Context-Free Detection of Events

Benedikt Kaiser[1] and Gunther Heidemann[2]

[1] University of Karlsruhe, Institute for Process Control and Robotics
Building 40.28, Kaiserstr. 12, D-76128 Karlsruhe, Germany
[2] University of Stuttgart, Intelligent Systems Group
Universitätsstr. 38, D-70569 Stuttgart, Germany

Abstract. The detection of basic events such as turning points in object trajectories is an important low-level task of image sequence analysis. We propose extending the SUSAN algorithm to the spatio-temporal domain for a context-free detection of salient events, which can be used as a starting point for further motion analysis. While in the static 2D-case SUSAN returns a map indicating edges and corners, we obtain in a straight forward extension of SUSAN a 2D+1D saliency map indicating edges and corners in both space and time. Since the mixture of spatial and temporal structures is still unsatisfying, we propose a modification better suited for event analysis.

1 Introduction

For the analysis of static images, the detection of regions of interest or points of interest (representing a region) is an important technique to direct the focus of attention. Thus the most relevant patches for further processing can be found. Such methods serve two purposes: Making computations more efficient, and pre-selecting relevant patterns. That is, even the close-to-signal algorithms are actually part of the pattern classification. Therefore, multiple cues such as colour and texture are in use (e.g. [1]).

The benefit of attentional techniques such as segmentation and interest point (IP) detection is that they are free of context [2], i.e. not adapted to a particular domain. Thus, they serve as a purely data driven starting point for the processing cycle. But in spite of the success in the static case, image sequence analysis rarely makes use of attentional methods, at least not of such that really process the spatio-temporal data. In other words, attention is directed to regions based on isolated frames. While for segmentation there are some approaches (e.g. [3]) that exploit the 2D+1D image data (time being the additional dimension), in the field of IP-detection so far only the Harris-detector [4] has been extended to spatio-temporal data [5].

This is astonishing, since for the decomposition of static scenes into meaningful components, IPs are a standard technique to filter out and represent areas which appear relevant at the signal level. Applications are image retrieval [6], active vision [7], object recognition [8], or image compression [9]. In the present

B.K. Ersbøll and K.S. Pedersen (Eds.): SCIA 2007, LNCS 4522, pp. 223–232, 2007.

paper, we will therefore transfer the concept to the spatio-temporal domain for the detection of baisc actions and events. In comparison to the common technique which is computation of the optical flow, IP-detection is less computationally costly. Certainly, it will not be possible to cover the entire complexity of natural actions by IP-detection, but in the same way as IPs offer cues for static patterns such as symmetry [2,8], basic events like turning points, acceleration, rotation, or approach of two objects (a closing gap) can be detected.

For static imagery, IPs are points which are "salient" or "distinctive" as compared to their neighbourhood. To be more precise, an IP is not necessarily an isolated salient pixel, rather, the IP at a pixel location which stands for a salient patch of the image. Most algorithms for IP detection are aimed at the detection of corners or edges in grey value images [10,4,11,12,13,14]. Methods of this kind which detect edges or corners are particularly promising for the spatio-temporal case, since 3D-corners may indicate the turning points of 2D-corners in a temporal sequence. Thus they would indicate saliency both in the geometrical sense and in the sense of an event.

Laptev and Lindeberg [5] have shown how the concept of spatial edge- and corner-based IPs can be extended to the spatio-temporal domain for the HARRIS detector [4]. They extend the detection of IPs from the eigenvalues of the 2D autocorrelation matrix of the signal to the 3D matrix in a straight forward approach, in addition, they propose a scale space approach to deal with different spatial and temporal scales. But since the 2D- and 3D-HARRIS detector depends on the often problematic computation of grey value derivatives, we chose to extend another IP-detector to the spatio-temporal domain: The SUSAN detector proposed by Smith and Brady [12], which detects edges and corners merely from the grey values.

We will first describe the spatial SUSAN detector (section 2), then its extension to the spatio-temporal domain is described in section 3. The new 3D-SUSAN detector is tested in section 4 using artificial image sequences displaying prototypical events. But since the tests uncover shortcomings of the straight forward extension from 2D to 3D, a modification is introduced in section 5. Finally, the new approach is tested on real image sequences.

2 The SUSAN-Detector for Static Images

Smith and Brady have proposed an approach to detect edges and corners, i.e., one- and two-dimensional image features [12]. While most algorithms of this kind rely on the (first) derivatives of the image matrix, the SUSAN-detector relies on the local binarisation of grey values. To compute the edge- or corner strength of a pixel (called the "nucleus"), a circular mask A around the pixel is considered. By choice of a brightness difference threshold ϑ, an area within the mask is selected which consists of pixels similar in brightness to the nucleus. This area is called USAN ("Univalue Segment Assimilating Nucleus"). To be more precise,

let $I(r)$ denote the grey value at pixel r, n the area (i.e. # pixels) of the USAN, and r_0 the nucleus. Then

$$n(r_0) = \sum_{r \in A} c(r, r_0), \qquad \text{with} \quad c(r, r_0) = \begin{cases} 1 \text{ for } |I(r) - I(r_0)| \leq \vartheta \\ 0 \text{ for } |I(r) - I(r_0)| > \vartheta \,. \end{cases} \tag{1}$$

The response of the SUSAN-detector at pixel r_0 is given by

$$R(r_0) = \begin{cases} g - n(r_0) \text{ for } n(r_0) < g \\ 0 \qquad\qquad \text{else} \,, \end{cases} \tag{2}$$

where g is called the *geometric threshold*. For edge detection, a suitable value is $g = \frac{3}{4} n_{max}$, for corner detection $g = \frac{1}{2} n_{max}$. It can be shown the these values are optimal in certain aspects for the assumption of a particular signal-to-noise ratio.

Smith and Brady [12] obtain a saliency map which indicates edge and corner strength as the inverted USAN area for each nucleus pixel. IPs are then found as the local minima of the USAN area, thus the name SUSAN (= Smallest Univalue Segment Assimilating Nucleus).

To find the local direction of an edge and to localize corners precisely, geometrical features of the USAN have to be exploited, see [12] for details.

The SUSAN-approach is well suited for a fast detection of one- and two-dimensional basic image features with the benefit that both localization precision and the implicitly built-in noise reduction are robust to changes of the size of the circular mask.

3 Spatio-temporal Extension of SUSAN

In this section we introduce the extension of the normal 2D-SUSAN-detector to the spatio-temporal ("3D") domain [15].

The generalization of an isotropic circular mask in two dimensions is a sphere in three dimensions. But since the spatial coordinates are independent of time, a (rotationally symmetric) ellipsoid around the time axis is betters suited for event detection since it allows suitable scaling (note the same physical event may, e.g., be captured using different frame rates). However, also other 3D-shapes with a circular cross section come into question. In the following, two algorithms using different 3D-masks M_E and M_Z are investigated:

$$M_E(x, y, t) = \begin{cases} 1 \text{ if } \frac{x^2+y^2}{R_{xy}^2} + \frac{t^2}{R_t^2} < 1 \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

$$M_Z(x, y, t) = \begin{cases} 1 \text{ if } \frac{x^2+y^2}{R_{xy}^2} \ \wedge \ -R_t \leq t \leq R_t \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

where R_{xy} denotes the radius in the x-y-plain, and R_t the extension of the mask in on the temporal $t-$axis. In the same way as the 2D-SUSAN-detector is applied to

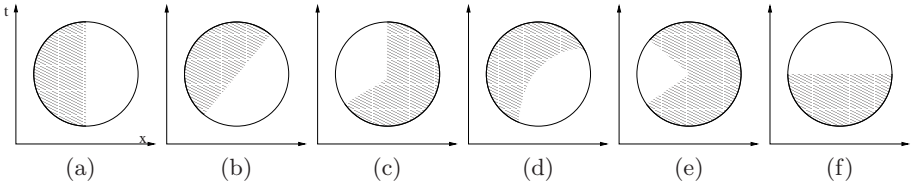


Fig. 1. A selection of different kinds of motion. The USANs are shown in the $x-t$-plain: (a) Rest, (b) constant velocity, (c) stop-event, (d) acceleration, (e) turn-around, (f) sudden appearance.

each spatial point, now the mask must be applied to each spatio-temporal point of an image sequence, and the grey value of the nucleus is compared to the other pixels within the mask to obtain the USAN-volume. Instead of the original binary decision function, we use the improved version as proposed by Smith and Brady:

$$c(r, r_0) = e^{-(\frac{I(r)-I(r_0)}{t})^6} \tag{5}$$

By this means, the robustness of the algorithm is improved since now slight variations of luminance may not lead to a large variation of the output. The USAN-volume can be calculated as

$$V(x, y, z) = \sum_{x',y',z'} I(x', y', z')M(x + x', y + y', z + z'), \tag{6}$$

Fig. 1 illustrates the way SUSAN3D processes different motion events (in only one spatial dimension). The x-axis points from left to right, the t-axis upwards. The USAN is the white area within the mask. While Fig. 1(a) shows an edge element at rest, Fig. 1(b) depicts motion at constant velocity. The result of $V = 0.5$ (of the area) is the same in both cases. Figs. 1(c) and 1(d) depict a stop-event and acceleration, respectively. For these cases, values V clearly below 0.5 are to be expected. The smallest value V is to be expected in the case depicted in Fig. 1e, which shows a turning point. Fig. 1(f) shows either a sudden appearance of the object or motion at a velocity to high to be resolved. Again, the volume is $V = 0.5$.

Summarizing, by evaluating the 3D-USAN values in the manner of the conventional 2D-SUSAN-detector "salient" events can be detected, such as acceleration and turn around (values the smaller the stronger curvature). However, rest, constant motion, and sudden appearance (all three $V = 0.5$) can not be discriminated. While this is still satisfactory for rest and constant motion ($V = 0.5$ being larger than for acceleration and turn-around), sudden appearance should get the smallest value (i.e. the largest saliency output).

4 Evaluation of the Naive Spatio-temporal Algorithm

In the following, we test the SUSAN3D-detector on artificial image sequences, using as a mask a cylinder of $2R_{xy} = 7$ pixels and $2R_t = 7$ frames, yielding a total volume of 224. The brightness threshold is set to 27 as in the 2D-version.

4.1 Simulations

The first test sequence shows squares moving in different ways. Fig. 2(a) shows, from top to bottom, the SUSAN3D response maps for an accelerating square, a moving one with high constant velocity, a moving one with low constant velocity, and a square at rest. Obviously, the response of the SUSAN3D is mainly governed by geometrical features, not by dynamical features. To reduce the influence of geometrical features, the response of SUSAN3D was tested in 2(b)-(h) by a moving filled circle, i.e. an object without any corners. Fig. 2(b) is the resulting spatio-temporal map of a circle at rest. Figs. 2(c)-(e) show the results for a circle moving to the right at a constant velocity of one, two, and three pixels per frame. In 2 (f), the circle is accelerating by a constant acceleration of $a = 2\,\text{pixels/frame}^2$, in 2 (g), acceleration grows exponentially Fig. 2 (h) shows a turn-around.

Now we searched for the minimum of the USAN. To compare the USAN values achieved at a certain spot of the moving circle, in a first test we searched for the minimum not within the entire response map but only in the area of the righthand circle border on a horizontal line through the middle of the map. The rest of the map was discarded. Results are listed in table 1. Remarkably, the minimal USAN-value is always 112 — except for the turn around — which is half of the mask volume (first line of table 1). So, the expectation that acceleration expresses itself in lower USAN values did not come true in this experiment. For a better analysis of this result, the seven circular "slices" of the mask cylinder have been analysed in separation in table 1 lines "-3" ... "3" ("0" corresponds to the central slice of the mask cylinder). Obviously, the contributions of the partial volumina are different. While the distribution differs for columns $v = 0, 1, 2$ and thus reflects the fact that the object moves at a different velocity, columns $v = 3, a = 2$ and *exp* do not exhibit any difference, because velocity is too large

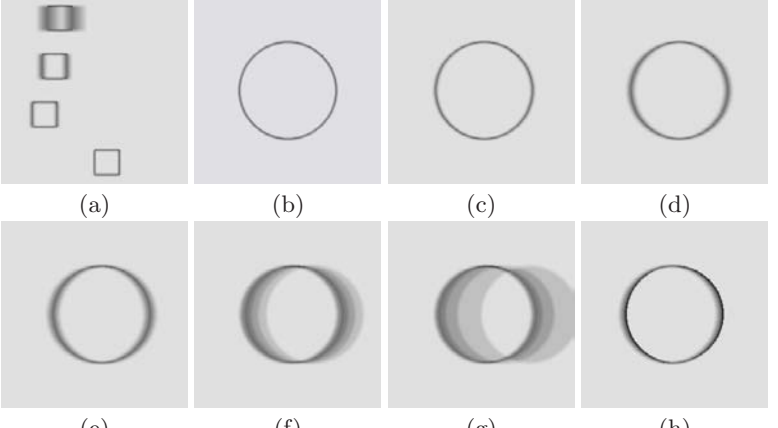


Fig. 2. Output of SUSAN3D at a single moment for different image sequences, see text

Table 1. Results of the first experiment with SUSAN3D, see Fig. 2

	$v=0$	$v=1$	$v=2$	$v=3$	$a=2$	exp	turn
USAN	112	112	112	112	112	112	36
-3	16	0	0	0	0	0	0
-2	16	4	0	0	0	0	0
-1	16	10	4	0	0	0	10
0	16	16	16	16	16	16	16
1	16	22	28	32	32	32	10
2	16	28	32	32	32	32	0
3	16	32	32	32	32	32	0

for the chosen mask size. I.e., the object is so fast that each of the three motions is equivalent to an "appearance out of nowhere" for the detector (cf. Fig. 1(f)). In column "turn", the turn around becomes visible both in a small USAN-value and in the distribution throughout the mask cylinder (cf. 1(e)).

4.2 Discussion on the Simulations

There is still a flaw in the "naive" extension of SUSAN: Geometrical and dynamical features are coupled implicitly. Therefore, an accelerating straight edge leads to smaller output in the sequence of saliency maps than a stationary corner. For event detection, however, the first case is more relevant, so the influence of the geometrical features should be attenuated. Fig. 3 illustrates that this is a non-trivial problem: In both cases, the corners move at a constant velocity from the left to the right, the only difference being the rotation of the object. The circles are the slices of the mask cylinder corresponding to successive frames of the sequence, which exhibit different intersections with the corners. In total, case Fig. 3(a) leads to a smaller USAN-volume than case Fig. 3(b), though, regarded in isolation, both the geometrical and the dynamical features are equal for both types of corners.

The contributions of the single circular slices of the mask cylinder first increase, then decrease in Fig. 3(a), whereas they continually increase for 3(b). Classification according to Fig. 1 yields "turning point" for Fig. 3(a) but "constant velocity" for Fig. 3(b). Thus, application of the SUSAN3D-detector is not feasible since it mixes geometrical and dynamical features.

Further, the size of the mask is difficult to choose: While it should be sufficiently small to overlap only corners or edges but no larger structures, it should be large enough to realize a reasonable resolution for the analysis of different velocities and accelerations. These opposing requirements refer both to the temporal and spatial dimension. In principle, the same problem exists for the spatial SUSAN-detector and in general for any windowed function, but it becomes more difficult in the spatio-temporal domain. While in the spatial domain a given window size simply selects a certain scale, in the spatio-temporal domain the coupling between typical spatial and temporal scales has to be dealt with.

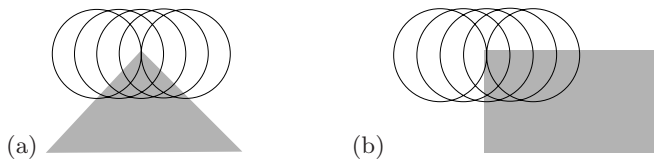


Fig. 3. A corner with constant velocity moving from left to right at different angles (a, b). The cylindrical mask of the object is indicated by its temporal "slices".

5 The SUSANinTime Detector

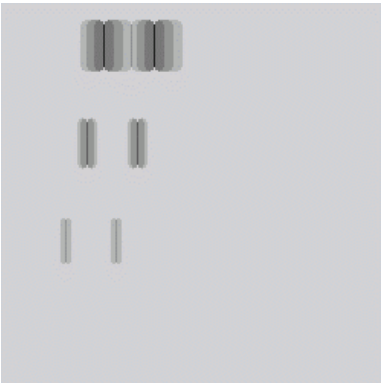
The discussion of the last section has shown that a straight forward extension of the SUSAN-detector to three dimensions is not satisfying. Of course, additional features could be computed to correct the detector response, but then the simple and elegant idea of using the USAN-volume as a feature for IP-detection would be more or less abandoned. Therefore, in the following we will outline an alternative approach, which applies the SUSAN principle only on the temporal dimension.

The first step is computation of the USAN-area within a cylindrical volume around the nucleus. The single x-y-slices of the cylindrical mask are evaluated to find the USAN-areas for every frame, these values are saved in a 1D-array (`areas[]`). Then the SUSAN principle is applied to the 1D-array `areas[]` in the following way: The USAN-area at the current time is considered to be a (second) nucleus value (`nucleus2`). Note the second nucleus value is an *area*, not a grey value. Then the array `areas[]` is binarised with respect to the `nucleus2` value, and the final detector response is the sum of the now binarised array.

In the pseudocode given in Fig. 4, `mask[x][y][t]` takes a value of 1 if `x,y,t` is inside the volume covered by the detector, else 0. c_1 and c_2 denote the thresholding functions for the spatial binarisation of the x-y-slices and the binarisation of the `areas[]` array, respectively.

```
SUSANinTime(x, y, t)
    nucleus <- getpixel(x,y,t)
    FOR tt FROM -R TO R DO
        areas[tt] <- 0
        FOR yy FROM -r TO r DO
            FOR xx FROM -r TO r DO
                IF mask[x][y][t] = 1 THEN
                    pixel <- getpixel(x + xx, y + yy, t + tt);
                    areas[tt] <- areas[tt] + c_1(nucleus, pixel)
    nucleus2 <- areas[0]
    value <- 0
    FOR tt FROM -R TO R DO
        value <- value + c_2(nucleus2, areas[tt])
    RETURN value
```

Fig. 4. Pseudocode of SUSANinTime, see text



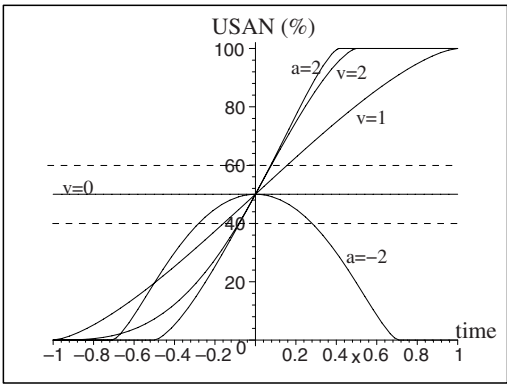


Fig. 5. Left: Output of the SUSANinTime-detector to the sequences of moving squares used in Fig. 2. Right: USAN-area as a function of time.

The idea of the SUSANinTime algorithm is to give a high response to such space-time volumes which exhibit a high activity, where "activity" is defined as a high temporal variation of the USAN-area. Fig. 5 illustrates the principle. It shows the USAN-area (as the percentage of a complete circular slice) as a function of time (to be more precise, as a function of the t-coordinate of the cylindrical mask). The nucleus value is 50% for all of the motion sequences. The SUSANinTime-detector computes for which span of time the USAN-areas are still within a surroundings the nucleus value. This time span takes a maximum for zero velocity ($v = 0$) and decreases with increasing velocity ($v = 1, v = 2$). Thus, it becomes also clear that the return value of SUSANinTime does not allow measurement of acceleration ($a = 2, a = -2$). Though the SUSANinTime-algorithm can not classify space-time events in categories velocity / acceleration, it has nevertheless highly useful properties. Fig. 5, left, shows the response of the SUSANinTime detector, where the input sequence is the one of Fig. 2(a). The detector response is approximately proportional to the velocity, for stationary regions, the detector is "blind" (here, small intensity values denote a high response). In contrast to Fig. 2(a), the corners of the squares yield no stronger response than the edges, though being geometrically more salient. So the response is determined by the dynamics, not stationary features — a major advantage, since now geometrical features detected by separate modules can be included in a final saliency map in a well-defined way.

First experiments on real world image sequences have shown that the algorithm yields robust results. Fig. 6 shows the output of the SUSANinTime detector for a sequence. At first, the hand accelerates in a movement to the right, then stops. Different intensities of the map reflect the different velocities. Note that the appearance or disappearance of (otherwise) static structures in the background is likewise detected as motion.

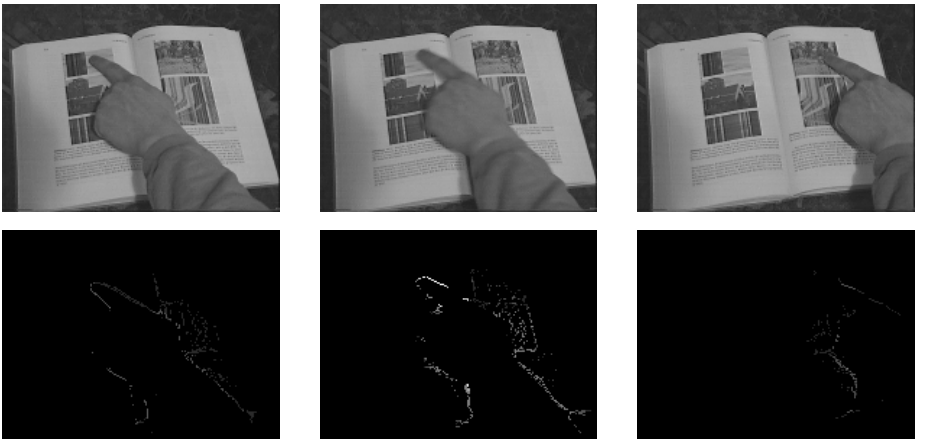


Fig. 6. Output of the SUSANinTime detector for a real image sequence. The pointing event can be clearly detected.

6 Summary and Conclusion

For the detection of generic events in image sequences such as turning points, we have introduced two extensions of the SUSAN IP-detector: A naive extension of SUSAN to a third dimension (time) is unsatisfactory, because dynamical features are less prominent in the computed sequence of saliency maps than the static edge- and corner-features. The SUSANinTime algorithms overcomes these problems both on artificial and real image sequences. Tests for human gestures and object manipulation by human hands have shown that important aspects of motion such as pointing events can be well captured.

In future work, we want to apply SUSANinTime for the classification of more complex events such as grasping an object. For this we plan to gather spatio-temporal IPs over a period of time long enough to capture the movement. Around each of the IPs local features will be extracted from the space-time volume. While in isolation features of this kind are not sufficient to characterize complex motion, we hope that a whole "cloud" of IPs provides sufficient information, which we intend to classify in a way similar to the (static) IP-based scene classification described in [16].

References

1. Martin, D.R., Fowlkes, C.C., Makik, J.: Learning to Detect Natural Image Boundaries Using Local Brightness, Color, and Texture Cues. IEEE Trans. on Pattern Analysis and Machine Intelligence vol. 26(1) (2004)
2. Reisfeld, D., Wolfson, H., Yeshurun, Y.: Context-Free Attentional Operators: The Generalized Symmetry Transform. of Computer Vision 14, 119–130 (1995)
3. Goldberger, J., Greenspan, H.: Context-Based Segmentation of Image Sequences. IEEE Trans. on Pattern Analysis and Machine Intelligence 28(3), 463–468 (2006)

4. Harris, C., Stephens, M.: A Combined Corner and Edge Detector. In: Proc. 4th Alvey Vision Conf. pp. 147–151 (1988)
5. Laptev, I., Lindeberg, T.: Space-time Interest Points. In: Proc. ICCV 2003. pp. 432–439 (2003)
6. Tian, Q., Sebe, N., Lew, M.S., Loupias, E., Huang, T.S.: Image Retrieval Using Wavelet-Based Salient Points. J. of Electronic Imaging 10(4), 835–849 (2001)
7. Backer, G., Mertsching, B., Bollmann, M.: Data- and Model-Driven Gaze Control for an Active-Vision System. IEEE Trans. on Pattern Analysis and Machine Intelligence 23(12), 1415–1429 (2001)
8. Heidemann, G.: Focus-of-Attention from Local Color Symmetries. IEEE Trans. on Pattern Analysis and Machine Intelligence 26(7), 817–830 (2004)
9. Privitera, C.M., Stark, L.W.: Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations. IEEE Trans. on Pattern Analysis and Machine Intelligence 22(9), 970–982 (2000)
10. Moravec, H.P.: Towards Automatic Visual Obstacle Avoidance. In: Proc. 5th Int'l Joint Conf. on Artificial Intelligence, Cambridge, Massachusetts, USA, pp. 584–587 (1977)
11. Schmid, C., Mohr, R.: Local Grayvalue Invariants for Image Retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence 19(5), 530–535 (1997)
12. Smith, S., Brady, J.: SUSAN – A New Approach to Low Level Image Processing. of Computer Vision 23(1), 45–78 (1997)
13. Zheng, Z., Wang, H., Teoh, W.: Analysis of Gray Level Corner Detection. Pattern Recognition Letters 20, 149–162 (1999)
14. Zitová, B., Kautsky, J., Peters, G., Flusser, J.: Robust detection of significant points in multiframe images. Pattern Recognition Letters 20(2), 199–206 (1999)
15. Heidemann, G., Kaiser, B., Bax, I., Bekel, H., Ritter, H.: Spatiotemporal Events and Action Sequences. Technical report, Bielefeld Univ., Neuroinformatics Group (2005)
16. Heidemann, G.: Unsupervised image categorization. Image and Vision Computing 23, 861–876 (2005)