Velocity adaptation of 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

The notion of local features in space-time has recently been proposed to capture and describe local events in video. When computing space-time descriptors, however, the result may strongly depend on the relative motion between the object and the camera. To compensate for this variation, we present a method that automatically adapts the features to the local velocity of the image pattern and, hence, results in a video representation that is stable with respect to different amounts of camera motion. Experimentally we show that the use of velocity adaptation substantially increases the repeatability of interest points as well as the stability of their associated descriptors. Moreover, for an application to human action recognition we demonstrate how velocity-adapted features enable recognition of human actions in situations with unknown camera motion and complex, non-stationary backgrounds.

1. Introduction

Video interpretation is a key component in many applications including video retrieval, surveillance, robot navigation and human-computer interaction. This motivates general-purpose approaches to video analysis that do not rely on restricting assumptions about particular types of motion patterns or environments.

Recently, several successive methods have been proposed for learning and recognizing a broad class of motion patterns directly from image measurements [13, 1]. Image measurements, such as optical flow and spatio-temporal gradients, however, depend on the recording conditions including the spatial resolution of the pattern and the motion relative to the camera. Whereas this problem can be solved in principle by external mechanisms for spatial segmentation and/or camera stabilization, such mechanisms might be unstable in complex situations and there is a need for alternative video representations that are stable with respect to the image acquisition conditions.

In this paper, we consider video representations in terms of local space-time features (interest points) [5]. We show how such features can be adapted to the relative velocity and the spatio-temporal scales of the pattern and obtain a video representation that is stable to the abovementioned variations. We demonstrate how this representation can be used for recognizing human actions in situations with complex backgrounds and unknown camera motion.

2. Theory

Space-time interest points [5] have recently been proposed to capture and represent local events in video. Such points have stable locations in space-time and correspond to moving two-dimensional image structures at moments of non-constant motion (see Figure 1a). A direct approach to detect such points consists of maximizing a measure of the local variations in the image sequence $f(x, y, t)$ over both space (x, y) and time t. For this purpose, consider a scale-space representation $L(\cdot, \Sigma) = f * g(\cdot, \Sigma)$ generated by the convolution of f with a spatio-temporal Gaussian kernel $g = \exp(\frac{-1}{2}(x, y, t)\Sigma^{-1}(x, y, t)^T)/\sqrt{(2\pi)^3 \det(\Sigma)}$ with a 3×3 covariance matrix Σ. The image variation in a Σ-neighborhood of a space-time point $(\cdot)$ can now be measured by a second-moment matrix composed from spatio-temporal gradients $\nabla L = (L_x, L_y, L_t)^T$

$$\mu(\cdot;\, \Sigma) = g(\cdot;\, s\Sigma) * \left(\nabla L (\nabla L)^T\right) = \begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{12} & \mu_{22} & \mu_{23} \\ \mu_{13} & \mu_{23} & \mu_{33} \end{pmatrix} \tag{1}$$

integrated within a Gaussian window. Neighborhoods with μ of rank 3 correspond to points with significant variations of image values over both space and time. Points that maximize these variations can be detected by maximizing all eigenvalues $\lambda_1, .., \lambda_3$ of μ or, similarly, by searching the

*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. We also thank Christian Schüldt and Barbara Caputo for their help in obtaining the experimental video data.

maxima of the interest point operator H [5] over (x, y, t)

$$H = \det(\mu) - k \operatorname{trace}^3(\mu) = \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3, \quad (2)$$

where k is a threshold on the discrepancy between $\lambda_1, .., \lambda_3$.

2.1. Galilean Transformation

The formulation of the interest operator H (2) in terms of eigenvalues implies invariance with respect to 3D rotations of the image pattern f. Whereas 2D rotations are common in the spatial domain, a 3D rotation in space-time does not correspond to any known physical transformation in the world. On the other hand, the temporal domain is highly effected by Galilean transformations caused by the constant motion between the camera and the observed pattern [7, 6] (see Figures 1a-b). A Galilean transformation G is defined by a velocity vector $(v_x, v_y)^T$

$$G = \begin{pmatrix} 1 & 0 & v_x \\ 0 & 1 & v_y \\ 0 & 0 & 1 \end{pmatrix} \quad (3)$$

and corresponds to a linear coordinate transformation $p' = Gp$ that has a skewing effect on the image function $L'(p';\, \Sigma') = L(p;\, \Sigma)$. It can be shown [7], that the covariance matrix Σ of the filter kernel g transforms under G according to $\Sigma' = G\Sigma G^T$ while the spatio-temporal gradient transforms as $\nabla L' = G^{-T}\nabla L$. Using these properties, the transformation of μ (1) can be derived [9] as

$$\mu'(p';\, \Sigma') = G^{-T}\mu(p;\, \Sigma)G^{-1} \quad (4)$$

and it follows that μ and, hence, the interest operator H (2) is not preserved under Galilean transformation.

2.2. Velocity adaptation

Our goal is to detect space-time interest points and to determine their neighborhoods independently of the relative velocity of the camera. When using the μ-descriptor, it is necessary to cancel the effect of a Galilean transformation and to re-define the interest operator in terms of a *velocity-adapted* descriptor $\mu'' = G^T\mu' G$. This, however, requires an estimate of G, which may not be available in advance. To estimate G from the data, we can note that the solution of

$$\begin{pmatrix} \mu'_{11} & \mu'_{12} \\ \mu'_{12} & \mu'_{22} \end{pmatrix} \begin{pmatrix} \tilde{v}_x \\ \tilde{v}_y \end{pmatrix} = \begin{pmatrix} \mu'_{13} \\ \mu'_{23} \end{pmatrix} \quad (5)$$

with respect to $(\tilde{v}_x, \tilde{v}_y)^T$ is structurally similar to the equation for computing optical flow according to Lukas and Kanade [10, 3]. Hence, we can use the matrix μ' for estimating the velocities $(\tilde{v}_x, \tilde{v}_y)^T$

$$\tilde{v}_x = \frac{\mu'_{22}\mu'_{13} - \mu'_{12}\mu'_{23}}{\mu'_{11}\mu'_{22} - \mu'^2_{12}}, \quad \tilde{v}_y = \frac{\mu'_{11}\mu'_{23} - \mu'_{12}\mu'_{13}}{\mu'_{11}\mu'_{22} - \mu'^2_{12}} \quad (6)$$

and obtaining a velocity-adapted descriptor μ''.[1] Such a method brings $\mu'' = G^T(\tilde{v}_x, \tilde{v}_y)\mu' G(\tilde{v}_x, \tilde{v}_y)$ into a block-diagonal form with elements $\mu''_{13} = \mu''_{23} = 0$ for any initial Galilean transformation $G(v_x, v_y)$. Hence, the block-diagonal form of μ can be used as a standard form for computing spatio-temporal image descriptors independently of the Galilean transformation. Moreover, the component

$$\mu''_{33} = \mu'_{33} + \frac{2\mu'_{13}\mu'_{12}\mu'_{23} - \mu'_{22}(\mu'_{13})^2 - \mu'_{11}(\mu'_{23})^2}{\mu'_{11}\mu'_{22} - (\mu'_{12})^2} \quad (7)$$

encodes all information in μ'' about the local non-linear motion. The block-diagonal form of μ'' then implies that μ''_{33} is the "pure temporal" eigenvalue of μ'' while the two other eigenvalues encode "pure spatial" variation in f. Hence, a *velocity-corrected* interest operator can now be re-defined in terms of μ'' as [8]:

$$H_{corr} = \det(\mu'') - k \operatorname{trace}^3(\mu''). \quad (8)$$

2.3. Velocity-adapted interest points

It should be noted, however, that the entities μ'' and H_{corr} are only approximately invariant under Galilean transformations, since the estimates $\tilde{v}_x, \tilde{v}_y$ depend on the shape of the filter kernels Σ' used for computing μ'. To obtain a strict invariance, filter kernels have to be adapted according to $\Sigma'' = G^{-1}\Sigma' G^{-T}$. Although this cannot be done directly, since G is unknown, we can iteratively estimate G using $\tilde{v}_x, \tilde{v}_y$ (6) and update Σ'' at each iteration until the velocities $\tilde{v}_x, \tilde{v}_y$ converge to stable values.

To detect velocity-adapted interest points, we start by detecting maxima of $H_{corr}(\cdot,\, \Sigma'_0)$ using filter kernels with

$$\Sigma'_0 = \begin{pmatrix} 1 & 0 & v_{x0} \\ 0 & 1 & v_{y0} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_0^2 & 0 \\ 0 & 0 & \tau_0^2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ v_{x0} & v_{y0} & 1 \end{pmatrix}$$

for some initial values of velocities v_{x0}, v_{y0} and spatial and temporal scale parameters σ_0^2, τ_0^2. The position and the neighborhood of each detected point is then updated iteratively. In each iteration k, the covariance Σ''_k is updated according to the recent velocity estimates $v_{x,k-1}, v_{y,k-1}$ and the position is updated by re-detecting maxima of $H_{corr}(\cdot,\, \Sigma''_k)$. Using the scale-selection approach in [5], we additionally update the scales of the interest points σ_k^2, τ_k^2 and stop iterating when the velocities, the scales and the position of the interest point converge to stable values.

This procedure bears close similarities to the adaptation of spatial interest points [2] with respect to affine transformations in the image plane [11]. In fact, the proposed velocity adaptation in space-time could be combined with affine adaptation in space by estimating the affine transformation from the spatial part of the second-moment matrix μ in (1).

[1] This idea is closely related to the local affine normalization of 2D image patches using a second-moment matrix in the spatial domain [9].

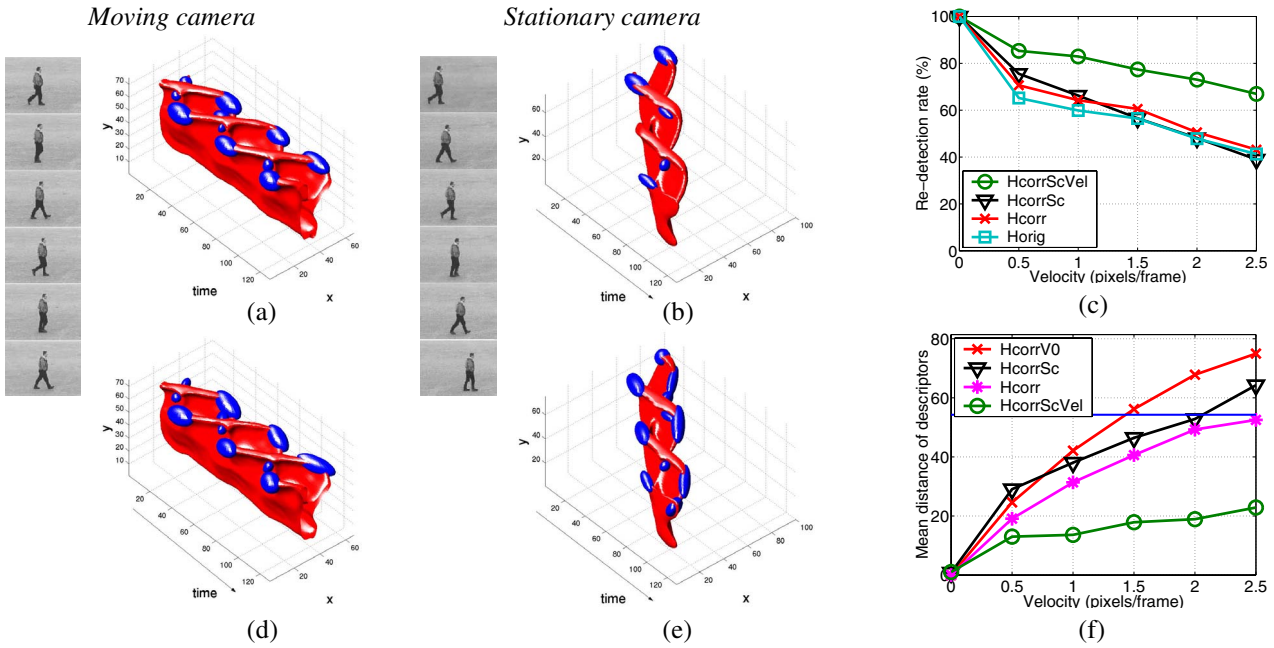

Figure 1. Detection and evaluation of space-time interest points for sequences with different camera motion. (a),(d): Spatio-temporal patterns of a walking person are shown by 3D plots (up-side-down) for manually stabilized camera and (b),(e): stationary camera. Interest points (ellipsoids) are detected using the original method without velocity adaptation in (a)-(b) and with the proposed method for iterative velocity adaptation in (d)-(e). (c): Repeatability of interest points evaluated for different (simulated) velocities of the camera and different detection methods (see the text). (f): Stability of descriptors under synthetic Galilean transformations. The stability is evaluated as the mean distance of jet descriptors of corresponding features.

3. Evaluation

In this section, we evaluate the influence of velocity adaptation on the detection and applications of interest points. To get an intuitive idea about the effect of velocity adaptation in practice, consider the two sequences of a walking person in Figure 1 recorded with a stabilized camera and a stationary camera. As can be seen from the spatio-temporal plots, the space-time structure of sequences as well as the result of interest point detection without velocity adaptation using [5] is highly influenced by the motion of the camera (see Figures 1a-b). The velocity-adapted interest points detected according to Section 2.3, on the other hand, have roughly corresponding positions and shapes in both sequences (see Figures 1d-e).

3.1. Repeatability

To evaluate the correspondence of interest points in sequences with different amounts of camera motion, we compared the repeatability of interest points computed for thirteen sequences of human actions (see Figure 2) which were Galilei-transformed according to a set of horizontal velocities $v_x = \{0.5, 1.0, 1.5, 2.0, 2.5\}$ using bilinear interpolation.[2] The positions of detected points were then transformed to the original sequence according to $p_{v_x} = G^{-1} p'_{v_x}$ and matched to the positions of the points detected in the original sequence. To evaluate different methods of feature detection, we detected space-time interest points using (Horig): Maxima of the space-time operator H (2) without neither scale nor velocity adaptation; (Hcorr): Maxima of the velocity-corrected operator H_c (8) without adaptation; (HcorrSc): Maxima of H_c with iterative adaptation of the spatio-temporal scales only [5]; (HcorrScVel): Maxima of H_c in combination with iterative scale and velocity adaptation according to Section 2.3.

As can be seen from the results in Figure 1c, the repeatability of velocity-adapted interest points (*HcorrScVel*) is stable for different values of velocities and outperforms the repeatability of three other methods without iterative velocity adaptation. By comparing the results of *Horig* and *Hcorr*, we also observe a slightly better repeatability of features detected using the velocity-corrected operator (*Hcorr*). To restrict the number of compared detectors, we will only use the velocity-corrected features in the following evaluation.

[2] Velocity $v_x = 1.5$ roughly corresponds to the velocity of walking persons in our image sequences.

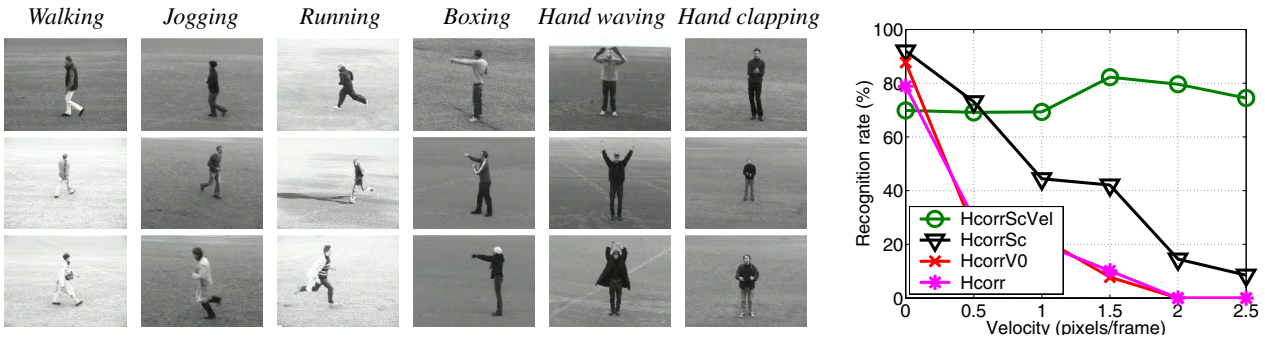

Figure 2. (Left): Examples of sequences from the action database [12]. All sequences have spatial resolution 160×120 pixels and contain 120 frames (25fps.) on average. (Right): Results of action recognition for different velocities of the camera.

3.2. Stability of descriptors

Image values in the neighborhoods of interest points can be used to describe and to compare local events in video [5]. Whereas similar events are expected to have similar descriptors, the values of any non-trivial type of descriptor will depend on the shape of the point neighborhoods and, hence, the relative motion of the camera.

To evaluate the stability of image descriptors under Galilean transformation, each interest point was described by a local jet of order four $j = (L_x, L_y, L_t, ..., L_{tttt})$. Gaussian spatio-temporal derivatives $L_{x^\alpha y^\beta t^\gamma} = \partial_x^\alpha \partial_y^\beta \partial_t^\gamma L(\cdot,\ \Sigma)$ were evaluated at positions of interest points with Σ obtained according to (i) iterative method of velocity adaptation, (ii) "one-step" velocity adaptation corresponding to the first step of adaptation method in Section 2.3 or (iii) no velocity adaptation. The first approach is truly invariant under velocity transformations and is natural when computing image descriptors for velocity-adapted features (*HcorrScVel*). The second approach is less demanding in terms of computations, at the cost of approximative invariance to velocity transformations. Such an approach is natural to combine with features detected without iterative velocity adaptation (*HcorrSc, Hcorr*). Finally, the case without velocity adaptation corresponds to jet descriptors computed with separable filter kernels. This method in combination with *Hcorr* features will be here denoted as *HcorrV0*.

The Euclidean distance between descriptors has been computed for all pairs of matched points obtained from the matching method and image sequences used in the previous section. The mean distance for different values of velocities and different methods is shown in Figure 1f. As can be seen, the descriptors of velocity-adapted interest points *HcorrScVel* are more stable than the descriptors of either features with one-step velocity adaptation or without velocity adaptation. Moreover, a comparison to the mean distance between all pairs of points in all the sequences (horizontal line in Figure 1f) also indicates a higher degree of discriminability of velocity-adapted interest points compared to interest points detected by the three other methods.

3.3. Application to recognition

Space-time interest points can be used for representing and recognizing motion patterns, such as human actions. A companion paper [12] presents an evaluation of recognizing human actions on a larger database. In this paper, we investigate the dependency of recognition on relative motions of the camera. To compare two image sequences represented in terms of interest points, we use a greedy matching strategy and repeatedly select and remove pairs of points (one from each sequence) with the minimum Euclidean distance evaluated for corresponding jet descriptors. The distance between two image sequences is then defined as the sum of the individual distances of the N strongest feature matches. Action recognition is achieved using Nearest Neighbor Classifier (NNC), where a sequence is considered as correctly recognized if its action coincides with the action of the nearest neighbor sequence of the training set.

In the first experiment, we used 192 sequences with six human actions (see Figure 2) taken with a stationary camera for training. For recognition, we used image sequences of human actions as in the previous experiments in Sections 3.1 and 3.2 with different subjects in the training set and in the test set. The motion of the camera was simulated by Galilei-transforming every test sequence according to a set of velocities $v_x = \{0.5, 1, 1.5, 2, 2.5\}$.

From the results presented in Figure 2(right) it follows that the recognition performance for velocity-adapted features (*HcorrScVel*) is rather stable for different velocities of the camera. The performance, however, decreases significantly for increasing camera velocity when using features with one-step velocity adaptation (*HcorrSc*, *Hcorr*) or features without velocity adaptation (*HcorrV0*). Iterative velocity adaptation, however, makes the features less discriminative and, hence, results in slightly worse performance in the case when the camera motions in the training and in

Figure 3. Image sequences of walking people recorded in scenes with complex backgrounds and a manually stabilized camera.

the test sets coincide ($v_x = 0$). This can be explained by the fact that velocity is an important cue when discriminating between, for example, a walking and a running person. Since velocity adaptation cancels this information from the local descriptors, the decreased performance of *HcorrScVel* is not surprising. Hence, the stability with respect to velocity transformations is here achieved at the cost of a slight decrease in the recognition performance when the relative velocity in the training set and the test set is the same.

In the second experiment, we aimed at recognizing walking actions in a city scenario (see Figure 3) while using the same training set as in the previous setting. For this purpose, 27 test sequences were recorded with a manually stabilized camera; hence, the velocities of the camera in the test and training sets were different. The results of recognition are presented with confusion matrices in Figure 4. As can be seen, the performance of the *HcorrScVel* features is significantly better (85.2%) than the performance of all the alternative methods without iterative velocity adaptation. This confirms the importance of velocity adaptation which was argued both in the theory and in the evaluation presented earlier. Moreover, these results also indicate the advantage of using space-time interest points for recognition in complex scenario where segmentation prior to recognition might be difficult. More discussion and experimental results on this topic can be found in [4]. Finally the results presented here could be further improved, for example, by a more efficient classification scheme such as SVM instead of NCC as presented in [12].

	Walk	Jog	Run	Box	Hclp	Hwav
Walk	**85.2**	3.7	11.1	0.0	0.0	0.0

HcorrScVel

	Walk	Jog	Run	Box	Hclp	Hwav
Walk	**25.9**	0.0	33.3	11.1	3.7	25.9

HcorrSc

	Walk	Jog	Run	Box	Hclp	Hwav
Walk	**18.5**	0.0	37.0	29.6	14.8	0.0

Hcorr

	Walk	Jog	Run	Box	Hclp	Hwav
Walk	**11.1**	0.0	63.0	22.2	3.7	0.0

HcorrV0

Figure 4. Results of classifying sequences with walking people in complex scenes using training sequences with different camera motion and different classes of actions. The results are shown for different types of features.

4. Summary

We have presented a method for adapting local space-time features to local Galilean transformations of video patterns. Adapted features provide a video representation that is stable under varying relative velocities of the camera. The stability of such a representation has been evaluated experimentally in terms of the repeatability of local features as well as by an application to recognition of human actions in situations with unknown camera motion.

References

[1] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc. ICCV*, pages 726–733, 2003.

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

[3] B. Jähne, H. Haußecker, and P. Geißler. 2. signal processing and pattern recognition. In *Handbook of Computer Vision and Applications*, page Chapter 13. Academic Press, 1999.

[4] I. Laptev. *Local Spatio-Temporal Image Features for Motion Interpretation*. PhD thesis, Department of Numerical Analysis and Computer Science (NADA), KTH, S-100 44 Stockholm, Sweden, 2004. ISBN 91-7283-793-4.

[5] I. Laptev and T. Lindeberg. Space-time interest points. In *Proc. ICCV*, pages 432–439, 2003.

[6] I. Laptev and T. Lindeberg. Velocity-adaptation of spatio-temporal receptive fields for direct recognition of activities: An experimental study. *IVC*, 22(2):105–116, February 2004.

[7] T. Lindeberg. Time-recursive velocity-adapted spatio-temporal scale-space filters. In *Proc. ECCV*, volume 2350 of *LNCS*, pages I:52–67. Springer, 2002.

[8] T. Lindeberg, A. Akbarzadeh, and I. Laptev. Galilean-corrected spatio-temporal interest operators. In *Proc. ICPR*, Cambridge, U.K., 2004.

[9] T. Lindeberg and J. Gårding. Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure. *IVC*, 15(6):415–434, June 1997.

[10] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *DARPA Image Understanding Workshop*, pages 121–130, 1981.

[11] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. ECCV*, volume 2350 of *LNCS*, pages I:128–142. Springer, 2002.

[12] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Proc. ICPR*, Cambridge, U.K., 2004.

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