Space-time Interest Points

Ivan Laptev and Tony Lindeberg*
Computational Vision and Active Perception Laboratory (CVAP)
Dept. of Numerical Analysis and Computer Science
KTH, SE-100 44 Stockholm, Sweden
{laptev, tony}@nada.kth.se

Abstract

Local image features or interest points provide compact and abstract representations of patterns in an image. In this paper, we propose to extend the notion of spatial interest points into the spatio-temporal domain and show how the resulting features often reflect interesting events that can be used for a compact representation of video data as well as for its interpretation.

To detect spatio-temporal events, we build on the idea of the Harris and Förstner interest point operators and detect local structures in space-time where the image values have significant local variations in both space and time. We then estimate the spatio-temporal extents of the detected events and compute their scale-invariant spatio-temporal descriptors. Using such descriptors, we classify events and construct video representation in terms of labeled space-time points. For the problem of human motion analysis, we illustrate how the proposed method allows for detection of walking people in scenes with occlusions and dynamic backgrounds.

1. Introduction

Analyzing and interpreting video is a growing topic in computer vision and its applications. Video data contains information about changes in the environment and is highly important for many visual tasks including navigation, surveillance and video indexing.

Traditional approaches for motion analysis mainly involve the computation of optic flow [1] or feature tracking [28, 4]. Although very effective for many tasks, both of these techniques have limitations. Optic flow approaches mostly capture first-order motion and often fail when the motion has sudden changes. Feature trackers often assume a constant appearance of image patches over time and may hence fail when this appearance changes, for example, in situations when two objects in the image merge or split.

*The support from the Swedish Research Council and from the Royal Swedish Academy of Sciences as well as the Knut and Alice Wallenberg Foundation is gratefully acknowledged.

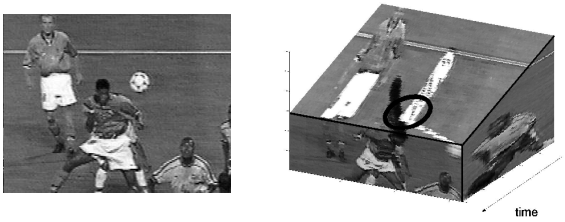


Figure 1: Result of detecting the strongest spatio-temporal interest point in a football sequence with a player heading the ball. The detected event corresponds to the high spatio-temporal variation of the image data or a "space-time corner" as illustrated by the spatio-temporal slice on the right.

Image structures in video are not restricted to constant velocity and/or constant appearance over time. On the contrary, many interesting events in video are characterized by strong variations of the data in both the spatial and the temporal dimensions. As example, consider scenes with a person entering a room, applauding hand gestures, a car crash or a water splash; see also the illustration in figure 1.

More generally, points with non-constant motion correspond to accelerating local image structures that might correspond to the accelerating objects in the world. Hence, such points might contain important information about the forces that act in the environment and change its structure.

In the spatial domain, points with a significant local variation of image intensities have been extensively investigated in the past [9, 11, 26]. Such image points are frequently denoted as "interest points" and are attractive due to their high information contents. Highly successful applications of interest point detectors have been presented for image indexing [25], stereo matching [30, 23, 29], optic flow estimation and tracking [28], and recognition [20, 10].

In this paper we detect interest points in the spatio-temporal domain and illustrate how the resulting space-time features often correspond to interesting events in video data. To detect spatio-temporal interest points, we build on the idea of the Harris and Förstner interest point operators [11, 9] and describe the detection method in section 2. To capture events with different spatio-temporal extents [32],

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

we compute interest points in spatio-temporal scale-space and select scales that roughly correspond to the size of the detected events in space and to their durations in time.

In section 3 we show how interesting events in video can be learned and classified using k-means clustering and point descriptors defined by local spatio-temporal image derivatives. In section 4 we consider video representation in terms of classified spatio-temporal interest points and demonstrate how this representation can be efficient for the task of video registration. In particular, we present an approach for detecting walking people in complex scenes with occlusions and dynamic background. Finally, section 5 concludes the paper with the discussion of the method.

2. Interest point detection

2.1. Interest points in spatial domain

The idea of the Harris interest point detector is to detect locations in a spatial image f^{sp} where the image values have significant variations in both directions. For a given scale of observation σ_l^2, such interest points can be found from a windowed second moment matrix integrated at scale $\sigma_i^2 = s\sigma_l^2$

$$\mu^{sp} = g^{sp}(\cdot;\, \sigma_i^2) * \begin{pmatrix} (L_x^{sp})^2 & L_x^{sp} L_y^{sp} \\ L_x^{sp} L_y^{sp} & (L_y^{sp})^2 \end{pmatrix} \tag{1}$$

where L_x^{sp} and L_y^{sp} are Gaussian derivatives defined as

$$\begin{aligned} L_x^{sp}(\cdot;\, \sigma_l^2) &= \partial_x (g^{sp}(\cdot;\, \sigma_l^2) * f^{sp}) \\ L_y^{sp}(\cdot;\, \sigma_l^2) &= \partial_y (g^{sp}(\cdot;\, \sigma_l^2) * f^{sp}), \end{aligned} \tag{2}$$

and where g^{sp} is the spatial Gaussian kernel

$$g^{sp}(x, y;\, \sigma^2) = \frac{1}{2\pi\sigma^2} \exp(-(x^2 + y^2)/2\sigma^2). \tag{3}$$

As the eigenvalues $\lambda_1, \lambda_2, (\lambda_1 \leq \lambda_2)$ of μ^{sp} represent characteristic variations of f^{sp} in both image directions, two significant values of λ_1, λ_2 indicate the presence of an interest point. To detect such points, Harris and Stephens [11] propose to detect positive maxima of the corner function

$$H^{sp} = \det(\mu^{sp}) - k \operatorname{trace}^2(\mu^{sp}) = \lambda_1 \lambda_2 - k(\lambda_1 + \lambda_2)^2. \tag{4}$$

2.2. Interest points in the space-time

The idea of interest points in the spatial domain can be extended into the spatio-temporal domain by requiring the image values in space-time to have large variations in both the spatial and the temporal dimensions. Points with such properties will be spatial interest points with a distinct location in time corresponding to the moments with non-constant motion of the image in a local spatio-temporal neighborhood [15].

To model a spatio-temporal image sequence, we use a function $f : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ and construct its linear scale-space representation $L : \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}_+^2 \mapsto \mathbb{R}$ by convolution of f with an anisotropic Gaussian kernel[1] with distinct spatial variance σ_l^2 and temporal variance τ_l^2

$$L(\cdot;\, \sigma_l^2, \tau_l^2) = g(\cdot;\, \sigma_l^2, \tau_l^2) * f(\cdot), \tag{5}$$

where the spatio-temporal separable Gaussian kernel is defined as

$$g(x, y, t;\, \sigma_l^2, \tau_l^2) = \frac{\exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/2\tau_l^2)}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}}. \tag{6}$$

Similar to the spatial domain, we consider the spatio-temporal second-moment matrix which is a 3-by-3 matrix composed of first order spatial and temporal derivatives averaged with a Gaussian weighting function $g(\cdot;\, \sigma_i^2, \tau_i^2)$

$$\mu = g(\cdot;\, \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}, \tag{7}$$

where the integration scales are $\sigma_i^2 = s\sigma_l^2$ and $\tau_i^2 = s\tau_l^2$, while the first-order derivatives are defined as $L_\xi(\cdot;\, \sigma_l^2, \tau_l^2) = \partial_\xi (g * f)$. The second-moment matrix μ has been used previously by Nagel and Gehrke [24] in the context of optic flow computation.

To detect interest points, we search for regions in f having significant eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of μ. Among different approaches to find such regions, we choose to extend the Harris corner function (4) defined for the spatial domain into the spatio-temporal domain by combining the determinant and the trace of μ in the following way

$$H = \det(\mu) - k \operatorname{trace}^3(\mu) = \lambda_1 \lambda_2 \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3. \tag{8}$$

To show that the positive local maxima of H correspond to points with high values of $\lambda_1, \lambda_2, \lambda_3$ $(\lambda_1 \leq \lambda_2 \leq \lambda_3)$, we define the ratios $\alpha = \lambda_2/\lambda_1$ and $\beta = \lambda_3/\lambda_1$ and rewrite $H = \lambda_1^3(\alpha\beta - k(1 + \alpha + \beta)^3)$. Then, from the requirement $H \geq 0$, we get $k \leq \alpha\beta/(1 + \alpha + \beta)^3$ and it follows that as k increases towards its maximal value $k = 1/27$, both ratios α and β tend to one. For sufficiently large values of k, positive local maxima of H correspond to points with high variation of the image gray-values in both the spatial and the temporal dimensions. Thus, spatio-temporal interest points of f can be found by detecting local positive spatio-temporal maxima in H.

[1] In general, convolution with a Gaussian kernel in the temporal domain violates causality constraints, since the temporal image data is available only for the past. Whereas for real-time implementations this problem can be solved using causal recursive filters [12, 19], in this paper we simplify the investigation and assume that the data is available for a sufficiently long period of time and that the image sequence can hence be convolved with a truncated Gaussian in both space and time. However, the proposed interest points can be computed using recursive filters in on-line mode.

2.3. Experiments on synthetic sequences

To illustrate the detection of spatio-temporal interest points on synthetic image sequences, we show the spatio-temporal data as 3-D space-time plots where the original signal is represented by a threshold surface while the detected interest points are presented by ellipsoids with semi-axes proportional to corresponding scale parameters σ_l and τ_l.

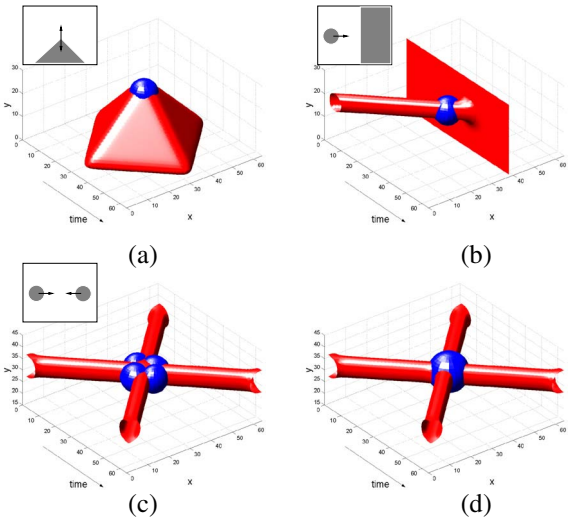


Figure 2: Results of detecting spatio-temporal interest points on synthetic image sequences: (a) Moving corner; (b) A merge of a ball and a wall; (c) Collision of two balls with interest points detected at scales $\sigma_l^2 = 8$ and $\tau_l^2 = 8$; (d) the same as in (c) but with interest points detected at scales $\sigma_l^2 = 16$ and $\tau_l^2 = 16$.

Figure 2a illustrates a sequence with a moving corner. The interest point is detected at the moment in time when the motion of the corner changes direction. This type of event occurs frequently in natural sequences such as sequences of articulated motion. Other typical types of events detected by the proposed method are splits and mergers of image structures. In figure 2b, the interest point is detected at the moment and the position corresponding to the collision of a ball and a wall. Similarly, interest points are detected at the moment of collision and bouncing of two balls as shown in figure 2c-d. Note, that different types of events are detected depending on the scale of observation.

In general, the result of interest point detector will depend on the scale parameters. Hence, the correct estimation of spatio-temporal extents of events is highly important for their detection and further interpretation.

2.4. Scale selection in space-time

To estimate the spatio-temporal extent of an event in space-time, we follow the idea of local scale selection proposed in the spatial domain by Lindeberg [18] as well as in the temporal domain [17]. As a prototype event we study a spatio-temporal Gaussian blob $f = g(x, y, t; \sigma_0^2, \tau_0^2)$ with spatial variance σ_0^2 and temporal variance τ_0^2. Using the semi-group property of the Gaussian kernel, it follows that the scale-space representation of f is $L(x, y, t; \sigma^2, \tau^2) = g(x, y, t; \sigma_0^2 + \sigma^2, \tau_0^2 + \tau^2)$.

To recover the spatio-temporal extent (σ_0, τ_0) of f we consider the scale-normalized spatio-temporal Laplacian operator defined by

$$\nabla^2_{norm} L = L_{xx,norm} + L_{yy,norm} + L_{tt,norm}, \quad (9)$$

where $L_{xx,norm} = \sigma^{2a}\tau^{2b} L_{xx}$ and $L_{tt,norm} = \sigma^{2c}\tau^{2d} L_{tt}$. As shown in [15], given the appropriate normalization parameters $a = 1, b = 1/4, c = 1/2$ and $d = 3/4$, the size of the blob f can be estimated from the scale values $\tilde{\sigma}^2$ and $\tilde{\tau}^2$ for which $\nabla^2_{norm} L$ assumes local extrema over scales, space and time. Hence, the spatio-temporal extent of the blob can be estimated by detecting local extrema of

$$\nabla^2_{norm} L = \sigma^2 \tau^{1/2} (L_{xx} + L_{yy}) + \sigma \tau^{3/2} L_{tt}. \quad (10)$$

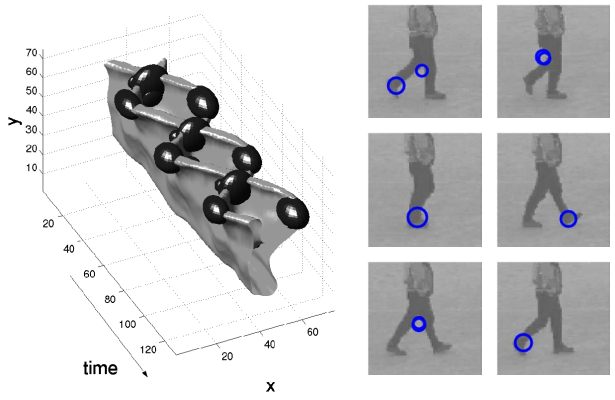
over both spatial and temporal scales.

2.5. Scale-adapted space-time interest points

Local scale estimation using the normalized Laplace operator has shown to be very useful in the spatial domain [18, 6]. In particularly, Mikolajczyk and Schmid [22] combined the Harris interest point operator with the normalized Laplace operator and derived the scale-invariant Harris-Laplace interest point detector. The idea is to find points in scale-space that are both maxima of the Harris function H^{sp} (4) in space and extrema of the scale-normalized spatial Laplace operator over scale.

Here, we extend this idea and detect interest points that are simultaneous maxima of the spatio-temporal corner function H (8) as well as extrema of the normalized spatio-temporal Laplace operator $\nabla^2_{norm} L$ (9). Hence, we detect interest points for a set of sparsely distributed scale values and then track these points in spatio-temporal scale-time-space towards the extrema of $\nabla^2_{norm} L$. We do this by iteratively updating the scale and the position of the interest points by (i) selecting the neighboring spatio-temporal scale that maximizes $(\nabla^2_{norm} L)^2$ and (ii) re-detecting the space-time location of the interest point at a new scale until the position and the scale converge to the stable values [15].

To illustrate the performance of the scale-adapted spatio-temporal interest point detector, let us consider a sequence with a walking person and non-constant image velocities due to the oscillating motion of the legs. As can be seen in figure 3, the pattern gives rise to stable interest points. Note that the detected points are well-localized both in space and time and correspond to events such as the stopping and starting feet. From the space-time plot in figure 3(a), we can also

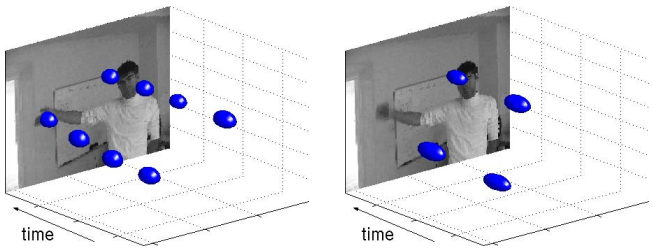


(a) (b)

Figure 3: Results of detecting spatio-temporal interest points for the motion of the legs of a walking person: (a) 3-D plot with a threshold surface of a leg pattern (up side down) and detected interest points; (b) interest points overlaid on single frames in the sequence.

observe how the selected spatial and temporal scales of the detected features roughly match the spatio-temporal extents of the corresponding image structures.

Hand waves with high frequency *Hand waves with low frequency*



(a) (b)

Figure 4: Result of interest point detection for a sequence with waving hand gestures: (a) Interest points for hand gestures with high frequency; (b) Interest points for hand gestures with low frequency.

The second example explicitly illustrates how the proposed method is able to estimate the temporal extent of detected events. Figure 4 shows a person making hand-waving gestures with high frequency on the left and low frequency on the right. The distinct interest points are detected at moments and at spatial positions where the hand changes its direction of motion. Whereas the spatial scales of the detected interest points remain roughly constant, the selected temporal scales depend on the frequency of the wave pattern.

3. Classification of events

The detected interest points have significant variations of image values in the local spatio-temporal neighborhood. To differentiate events from each other and from noise, one approach is to compare their local neighborhoods and assign points with similar neighborhoods to the same class of events. Similar approach has proven to be successful in the spatial domain for the task of image representation [21] indexing [25] and recognition [10, 31, 16]. In the spatio-temporal domain local descriptors have been used previously by [7] and others.

To describe a spatio-temporal neighborhood we consider normalized spatio-temporal Gaussian derivatives defined as

$$L_{x^m y^n t^k} = \sigma^{m+n} \tau^k (\partial_{x^m y^n t^k} g) * f, \qquad (11)$$

computed at the scales used for detecting the corresponding interest points. The normalization with respect to scale parameters guarantees the invariance of the derivative responses with respect to image scalings in both the spatial domain and the temporal domain. Using derivatives, we define event descriptors from the third order local jet[2] [13]

$$j = (L_x, L_y, L_t, L_{xx}, ..., L_{ttt}). \qquad (12)$$

To compare two events we compute the Mahalanobis distance between their descriptors as

$$d^2(j_1, j_2) = (j_1 - j_2)\Sigma^{-1}(j_1 - j_2)^T, \qquad (13)$$

where Σ is a covariance matrix corresponding to the typical distribution of interest points in the data.

To detect similar events in the data, we apply k-means clustering [8] in the space of point descriptors and detect groups of points with similar spatio-temporal neighborhoods. The clustering of spatio-temporal neighborhoods is similar to the idea of textons [21] used to describe image texture as well as to detect object parts for spatial recognition [31]. Given training sequences with periodic motion, we can expect repeating events to give rise to populated clusters. On the contrary, sporadic interest points can be expected to be sparsely distributed over the descriptor space giving rise to weakly populated clusters. To prove this idea we applied k-means clustering with $k = 15$ to the sequence of a walking person in the upper row of figure 5. We found out that four of the most densely populated clusters $c1, ..., c4$ indeed corresponded to the stable interest points of the gait pattern. Local spatio-temporal neighborhoods of these points are shown in figure 6, where we can confirm the similarity of patterns inside the clusters and their difference between clusters.

[2]Note that our representation is currently not invariant with respect to planar image rotations. Such invariance could be added if considering steerable derivatives or rotationally invariant operators in space.

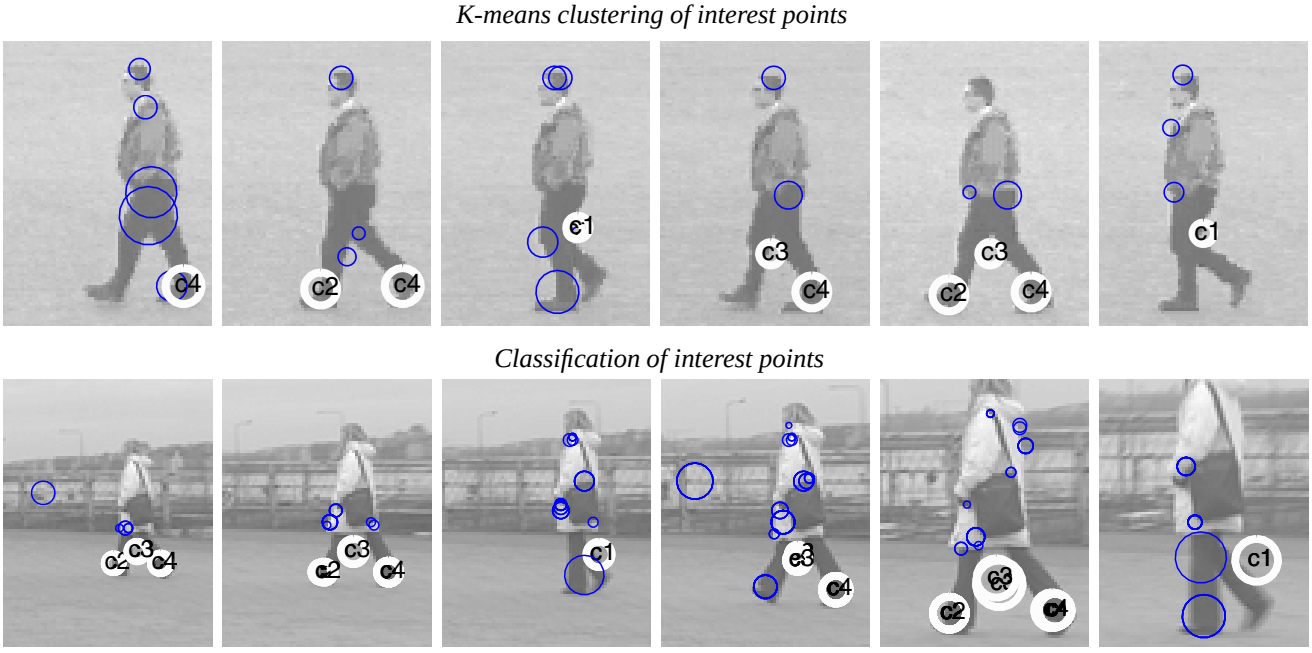


Figure 5: Interest points detected for sequences of walking persons. First row: result of clustering spatio-temporal interest points. Labeled points correspond to the four most populated clusters; Second row: result of classification of interest points with respect to the clusters found in the first sequence.

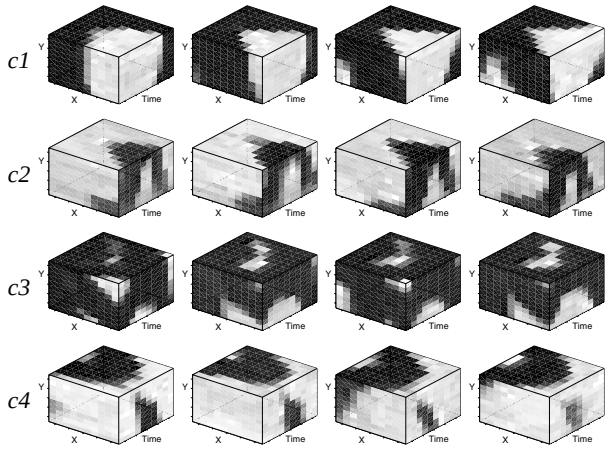


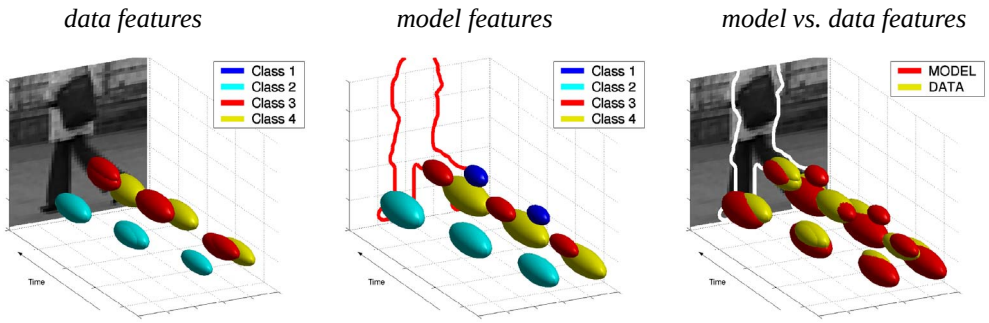
Figure 6: Local spatio-temporal neighborhoods of interest points corresponding to the first four most populated clusters.

To represent characteristic repetitive events in video, we compute cluster means $m_i = \frac{1}{n_i} \sum_{k=1}^{n_i} j_k$ for each significant cluster c_i consisting of n_i points. Then, in order to classify an event on an unseen sequence, we assign the detected point to the cluster c_i if it minimizes the distance $d(m_i, j_0)$ (13) between the jet of the interest point j_0 and the cluster mean m_i. If the distance is above a threshold, the point is classified as the background. Application of this classification scheme is demonstrated in the second row of figure 5. As can be seen, most of the points corresponding to the gait pattern are correctly classified while the other interest points are discarded. Observe that the person in the second sequence of figure 5 undergoes significant size changes in the image. Due to the scale-invariance of interest points as well as the jet responses, size transformations do not effect neither the result of event detection nor the result of classification.

4. Application to video interpretation

In this section, we illustrate how the representation of video sequences by classified spatio-temporal interest points can be used for video interpretation. We consider the problem of detecting walking people and estimating their poses when viewed from the side in outdoor scenes. Such a task is complicated, since the variations in appearance of people together with the variations in the background may lead to ambiguous interpretations. Human motion is a strong cue that has been used to resolve this ambiguity in a number of previous works. Some of the works rely on pure spatial image features while using sophisticated body models and tracking schemes to constrain the interpretation [2, 5, 27]. Other approaches use spatio-temporal image cues such as optical flow [3] or motion templates [2].

The idea of this approach is to represent both the model and the data using local and discriminative spatio-temporal features and to match the model by matching its features



(a) (b) (c)

Figure 7: Matching of spatio-temporal data features with model features: (a) Features detected from the data sequence in the time interval corresponding to three periods of the gait cycle; (b) Model features minimizing the distance to the features in (a); (c) Model features and data features overlaid. The estimated silhouette overlayed on the current frame confirms the correctness of the method.

to the correspondent features of the data inside a spatio-temporal window (see figure 7).

4.1. Walking model

To obtain the model of a walking person, we consider the upper sequence in figure 5 and manually select the time interval $(t_0, t_0 + T)$ corresponding to the period T of the gait pattern. Then, given n features $f_i = (x_i^m, y_i^m, t_i^m, \sigma_i^m, \tau_i^m, c_i^m)$, $i = 1, ..., n$ defined by the positions (x_i^m, y_i^m, t_i^m), scales (σ_i^m, τ_i^m) and classes c_i^m of interest points detected in the selected time interval, i.e. $t_i^m \in (t_0, t_0 + T)$, we define the walking model by the set of periodically repeating features $M = \{f_i + (0, 0, kT, 0, 0, 0, 0) | i = 1, ..., n, k \in \mathcal{Z}\}$. Furthermore, to account for variations of the position and the size of a person in the image, we introduce the state of the model determined by the vector $X = (x, y, \theta, s, v_x, v_y, v_s)$. The components of X describe the position of the person in the image (x, y), his size s, the phase θ of the gait cycle at the current time moment as well as the temporal variations (v_x, v_y, v_s) of (x, y, s). Given the state X, the parameters of each model feature $f \in M$ transform according to

$$\begin{aligned} \tilde{x}^m &= sx^m + x + v_x(t^m + \theta) + sx^m v_s(t^m + \theta) \\ \tilde{y}^m &= sy^m + y + v_y(t^m + \theta) + sy^m v_s(t^m + \theta) \\ \tilde{t}^m &= t^m + \theta \\ \tilde{\sigma}^m &= s\sigma^m + v_s s\sigma^m(t^m + \theta)) \\ \tilde{\tau}^m &= \tau^m \\ \tilde{c}^m &= c^m \end{aligned} \tag{14}$$

It follows that the current scheme does not allow for scalings of the model in the temporal direction and enables only the first-order variations of positions and sizes of the model features over time. These restrictions have not caused problems in our experiments and can be easily removed by introducing additional parameters in the state vector X and corresponding rules for updating the model features.

To estimate the boundary of the person, we extract silhouettes $S = \{x^s, y^s, \theta^s | \theta^s = 1, ..., T\}$ on the model sequence (see figure 5) one for each frame corresponding to the discrete value of the phase parameter θ. The silhouettes here are used only for the visualization purposes and enable to approximate the boundary of the person for the current frame and model state X by points $\{(x^s, y^s, \theta^s) \in S | \theta^s = \theta\}$ transformed according to $\tilde{x}^s = sx^s + x$, $\tilde{y}^s = sy^s + y$.

4.2. Model matching

Given the model state X, the current time t_0, the length of the time window t_w, and the data features $D = \{f^d = (x^d, y^d, t^d, \sigma^d, \tau^d, c^d) | t^d \in (t_0, t_0 - t_w)\}$ detected from the the recent time window of the data sequence, the match between the model and the data is defined by the weighted sum of distances h between the model and the data features

$$\mathcal{H}(\tilde{M}(X), D, t_0) = \sum_i^n h(\tilde{f}_i^m, f_j^d) e^{-(\tilde{t}_i^m - t_0)^2/\xi}, \tag{15}$$

where $\tilde{M}(X)$ is a set of n model features in the time window $(t_0, t_0 - t_w)$ transformed according to (14), i.e. $\tilde{M} = \{\tilde{f}^m | t^m \in (t_0, t_0 - t_w)\}$, $f_j^d \in D$ is a data feature minimizing the distance h for a given f_i^m and ξ is the variance of the exponential weighting function that intends to give more importance to recent features.

The distance h between two features of the same class is defined as a Euclidean distance between two points in space-time, where the spatial and the temporal dimensions are weighted with respect to parameter ν as well as by extents of features in space-time

$$h^2(f^m, f^d) = \frac{(x^m - x^d)^2 + (y^m - y^d)^2}{(1 - \nu)(\sigma^m)^2} + \frac{(t^m - t^d)^2}{\nu(\tau^m)^2}. \tag{16}$$

The distance between features of different classes is regarded as infinite.

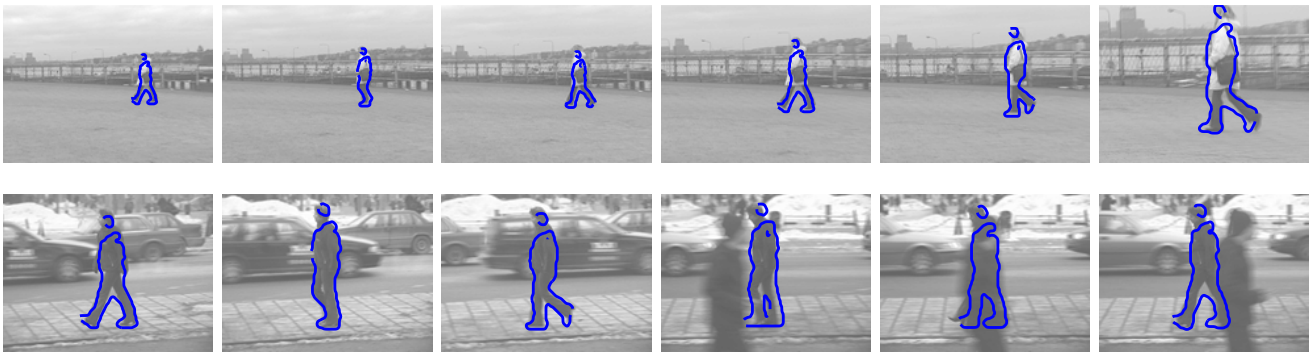


Figure 8: The result of matching spatio-temporal walking model to the sequences of outdoor scenes.

To find the best match between the model and the data, we search for the model state $\tilde{X}$ that minimizes $\mathcal{H}$ in (15)

$$\tilde{X} = \text{argmin}_X \, \mathcal{H}(\tilde{M}(X), D, t_0) \qquad (17)$$

using a standard Gauss-Newton optimization method. The result of such an optimization for a sequence with data features in figure 7(a) is illustrated in figure 7(b). Here the match between the model and the data features was searched over a time window corresponding to three periods of the gait pattern or approximately 2 seconds of video. As can be seen from figure 7(c), the overlaps between the model features and the data features confirm the match between the model and the data. Moreover, the model silhouette transformed according to $\tilde{X}$ matches with the contours of the person in the current frame and confirms a reasonable estimation of model parameters.

4.3. Results

Figure 8 presents results of the described approach applied to two outdoor sequences. The first sequence illustrates the invariance of the method with respect to size variations of the person in the image plane. The second sequence shows the successful detection and pose estimation of a person despite the presence of a complex non-stationary background and occlusions. Note that these results have been obtained by re-initializing model parameters before optimization at each frame. Hence, the approach is highly stable and could be improved even more by tracking the model parameters $\tilde{X}$ over time.

The need of careful initialization and/or simple background have been frequent obstacles in previous approaches for human motion analysis. The success of our method is due to the low ambiguity and simplicity of the matching scheme originating from the distinct and stable nature of the spatio-temporal features. In this respect, direct detection of spatio-temporal events constitutes an interesting alternative when representing and interpreting video data.

5. Summary

We have described an interest point detector that finds local image features in space-time characterized by high variation of the image values in space and non-constant motion in time. From the presented examples, it follows that many of the detected points indeed correspond to meaningful events. Moreover, estimation of characteristic local scales provides information about the spatio-temporal extents of events and enables a computation of scale-invariant descriptors.

Using differential descriptors computed at interest points we addressed the problem of event classification and illustrated how classified spatio-temporal interest points constitute distinct and stable descriptors of events in video that can be used for video representation and interpretation. In particular, we have shown how video representation by spatio-temporal interest points enables the detection and pose estimation of walking people in the presence of occlusions and highly cluttered and dynamic background. Note that this result was obtained using a standard optimization method without careful manual initialization nor tracking.

In future work we plan to extend application of interest points to the field of motion-based recognition. Moreover, as the current detection scheme is not invariant under Galilean transformations, future work should investigate the possibilities of including such an invariance and making the approach independent of the relative camera motion [14]. Another extension should consider the invariance of spatio-temporal descriptors with respect to the direction of motion, changes in image contrast and rotations.

6 Acknowledgments

We thank Anastasiya Syromyatnikova and Josephine Sullivan for their help in obtaining video data for the experiments.

References

[1] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *IJCV*, 12(1):43–77, February 1994.

[2] A. M. Baumberg and D.C. Hogg. Generating spatiotemporal models from examples. *Image and Vision Computing*, 14(8):525–532, August 1996.

[3] M.J. Black, Y. Yacoob, A.D. A.D. Jepson, and D.J. D.J. Fleet. Learning parameterized models of image motion. In *Proc. CVPR*, pages 561–567, 1997.

[4] A. Blake and M. Isard. Condensation – conditional density propagation for visual tracking. *IJCV*, 29(1):5–28, August 1998.

[5] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. CVPR*, pages 8–15, Santa Barbara, CA, 1998.

[6] O. Chomat, V.C. de Verdiere, D. Hall, and J.L. Crowley. Local scale selection for Gaussian based description techniques. In *Proc. ECCV*, volume 1842 of *Lecture Notes in Computer Science*, pages I:117–133, Dublin, Ireland, 2000. Springer Verlag, Berlin.

[7] O. Chomat, J. Martin, and J.L. Crowley. A probabilistic sensor for the perception and recognition of activities. In *Proc. ECCV*, volume 1842 of *Lecture Notes in Computer Science*, pages I:487–503, Dublin, Ireland, 2000. Springer Verlag, Berlin.

[8] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, 2001.

[9] W. A. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centers of circular features. In *ISPRS*, 1987.

[10] D. Hall, V.C. de Verdiere, and J.L. Crowley. Object recognition using coloured receptive fields. In *Proc. ECCV*, volume 1842 of *Lecture Notes in Computer Science*, pages I:164–177, Dublin, Ireland, 2000. Springer Verlag, Berlin.

[11] C. Harris and M.J. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–152, 1988.

[12] J. J. Koenderink. Scale-time. *Biol. Cyb.*, 58:159–162, 1988.

[13] J.J. Koenderink and A.J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.

[14] I. Laptev and T. Lindeberg. Velocity-adaptation of spatio-temporal receptive fields for direct recognition of activities: An experimental study. In D. Suter, editor, *Proc. ECCV'02 workshop on Statistical Methods in Video Processing*, pages 61–66, Copenhagen, Denmark, 2002.

[15] I. Laptev and T. Lindeberg. Interest point detection and scale selection in space-time. In L.D. Griffin and M Lillholm, editors, *Scale-Space'03*, volume 2695 of *Lecture Notes in Computer Science*, pages 372–387. Springer Verlag, Berlin, 2003.

[16] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, June 2001.

[17] T. Lindeberg. On automatic selection of temporal scales in time-causal scale-space. In *AFPAC'97: Algebraic Frames for the Perception-Action Cycle*, volume 1315 of *Lecture Notes in Computer Science*, pages 94–113. Springer Verlag, Berlin, 1997.

[18] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):77–116, 1998.

[19] T. Lindeberg and D. Fagerström. Scale-space with causal time direction. In *Proc. ECCV*, volume 1064 of *Lecture Notes in Computer Science*, pages I:229–240, Cambridge, UK, 1996. Springer Verlag, Berlin.

[20] D.G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, pages 1150–1157, Corfu, Greece, 1999.

[21] J. Malik, S. Belongie, J. Shi, and T. Leung. Textons, contours and regions: Cue integration in image segmentation. In *Proc. ICCV*, pages 918–925, Corfu, Greece, 1999.

[22] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proc. ICCV*, pages I:525–531, Vancouver, Canada, 2001.

[23] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. ECCV*, volume 2350 of *Lecture Notes in Computer Science*, pages I:128–142, Copenhagen, Denmark, 2002. Springer Verlag, Berlin.

[24] H.H. Nagel and A. Gehrke. Spatiotemporal adaptive filtering for estimation and segmentation of optical flow fields. In H. Burkhardt and B. Neumann, editors, *Proc. ECCV*, volume 1407 of *Lecture Notes in Computer Science*, pages II:86–102, Freiburg, Germany, 1998. Springer Verlag, Berlin.

[25] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE-PAMI*, 19(5):530–535, May 1997.

[26] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *IJCV*, 37(2):151–172, June 2000.

[27] H. Sidenbladh, M.J. Black, and D.J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *Proc. ECCV*, volume 1843 of *Lecture Notes in Computer Science*, pages II:702–718, Dublin, Ireland, 2000. Springer Verlag, Berlin.

[28] S.M. Smith and J.M. Brady. ASSET-2: Real-time motion segmentation and shape tracking. *IEEE-PAMI*, 17(8):814–820, 1995.

[29] D. Tell and S. Carlsson. Combining topology and appearance for wide baseline matching. In *Proc. ECCV*, volume 2350 of *Lecture Notes in Computer Science*, pages I:68–83, Copenhagen, Denmark, 2002. Springer Verlag, Berlin.

[30] T. Tuytelaars and L.J. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *British Machine Vision Conference*, pages 412–425, 2000.

[31] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for visual object class recognition. In *Proc. ECCV*, volume 1842 of *Lecture Notes in Computer Science*, pages I:18–32, Dublin, Ireland, 2000. Springer Verlag, Berlin.

[32] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *Proc. CVPR*, pages II:123–130, Kauai Marriott, Hawaii, 2001.