Fusion of Multiple Visual Cues for Visual Saliency Extraction from Wearable Camera Settings with Strong Motion

Hugo Boujut[1], Jenny Benois-Pineau[1], and Remi Megret[2]

[1] University of Bordeaux, LaBRI, UMR 5800, F-33400 Talence, France
{hugo.boujut,benois-p}@labri.fr
[2] University of Bordeaux, IMS, UMR 5218, F-33400 Talence, France
remi.megret@ims-bordeaux.fr

Abstract. In this paper we are interested in the saliency of visual content from wearable cameras. The subjective saliency in wearable video is studied first due to the psycho-visual experience on this content. Then the method for objective saliency map computation with a specific contribution based on geometrical saliency is proposed. Fusion of spatial, temporal and geometric cues in an objective saliency map is realized by the multiplicative operator. Resulting objective saliency maps are evaluated against the subjective maps with promising results, highlighting interesting performance of proposed geometric saliency model.

1 Introduction

Since recently, the focus of attention in video content understanding, presentation, and assessment has moved toward incorporating of visual saliency information to drive local analysis process. Hence in the paper from A. Benoit [1], the task is to classify animated movies from low motion content. The *global* saliency is therefore expressed by rhythm and other motion descriptors. In another task such as content viewing on a mobile screen, the most salient regions are selected according to the perceptual model of L. Itti and C. Koch [2]. If we simplify the concept of saliency to its very basic definition, we can reasonably say that visual saliency is what attracts human gaze. Numerous psycho-visual studies which have been conducted since the last quarter of 20th century uncovered some factors influencing it. Considering only signal features, the sensitivity to color contrasts, contours, orientation and motion observed in image plane has been stated by numerous authors [3,4]. Nevertheless, only these features are not sufficient to delimit the area in the image plane which is the strongest gaze attractor. In [5], the author states, for still images, that *observers show a marked tendency to fixate the center of the screen when viewing scenes on computer monitors*. The authors of [6] come to the same conclusion for dynamic general video content such as movies and Hollywood trailers. This is why the authors of [7] propose the third cue which is the geometrical saliency modelled by a 2D Gaussian located at the image center. While signal based cues remain valuable

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 436–445, 2012.

saliency indicators, we claim that geometrical saliency depends on global motion and camera settings in the dynamic scene. Nowadays, the attention of computer vision community is more and more turned to the new forms of video content: such as wearable video cameras, or "egocentric" view of the world [8,9]. Some attempts to identify visual saliency mainly on the basis of the frequency of repetition of visual objects and regions in the wearable video content have recently been made in [10]. We are specifically interested in building visual saliency maps by fusion of all cues in the pixel domain for the case of "egocentric" video content recorded with wearable cameras. Hence in this paper, we propose an automatic method of spatio-temporal saliency extraction for wearable camera videos with a specific accent on geometrical saliency dependent on strong wearable camera motion. We evaluate the proposed method with regard to subjective saliency maps obtained from gaze tracking.

The rest of the paper is organized as follows. In section 2 we will introduce the context and motivation of our work. In section 3 we report the method and our psycho-visual experiments for reference subjective saliency map construction. The objective saliency map will be presented in Section 4. The evaluation of the latter will be described in Section 5. Section 6 will conclude our work and outline its perspectives.

2 Motivations

The context of actual work is the multi-disciplinary research on Alzheimer disease [11,12]. The goal here is to ensure an objective assessment of the capacity of patients to conduct the IADL (Instrumental Activities of Daily Living). In [11] the framework for video monitoring with wearable camera was designed. The wearable camera is fixed in an ergonomic position on the patient's shoulder (see Fig. 1) and the recording is realized at patient's home. The computer vision task consists in an automatic recognition and indexing of IADLS from a taxonomy proposed to patients in each recording. We seek to limit the automatic analysis of the observed dynamic scene to the area of interest which is visually salient for the medical practitioner. Hence in the next section we study the subjective saliency of this type of content. Then we propose an automatic, objective saliency model from video data and assess it with regard to the built subjective saliency ground truth.

3 Subjective Saliency Map

In this section we present our approach for subjective saliency extraction on wearable video.

3.1 Eye-Tracker Experiment

The subjective saliency maps expressing user attention are obtained on the basis of psycho-visual experiment consisting in measuring the gaze positions on videos

Fig. 1. Wearable camera setup

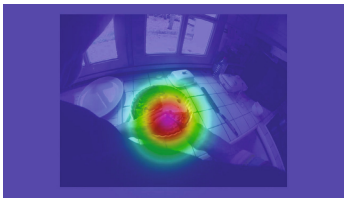

Fig. 2. Subjective saliency map example

from wearable video camera. The map of the visual attention has to be built on each frame of these videos. Videos from wearable camera differ from *traditional* video scenes: the camera films the user point of view, including his hands. Unlike *traditional* videos, wearable camera videos have a very high temporal activity due to the strong ego-motion of the wearer.

The gaze positions are recorded with an eye-tracker. We used HS-VET 250Hz from Cambridge Research Systems Ltd. This device is able to record 250 eye positions per second. The videos we display in this experiment have a frame-rate of 29.97 frames per second. A total of 28 videos filming the activities of daily living of patients and healthy volunteers are displayed to each participant of the experiment. This represents 17 minutes and 30 seconds of video. The resolution of the videos is 1280x960 pixels and the storage format is raw YUV 4:2:0. The experiment conditions and the experiment room is compliant to the recommendation ITU-R BT.500-11 [13]. Videos are displayed on a 23 inches LCD monitor with a native resolution of 1920x1080 pixels. To avoid image distortions, videos are not resized to screen resolution. A mid-gray frame is inserted around the displayed video. 25 participants were gathered for this experiment, 10 women and 15 men. For 5 participants some problems occurred in the eye-tracking recording process. So we decided to exclude these 5 records.

After looking at gaze position records on video frames, we stated that gaze anticipated camera motion and user actions. This phenomenon has been already reported by M. Land et al. in [14]. They state that *visual fixation does precede motor manipulation, putting eye movements in the vanguard of each motor act, rather than as adjuncts to it.*

Nevertheless, gaze positions cannot directly be applied as *ground truth* to compare automatic saliency model we aim at. They must be processed in order to get the subjective saliency map. The next section describes how to build subjective saliency maps from gaze tracking records.

3.2 Subjective Saliency Map Build Method

Any *objective* human visual perception model has to be validated and evaluated with regard to a *ground truth*. The *ground truth* is the *subjective* saliency in this case. The subjective saliency is built from eye position measurements. There are two reasons for which eye positions cannot be directly used to represent the

visual attention. First, the eye positions are only spots on the frame and do not represent the field of view. Secondly, to get accurate results, the saliency map is not built with the eye tracking data from one subject, but from many subjects. So the subjective saliency map should provide an information about the density of eye positions.

The method proposed by D. S. Wooding [15] fulfils these two constraints. In the case of video sequences, the method is applied on each frame I of a video sequence K. The process result is a subjective saliency map $S_{subj}(I)$ for each frame I. With this method, the saliency map is computed in three steps. In the first step, for each eye measure m of frame I, a two dimensional Gaussian is applied at the center of the eye measure $(x_0, y_0)_m$. The two dimensional Gaussian depicts the fovea projection on the screen. The fovea is the central retina part where the vision is the most accurate. In the *Sensibility to Light* [16] book chapter from D.C. Hood and M.A. Finkelstein (1986), the authors stated that the fovea covers an area from 1.5° to 2° in diameter at the retina center. D.S. Wooding proposed to set the Gaussian spread σ to an angle of 2°.

For the eye measure m of the frame I, a partial saliency map $S_{subj}(I, m)$ is computed (1).

$$S_{subj}(I, m) = Ae^{-\left(\frac{(x-x_{0_m})^2}{2\sigma_x^2}+\frac{(y-y_{0_m})^2}{2\sigma_y^2}\right)} \quad \text{with } \sigma_x = \sigma_y = \sigma \text{ and } A = 1 \tag{1}$$

Then, at the second step, all the partial saliency maps $S_{subj}(I, m)$ of frame $S_i I$ are added into $S_{subj}\prime(I)$ (2).

$$S_{subj}\prime(I) = \sum_{m=0}^{N_I} S_{subj}(I, m) \tag{2}$$

where N_I is the number of eye measures recorded on all the subjects for the frame I. Finally, at the third step, the saliency map $S_{subj}\prime(I)$ is normalized by the highest value $argmax$ of $S_{subj}\prime(I)$. The normalized subjective saliency map is stored in $S_{subj}(I)$. An example of a subjective saliency map is presented in Fig. 2. Subjective saliency maps cannot be used for real-world applications in video analysis, as requiring psycho-visual experiments for each video to process. We are thus interested in an automatic *objective* saliency maps. The subjective saliency maps will be used as the ground truth to asses the objective maps automatically built.

4 Objective Saliency Map

To delimit the area of video analysis in video frames to the regions which are potentially interesting to human observers we need to model visual saliency on the basis of video signal features. Here we follow the results of community research we reported in Section 1 proposing fusion of spatial, temporal and

geometric cues. We extend the state-of-the art approaches by a specific modelling of geometrical saliency and propose multiplicative fusion of all three cues.

4.1 Spatial Saliency Map

The spatial saliency map S_{sp} is mainly based on color contrasts [17]. We used the method from O. Brouard, V. Ricordel and D. Barba [7]. The spatial saliency map extraction is based on seven color contrast descriptors. These descriptors are computed in the HSI color space. On the contrary to RGB color system, the HSI color space is well suited to describe color interpretation by humans. The spatial saliency is defined according to the following seven local color contrasts V in the HSI domain : the *Contrast of Saturation*, *Contrast of Intensity*, *Contrast of Hue*, *Contrast of Opponents*, *Contrast of Warm and Cold Colors*, *Dominance of Warm Colors*, and *Dominance of Brightness and Saturation*.
The spatial saliency value $S'_{sp}(I,i)$ for pixel i from frame I is computed by mean fusion operator from seven color contrast descriptors (3) :

$$S'_{sp}(I,i) = \frac{1}{7}\sum_{\varsigma=1}^{7} V_{\varsigma}(I,i) \tag{3}$$

Finally, $S'_{sp}(I,i)$ is normalized between 0 and 1 to $S_{sp}(I,i)$according to its maximum value.

4.2 Temporal Saliency Map

The objective spatio-temporal saliency map model requires a temporal saliency dimension. This section will describe how to build temporal saliency maps. The temporal saliency map S_t models the attraction of attention to motion singularities in a scene. The visual attention is not grabbed by the motion itself. The gaze is attracted by the motion difference between the *real* motion scene and the global motion scene. The motion difference is called the residual motion. O. Brouard et al. [7] and S. Marat [18] propose a temporal saliency map model that takes advantage of the residual motion. In this paper, we have implemented the model from O. Brouard et al. [7].

The temporal saliency map is computed in three steps. The first one is the optical flow estimation. Then the global motion is estimated in order to get the residual motion. Finally a psycho-visual filter is applied on the residual motion.

To compute the optical flow, we have applied the Lucas Kanade method from OpenCV library [19]. The optical flow was sparsely computed on 4x4 blocks, as good results were reported in [20] when using 4×4 macro-block motion vectors from the H.264 AVC compressed stream. The next step in temporal saliency computation is the global motion estimation.

The goal here is to estimate a global motion model to differentiate then local motion from camera motion. In this work, we follow the preliminary study from [20] and use a complete first order affine model (4) :

$$\begin{aligned} dx_i &= a_1 + a_2 x + a_3 y \\ dy_i &= a_4 + a_5 x + a_6 y \end{aligned} \tag{4}$$

Here $\theta = (a_1, a_2, \ldots, a_6)^T$ is the parameter vector of the global model (4) and $(dx_i, dy_i)^T$ is the motion vector of a block. To estimate this model, we used robust least square estimator presented in [21]. We denote this motion vector $\boldsymbol{V}_\theta(I, i)$. Our goal is now to extract the local motion in video frames i.e. residual motion with regard to model (4). We denote the macro-block optical flow motion vector $\boldsymbol{V}_c(I, i)$. The residual motion $\boldsymbol{V}_r(I, i)$ is computed as a difference between block motion vectors and estimated global motion vectors.

Finally, the temporal saliency map $S_t(I, i)$ is computed by filtering the amount of residual motion in the frame. The authors of [7] reported, as established by S. Daly, that the human eye cannot follow objects with a velocity higher than $80°/s$ [22]. In this case, the saliency is null. S. Daly has also demonstrated that the saliency reaches its maximum with motion values between $6°/s$ and $30°/s$. According to this psycho-visual constraints, the filter proposed in [7] is given by (5).

$$S_t(s_i) = \begin{cases} \frac{1}{6}\boldsymbol{V}_r(I,i), & \text{if } 0 \leq \boldsymbol{V}_r(I,i) < \boldsymbol{v}_1 \\ 1, & \text{if } \boldsymbol{v}_1 \leq \boldsymbol{V}_r(I,i) < \boldsymbol{v}_2 \\ -\frac{1}{50}\boldsymbol{V}_r(I,i) + \frac{8}{5}, & \text{if } \boldsymbol{v}_2 \leq \boldsymbol{V}_r(I,i) < \boldsymbol{v}_{max} \\ 0, & \text{if } \boldsymbol{v}_{max} \leq \boldsymbol{V}_r(I,i) \end{cases} \tag{5}$$

with $\boldsymbol{v}_1 = 6°/s$, $\boldsymbol{v}_2 = 30°/s$ and $\boldsymbol{v}_{max} = 80°/s$. We follow this filtering scheme in temporal saliency map computation.

4.3 Geometric Saliency Map

As stated in the introduction, many studies have showed that the observers are attracted by the screen center. In [7], the geometrical saliency map is a 2D Gaussian located at the screen center with a spread $\sigma_x = \sigma_y = 5°$. In our psycho-visual experiments we stated that in a shoulder-fixed wearable camera video the gaze is always located in the first upper third of video frames, see the scattered plot of subjective saliency peaks in Fig. 3. Therefore, we have set the 2D Gaussian center at $x_0 = \frac{width}{2}$ and $y_0 = \frac{height}{3}$. The geometrical saliency S_g map equation is given by (6).

$$S_g(I) = e^{-\left(\frac{(x-x_0)^2}{2\sigma_x^2} + \frac{(y-y_0)^2}{2\sigma_y^2}\right)} \tag{6}$$

However, this attraction may change with the camera motion. This is explained by the anticipation phenomenon [14], see section 3.1. Hence we propose to simulate this phenomenon by moving the 2D Gaussian centred on initial "geometric saliency point" in the direction of the camera motion projected in the image plane. A rough approximation of this projection is the motion of image center computed with the global motion estimation model, equation (4), where $x = \frac{width}{2}$ and $y = \frac{height}{2}$.

4.4 Saliency Map Fusion

In the previous sections we explained how to compute spatial temporal and geometric saliency maps. In this section we describe the method that merges

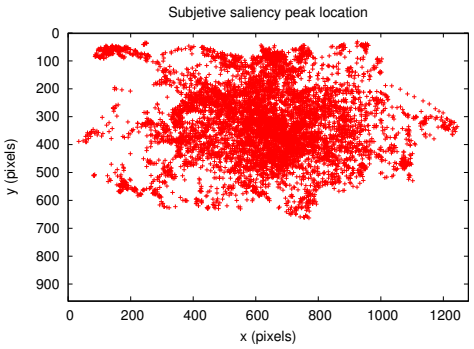

Fig. 3. Scattered plot of subjective saliency peaks for all database frames

these three saliency maps in the target *objective* saliency map. The fusion result is a spatio-temporal-geometric saliency map. In [20], several fusion methods for the spatio-temporal saliency without geometric component were proposed. We have tested these fusion methods on wearable video database. The results show that the multiplicative fusion performs the best. So for the full spatio-temporal-geometric saliency we compute multiplicative S^{mul}_{sp-t-g} (7).

$$S^{mul}_{sp-t-g}(I) = S_{sp}(I) \times S_t(I) \times S_g(I) \tag{7}$$

5 Evaluation

5.1 Normalized Scanpath Saliency

In this section, we compare the *objective* spatio-temporal saliency maps with *subjective* saliency map obtained from gaze tracking S_{subj}.

Here, we use the Normalized Scanpath Saliency (NSS) metric that was proposed in [18]. The NSS is a Z-Score that expresses the divergence of the subjective saliency maps from the objective saliency maps. The NSS computation for a frame I is depicted by (8). Here, S^N_{obj} denotes the objective saliency map S_{obj} normalized to have a zero mean and a unit standard deviation, $\bar{X}$ means an average. When $\overline{S_{subj} \times S^N_{obj}}$ is higher than the average objective saliency, the NSS is positive; it means that the gaze locations are inside the saliency depicted by the objective saliency map. In other words, higher the NSS is, more objective and subjective saliency maps are similar.

$$NSS = \frac{\overline{S_{subj} \times S^N_{obj}} - \overline{S_{obj}}}{\sigma(S_{obj})} \tag{8}$$

The NSS score for a video sequence is obtained by computing the average of NSS for all frames as in [18]. Then the overall NSS score on each video database is the average NSS of all video sequences. Results are presented in the next section.

5.2 Results

In this section, we compare the correlation of three automatic saliency maps with the subjective saliency. These three saliency maps are the spatio-temporal saliency map, the spatio-temporal-geometrical without camera motion, and the proposed method the spatio-temporal-geometrical with camera motion, expressing the anticipation phenomenon. The 28 video sequences described in Section 2 from wearable cameras are all characterized by strong camera motion which is up to 50 pixels magnitude in the center of frames. As it can be seen from the Fig. 4 the proposed method with moving of geometrical Gaussian almost systematically outperforms the base-line spatio-temporal saliency model and the spatio-temporal-geometrical saliency with a fixed Gaussian. For few sequences (e.g. number 2), the performance is poorer than obtained by geometric saliency with a fixed Gaussian. In these visual scenes, the distractors appear in the field of view. The resulting subjective saliency map then contains multiple maxima due to the unequal perception of scenes by the subjects. This is more "semantic saliency" phenomenon (faces, etc) which can not be handled with the proposed model. The average NSS on the whole database also shows the interest of proposed moving geometrical saliency. The mean NSS scores are respectively 1.832 for spatio-temporal, 2.607 for spatio-temporal with still geometrical Gaussian, and 2.791 with moving geometrical Gaussian. Which means 52.37% improvement of correspondence with subjective visual saliency map, which was our goal.

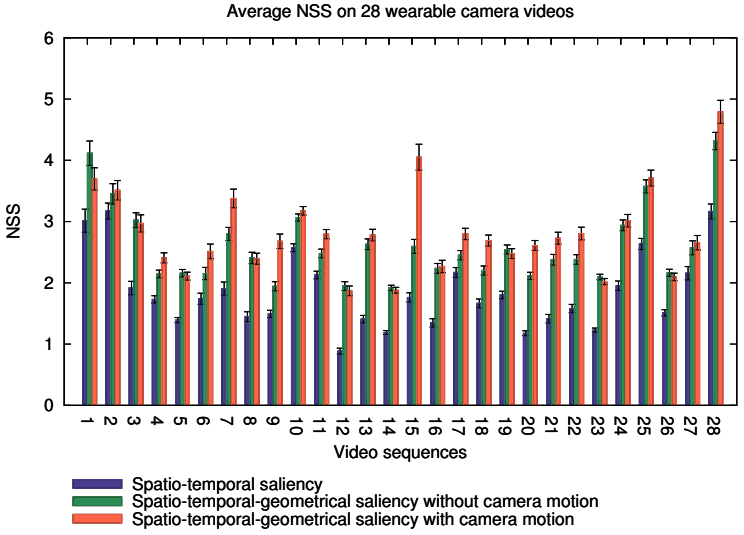

Fig. 4. Average NSS computed on 28 videos from a wearable camera

6 Conclusion

Hence in this work we proposed new objective visual saliency model and computation method by fusing spatial, temporal and geometric cues in video frames. The method was proposed for video with a strong motion recorded with cameras worn by subjects in the context of recording of Instrumental Activities of Daily Living. To our knowledge, this is the first attempt to define saliency in such a content taking into account psycho-visual models known for other types of traditional video content. First of all, conducting psycho-visual experiment on a representative set of test subjects, we stated the anticipation phenomenon in gaze positioning which was obviously transmitted to the subjective saliency map built according to Wooding method. While in previous research, the geometrical saliency was systematically frame centred, we recorded experimental evidence of the dependence of the geometric saliency of camera fixation on a body. Finally, in order to incorporate the anticipation phenomenon into the automatic construction of objective saliency map we expressed it by moving geometric saliency Gaussian in direction of camera motion projected into the image plane. These results are encouraging. In some video sequences moving of geometric Gaussian allows to improve the NSS up to 40% compared to fixed Gaussian and up-to 50% compared to base-line spatio-temporal psycho-visual saliency model. In the future of this work we will work on incorporating distractors and on saliency -based feature weighting in the problem of scene recognition.

Acknowledgments. This research is supported by the EU FP7 PI Dem@Care project #288199.

References

1. Ionescu, B., Vertan, C., Lambert, P., Benoit, A.: A color-action perceptual approach to the classification of animated movies. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR 2011, pp. 10:1–10:8. ACM, New York (2011)
2. Itti, L., Koch, C.: Computational modelling of visual attention. Nature Review Neuroscience 2, 194–203 (2001)
3. Itti, L.: Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. Visual Cognition 12, 1093–1123 (2005)
4. Le Meur, O., Le Callet, P., Barba, D.: Predicting visual fixations on video based on low-level video features. Vision Research 47, 1057–1092 (2007)
5. Tatler, B.W.: The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. Journal of Vision 7, 1–17 (2007)
6. Dorr, M., Martinetz, T., Gegenfurtner, K.R., Barth, E.: Variability of eye movements when viewing dynamic natural scenes. Journal of Vision 10 (2010)
7. Brouard, O., Ricordel, V., Barba, D.: Cartes de Saillance Spatio-Temporelle basées Contrastes de Couleur et Mouvement Relatif. In: Compression et Representation des Signaux Audiovisuels, CORESA 2009, Toulouse, France, 6 pages (2009)

8. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012, pp. 1–8 (2012)
9. Starner, T., Schiele, B., Pentland, A.: Visual contextual awareness in wearable computing. In: ISWC, pp. 50–57 (1998)
10. Ren, X., Philipose, M.: Egocentric recognition of handled objects: Benchmark and analysis. In: Computer Vision and Pattern Recognition Workshop, pp. 1–8 (2009)
11. Karaman, S., Benois-Pineau, J., Mégret, R., Dovgalecs, V., Dartigues, J.F., Gaëstel, Y.: Human Daily Activities Indexing in Videos from Wearable Cameras for Monitoring of Patients with Dementia Diseases. In: ICPR 2010, Istanbul, Turquie, pp. 4113–4116 (2010) ANR-09-BLAN-0165-02
12. Szolgay, D., Benois-Pineau, J., Mégret, R., Gaëstel, Y., Dartigues, J.F.: Detection of moving foreground objects in videos with strong camera motion. Pattern Analysis and Applications 14, 311–328 (2011)
13. International Telecommunication Union: Methodology for the subjective assessment of the quality of television pictures. Recommendation BT.500-11, International Telecommunication Union (2002)
14. Land, M., Mennie, N., Rusted, J.: The roles of vision and eye movements in the control of activities of daily living. Perception 28, 1311–1328 (1999)
15. Wooding, D.: Eye movements of large populations: Ii. Deriving regions of interest, coverage, and similarity using fixation maps. Behavior Research Methods 34, 518–528 (2002), doi:10.3758/BF03195481
16. Hood, D.C., Finkelstein, M.A.: Sensitivity to light. In: Boff, K.R., Kaufman, L., Thomas, J.P. (eds.) Handbook of Perception and Human Performance. Sensory processes and perception, vol. 1, pp. 5-1–5-66. John Wiley & Sons, New York (1986)
17. Aziz, M., Mertsching, B.: Fast and robust generation of feature maps for region-based visual attention. IEEE Transactions on Image Processing 17, 633–644 (2008)
18. Marat, S., Ho Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., Guérin-Dugué, A.: Modelling spatio-temporal saliency to predict gaze direction for short videos. International Journal of Computer Vision 82, 231–243 (2009), Département Images et Signal
19. Bouguet, J.Y.: Pyramidal implementation of the lucas kanade feature tracker. Intel Corporation, Microprocessor Research Labs (2000)
20. Boujut, H., Benois-Pineau, J., Ahmed, T., Hadar, O., Bonnet, P.: A metric for no-reference video quality assessment for hd tv delivery based on saliency maps. In: 2011 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–5 (2011)
21. Kraemer, P., Benois-Pineau, J., Domenger, J.P.: Scene Similarity Measure for Video Content Segmentation in the Framework of Rough Indexing Paradigm, Espagne, pp. 141–155 (2004)
22. Daly, S.J.: Engineering observations from spatiovelocity and spatiotemporal visual models. In: IS&T/SPIE Conference on Human Vision and Electronic Imaging III, vol. 3299, pp. 180–191 (1998)