based on characteristics in 2D: 1. GBVS / 2. NMPT / 3. SSOV / 4. SDSR / 5. VICO / 6. SMVQA / 7. SMAMS / 8. STVSM / 9. STRARE / 10. STRAP. The first coordinate substantially corresponds to the local/global class while the second substantially represents color/grayscale as input.

We can see from Fig. 1.12, a 2D MDS models representation based on video characteristics. The coordinates of this representation are components that represent a combination of characteristics. The first coordinate substantially corresponds to the first characteristic. Indeed, on one side (right) saliency models with local approach appear to have distances in the same range relatively to other models. On the other side (left) saliency models with global approach also seem to have distances in the same range. The second coordinate substantially corresponds to the last characteristic. Indeed, on one side (bottom) saliency models with color stimuli as input are very close while on the other side (top), saliency models with grayscale stimuli as input appear to have distances in the same range. These observations divide the presented models into 4 categories (from C1 to C4 on Fig. 1.12).

1.3 Summary

- 10 models for videos are described using descriptive sheets and will be used in the validation framework in the next chapters.
- Some models like GBVS, SSOV are extensions from 2D, while others are temporal models.
- In order to compare the models, different characteristics have been chosen and classified them into several classes.
- A list of dynamic state-of-the-art saliency models which are available online can be found from the Computational Attention Group of TCTS lab at <http://tcts.fpms.ac.be/attention>.

Bibliography

- [1] T. J. Smith and P. K. Mital, “Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes”, *Journal of Vision*, vol. 13, no. 8, p. 16, 2013.
- [2] T. V. Nguyen, M. Xu, G. Gao, M. Kankanhalli, Q. Tian, and S. Yan, “Static saliency vs. dynamic saliency: a comparative study”, in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 987–996.
- [3] C. K. J. Harel and P. Perona, “Graph-based visual saliency”, *Proceedings of Neural Information Processing Systems (NIPS)*, 2006.
- [4] N. J. Butko, L. Zhang, G. W. Cottrell, and J. R. Movellan, “Visual saliency model for robot cameras”, in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2008, pp. 2398–2403.
- [5] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, “Segmenting salient objects from images and videos”, in *The European Conference on Computer Vision (ECCV)*. Springer, 2010, pp. 366–379.
- [6] H. J. Seo and P. Milanfar, “Static and space-time visual saliency detection by self-resemblance”, *Journal of Vision*, vol. 9(12), no. 15, pp. pp. 1–27, 2009.
- [7] M. P. Da Silva, V. Courboulay, A. Prigent, and P. Estraillier, “Evaluation of preys/predators systems for visual attention simulation”, in *The International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2. INSTICC, 2010, pp. 275–282.

-
- [8] D. Culibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic, and D. Kukolj, “Salient motion features for video quality assessment”, *IEEE Transactions on Image Processing (TIP)*, vol. 20, no. 4, pp. 948–958, 2011.
- [9] M. Mancas, N. Riche, J. Leroy, and B. Gosselin, “Abnormal motion selection in crowds using bottom-up saliency”, in *International Conference on Image Processing (ICIP)*. IEEE, 2011, pp. 229–232.
- [10] S. Marat, A. Rahman, D. Pellerin, N. Guyader, and D. Houzet, “Improving visual saliency by adding ‘face feature map’ and ‘center bias’”, *Cognitive Computation*, vol. 5, no. 1, pp. 63–75, 2013.
- [11] 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*, vol. 8, no. 7, p. 32, 2008.
- [12] E. Rahtu and J. Heikkilä, “A simple and efficient saliency detector for background subtraction”, in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1137–1144.
- [13] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof, “Anisotropic huber-l1 optical flow.” in *Proceedings of the British Machine Vision Conference (BMVC)*, vol. 1, no. 2, 2009, p. 3.
- [14] H. J. Seo and P. Milanfar, “Nonparametric bottom-up saliency detection by self-resemblance”, *Conference on Computer Vision and Pattern Recognition (CVPR), 1st International Workshop on Visual Scene Understanding (ViSU)*, June 2009.
- [15] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis”, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [16] V. Crnojević, B. Antić *et al.*, “Multiscale background modelling and segmentation”, in *International Conference on Digital Signal Processing (DSP)*. IEEE, 2009, pp. 1–6.
- [17] M. Decombas, N. Riche, F. Dufaux, B. Pesquet-Popescu, M. Mancas, B. Gosselin, and T. Dutoit, “Spatio-temporal saliency based on rare model”, in *International Conference on Image Processing (ICIP)*. IEEE,

2013, pp. 3451–3455.

- [18] S. Hamel, N. Guyader, D. Pellerin, and D. Houzet, “Color information in a model of saliency”, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*. IEEE, 2014, pp. 226–230.