Recognizing Human Actions: A Local SVM Approach[*]

Christian Schüldt   Ivan Laptev   Barbara Caputo
Computational Vision and Active Perception Laboratory (CVAP)
Dept. of Numerical Analysis and Computer Science
KTH, SE-100 44 Stockholm, Sweden
{crilla, laptev, caputo}@nada.kth.se

Abstract

Local space-time features capture local events in video and can be adapted to the size, the frequency and the velocity of moving patterns. In this paper we demonstrate how such features can be used for recognizing complex motion patterns. We construct video representations in terms of local space-time features and integrate such representations with SVM classification schemes for recognition. For the purpose of evaluation we introduce a new video database containing 2391 sequences of six human actions performed by 25 people in four different scenarios. The presented results of action recognition justify the proposed method and demonstrate its advantage compared to other relative approaches for action recognition.

1. Introduction

Applications such as surveillance, video retrieval and human-computer interaction require methods for recognizing human actions in various scenarios. Typical scenarios include scenes with cluttered, moving backgrounds, nonstationary camera, scale variations, individual variations in appearance and cloth of people, changes in light and view point and so forth. All of these conditions introduce challenging problems that have been addressed in computer vision in the past (see [1, 11] for a review).

Recently, several successive methods for learning and recognizing human actions directly from image measurements have been proposed [6, 3, 4, 15, 7]. When using image measurements in terms of optic flow or spatio-temporal gradients, however, these measurements and therefore the results of recognition may depend on the recording conditions such as position of the pattern in the frame, spatial resolution and relative motion with respect to the camera. Moreover, global image measurements can be influenced by motions of multiple objects and variations in the background. Whereas these problems can be solved in principle by external mechanisms for spatial segmentation and/or camera stabilization, such mechanisms might be unstable in complex situations. This motivates the need of alternative video representations that are stable with respect to changes of recording conditions.

In this paper we demonstrate that action recognition can be achieved using local measurements in terms of spatio-temporal interest points (local features) [9]. Such features capture local motion events in video and can be adapted to the size, the frequency and the velocity of moving patterns, hence, resulting in video representations that are stable with respect to corresponding transformations.

In spatial recognition, local features have recently been combined with SVM in a robust classification approach [12]. In a similar manner, here, we explore the combination of local space-time features and SVM and apply the resulting approach to the recognition of human actions. For the purpose of evaluation we introduce a new video database and present results of recognizing six types of human actions performed by 25 different people in different scenarios.

2. Representation

To represent motion patterns we use local space-time features [9] which can be considered as primitive events corresponding to moving two-dimensional image structures at moments of non-constant motion (see Figure 1).

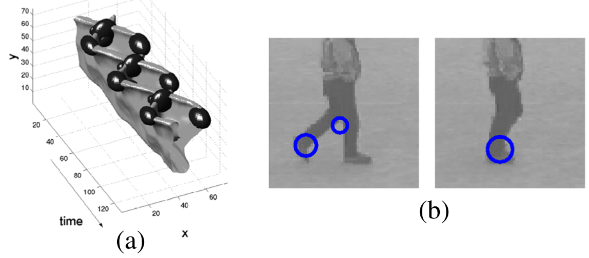


Figure 1. Local space-time features detected for a walking pattern: (a) 3-D plot of a spatio-temporal leg motion (up side down) and corresponding features (in black); (b) Features overlaid on selected frames of a sequence.

[*] The support from the Swedish Research Council is gratefully acknowledged.

To detect local features in image sequence $f(x, y, t)$, we construct its scale-space representation $L(\cdot, \sigma^2, \tau^2) = f * g(\cdot, \sigma^2, \tau^2)$ using Gaussian convolution kernel $g = \exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/2\tau_l^2)/\sqrt{(2\pi)^3\sigma_l^4\tau_l^2}$. We compute the second-moment matrix using spatio-temporal image gradients $\nabla L = (L_x, L_y, L_t)^T$ within a Gaussian neighborhood of each point

$$\mu(\cdot;\, \sigma^2, \tau^2) = g(\cdot;\, s\sigma^2, s\tau^2) * \left(\nabla L(\nabla L)^T\right) \qquad (1)$$

and define positions of features by local maxima of $H = \det(\mu) - k\, \text{trace}^3(\mu)$ over (x, y, t). The spatio-temporal neighborhood of features in space and time is then defined by spatial and temporal scale parameters (σ, τ) of the associated Gaussian kernel. As shown in [9], the size of features can be adapted to match the spatio-temporal extent of underlying image structures by automatically selecting scales parameters (σ, τ). Moreover, the shape of the features can be adapted to the velocity of the local pattern, hence, making the features stable with respect to different amounts of camera motion [10]. Here we use both of these methods and adapt features with respect to scale and velocity to obtain invariance with respect to the size of the moving pattern in the image as well as the relative velocity of the camera.

Spatio-temporal neighborhoods of local features contain information about the motion and the spatial appearance of events in image sequences. To capture this information, we compute spatio-temporal jets

$$\boldsymbol{l} = (L_x, L_y, L_t, L_{xx}, ..., L_{tttt}) \qquad (2)$$

at the center of each feature using normalized derivatives $L_{x^m y^n t^k} = \sigma^{m+n}\tau^k(\partial_{x^m y^n t^k} g) * f$ computed using selected scale values (σ^2, τ^2) [9]. To enable invariance with respect to relative camera motions, we also warp the neighborhoods of features using estimated velocity values prior to computation of $\boldsymbol{l}$ (see [8] for more details).

K-means clustering of descriptors $\boldsymbol{l}$ in the training set gives a vocabulary of primitive events h_i. The numbers of features with labels h_i in a particular sequence define a feature histogram $\boldsymbol{H} = (h_1, ..., h_n)$. We use such histograms as one alternative representation when recognizing motions in image sequences.

3. Classification: Support Vector Machines

Support Vector Machines (SVMs) are state-of-the-art large margin classifiers which have recently gained popularity within visual pattern recognition ([13, 14] and many others). In this section we provide a brief review of the theory behind this type of algorithm; for more details we refer the reader to [5, 12].

Consider the problem of separating the set of training data $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots (\boldsymbol{x}_m, y_m)$ into two classes, where $\boldsymbol{x}_i \in \Re^N$ is a feature vector and $y_i \in \{-1, +1\}$ its class label. If we assume that the two classes can be separated by a hyperplane $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$ in some space $\mathcal{H}$, and that we have no prior knowledge about the data distribution, then the optimal hyperplane is the one which maximizes the margin [12]. The optimal values for $\boldsymbol{w}$ and b can be found by solving a constrained minimization problem, using Lagrange multipliers $\alpha_i (i = 1, \ldots m)$.

$$f(\boldsymbol{x}) = \text{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right) \qquad (3)$$

where α_i and b are found by using an SVC learning algorithm [12]. Those $\boldsymbol{x}_i$ with nonzero α_i are the “support vectors”. For $K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y}$, this corresponds to constructing an optimal separating hyperplane in the input space $\Re^N$.
Based on results reported in the literature, in this paper we use the kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \exp\{-\gamma\chi^2(\boldsymbol{x}, \boldsymbol{y})\}$ [2] for histogram features $\boldsymbol{H}$, and for local features we use the kernel $K_L(\boldsymbol{L}_h, \boldsymbol{L}_k) = 1/2[\hat{K}(\boldsymbol{L}_h, \boldsymbol{L}_k) + \hat{K}(\boldsymbol{L}_k, \boldsymbol{L}_h)]$, with

$$\hat{K}(\boldsymbol{L}_h, \boldsymbol{L}_k) = \frac{1}{n_h}\sum_{j_h=1}^{n_h} \max_{j_k=1,\ldots n_k}\{K_l(\boldsymbol{l}_{j_h}, \boldsymbol{l}_{j_k})\} \qquad (4)$$

where $\boldsymbol{L}_i = \{\boldsymbol{l}_{j_i}\}_{j=1}^{n_i}$ and $\boldsymbol{l}_{j_i}$ is a jet descriptor of interest point j in sequence i and

$$K_l(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{-\rho\left(1 - \frac{\langle \boldsymbol{x} - \boldsymbol{\mu_x} | \boldsymbol{y} - \boldsymbol{\mu_y}\rangle}{||\boldsymbol{x} - \boldsymbol{\mu_x}|| \cdot ||\boldsymbol{y} - \boldsymbol{\mu_y}||}\right)\right\}, \qquad (5)$$

where $\boldsymbol{\mu_x}$ is the mean of $\boldsymbol{x}$ (consider [13] for more details).

4. Experiments

SVM classification combined with motion descriptors in terms of local features (LF) and feature histograms (HistLF) define two novel methods for motion recognition. In this section we evaluate both methods on the problem of recognizing human actions and compare the performance to other approaches using alternative techniques for representation and/or classification.

4.1. Experimental setup

For the evaluation, we recorded a video database containing six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors $s1$, outdoors with scale variation $s2$, outdoors with different clothes $s3$ and indoors $s4$ (see Figure 2). Currently the database contains 2391 sequences. All sequences were taken over homogeneous backgrounds with a static camera with 25fps frame rate. The sequences were downsampled to

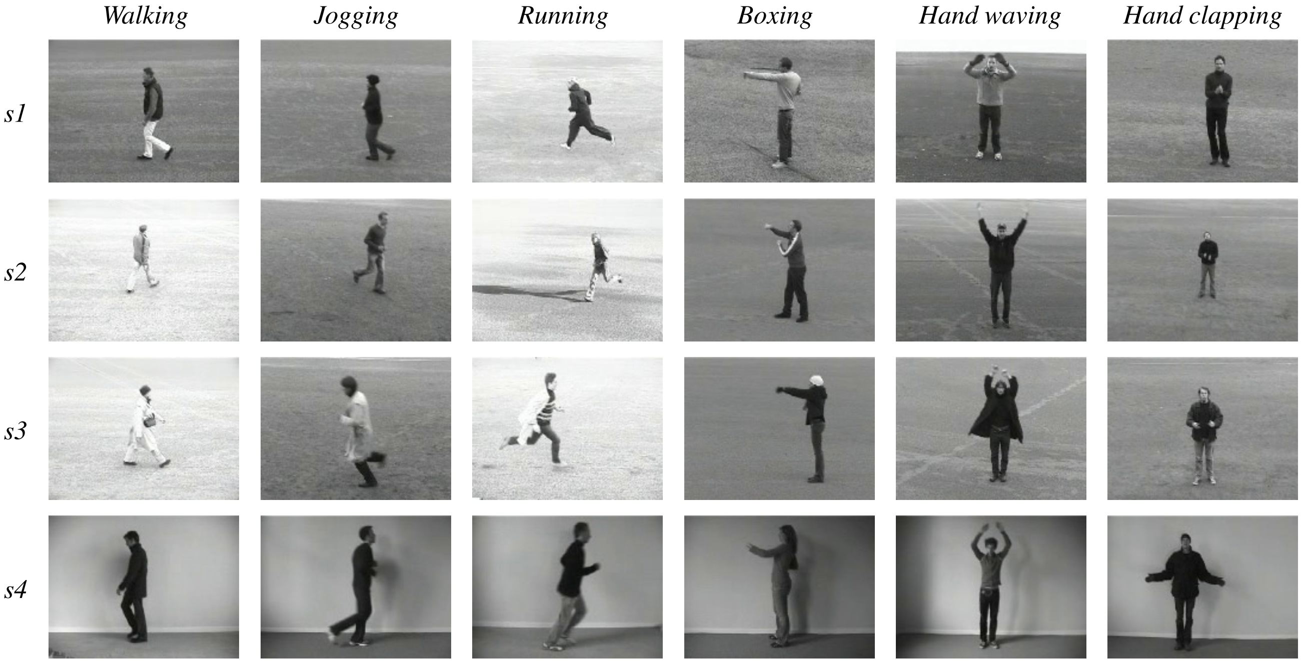


Figure 2. Action database (available on request): examples of sequences corresponding to different types of actions and scenarios.

the spatial resolution of 160×120 pixels and have a length of four seconds in average. To the best of our knowledge, this is the largest video database with sequences of human actions taken over different scenarios.

All sequences were divided with respect to the subjects into a training set (8 persons), a validation set (8 persons) and a test set (9 persons). The classifiers were trained on a training set while the validation set was used to optimize the parameters of each method. The presented recognition results were obtained on the test set.

4.2. Methods

We compare results of combining three different representations and two classifiers. The representations are *i)* local features described by spatio-temporal jets $\boldsymbol{l}$ (2) of order four (LF), *ii)* 128-bin histograms of local features (HistLF), see Section 2 and *iii)* marginalized histograms of normalized spatio-temporal gradients (HistSTG) computed at 4 temporal scales of a temporal pyramid [15]. In the latest approach we only used image points with temporal derivative higher than some threshold which value was optimized on the validation set.

For the classification we use *i)* SVM with either local feature kernel [13] in combination with LF or SVM with χ^2 kernel for classifying histogram-based representations HistLF and HistSTG, *ii)* nearest neighbor classification (NNC) in combination with with HistLF and HistSTG.

4.3. Results

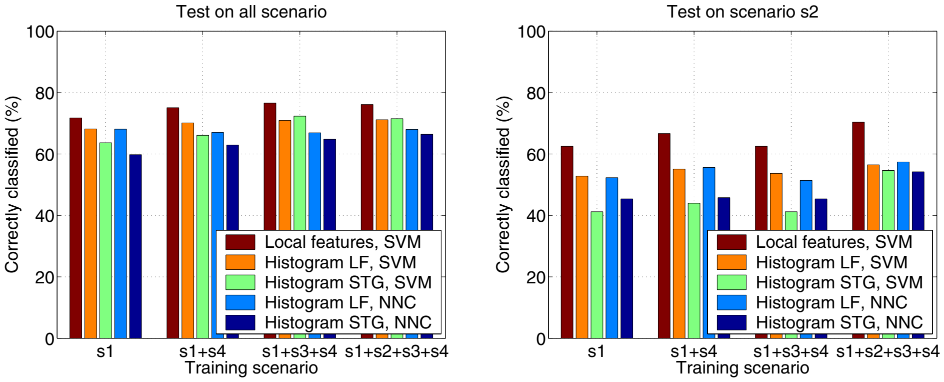
Figure 3(top) shows recognition rates for all of the methods. To analyze the influence of different scenarios we performed training on different subsets of $\{s1\}$, $\{s1, s4\}$, $\{s1, s3, s4\}$ and $\{s1, s2, s3, s4\}$. It follows that LF with local SVM gives the best performance for all training sets while the performance of all methods increases with the number of scenarios used for training. Concerning histogram representations, SVM outperforms NNC as expected, while HistLF gives a slightly better performance than HistSTG.

Figure 3(bottom) shows confusion matrices obtained with LF+SVM method. As can be seen, there is a clear separation between leg actions and arm actions. The most of confusion occurs between jogging and running sequences as well as between boxing and hand clapping sequences. We observed similar structure for all other methods as well.

Scenario with scale variations ($s2$) is the most difficult one for all methods. Recognition rates and the confusion matrix when testing on $s2$ only are shown in Figure 3(right).

4.4. Matching of local features

A necessary requirement for action recognition using the local feature kernel in Equation (5) is the match between corresponding features in different sequences. Figure 4 presents a few pairs of matched features for different sequences with human actions. The pairs correspond to features with jet descriptors $\boldsymbol{l}_{j_h}$ and $\boldsymbol{l}_{j_k}$ selected by maximizing



	Walk	Jog	Run	Box	Hclp	Hwav
Walk	**83.8**	16.2	0.0	0.0	0.0	0.0
Jog	22.9	**60.4**	16.7	0.0	0.0	0.0
Run	6.3	38.9	**54.9**	0.0	0.0	0.0
Box	0.7	0.0	0.0	**97.9**	0.7	0.7
Hclp	1.4	0.0	0.0	35.4	**59.7**	3.5
Hwav	0.7	0.0	0.0	20.8	4.9	**73.6**

Confusion matrix, all scenarios (LF+SVM)

	Walk	Jog	Run	Box	Hclp	Hwav
Walk	**100.0**	0.0	0.0	0.0	0.0	0.0
Jog	66.7	**33.3**	0.0	0.0	0.0	0.0
Run	13.9	69.4	**16.7**	0.0	0.0	0.0
Box	0.0	0.0	0.0	**97.2**	2.8	0.0
Hclp	0.0	0.0	0.0	36.1	**58.3**	5.6
Hwav	0.0	0.0	0.0	25.0	5.6	**69.4**

Confusion matrix, $s2$ scenario (LF+SVM)

Figure 3. Results of action recognition for different methods and scenarios. (top,left): recognition rates for test sequences in all scenarios; (top,right): recognition rates for test sequences in $s2$ scenario; (bottom,left): confusion matrix for Local Features + SVM for test sequences in all scenarios; (bottom,left): confusion matrix for Local Features + SVM for test sequences in $s2$ scenario;

the feature kernel over j_k in Equation (4). As can bee seen, matches are found for similar parts (legs, arms and hands) at moments of similar motion. The locality of descriptors allows for matching of similar events in spite of variations in clothing, lighting and individual patterns of motion. Due to the local nature of features and corresponding jet descriptors, however, some of the matched features correspond to different parts of (different) actions which are difficult to distinguish based on local information only. Hence, there is an obvious possibility for improvement of our method by taking the spatial and the temporal consistency of local features into account.

The locality of our method also allow for matching similar events in sequences with complex non-stationary backgrounds as illustrated in Figure 5. This indicates that local space-time features could be used for motion interpretation in complex scenes. Successful application of local features for action recognition in unconstrained scenes with moving heterogeneous backgrounds has recently been presented in [8].

4.5. Discussion

Confusion between walking and jogging as well as between jogging and running can partly be explained by high similarities of these classes (running of some people may appear very similar to the jogging of the others).

Global motion of subjects in the database is a strong cue for discriminating between the leg and the arm actions when using histograms of spatio-temporal gradients (HistSTG). This information, however, is (at least partly) canceled when representing the actions in terms of velocity-adapted local features. Hence, LF and HistLF representations can be expected to give similar recognition performance disregarding global motion of the person relative to the camera [10].

As can be seen from Figure 3(top,right), the performance of local features (LF) is significantly better than the performance of HistSTG for all training subsets that do not include sequences with scale variations ($s2$). This indicates the stability of recognition with respect to scale variations in image sequences when using local features for action representation. This behavior was expected from the scale-adaptation of features discussed in Section 2.

5. Summary

We have demonstrated how local spatio-temporal features can be used for representing and recognizing motion patterns such as human actions. By combining local features with SVM we derived a novel method for motion

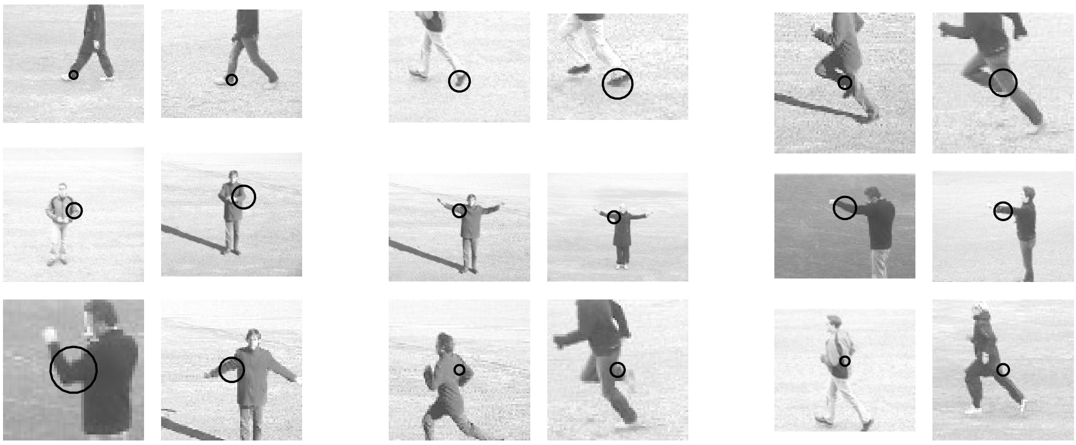


Figure 4. Examples of matched features in different sequences. (top): Correct matches in sequences with leg actions; (middle): Correct matches in sequences with arm actions; (bottom): false matches.

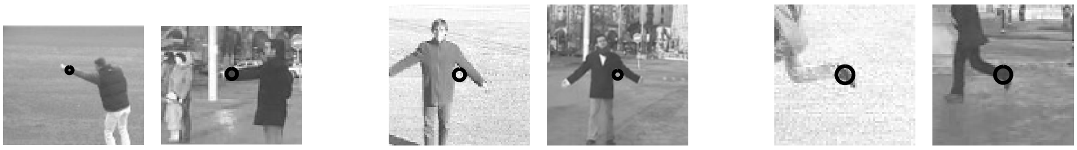


Figure 5. Examples of matching local features for pairs of sequences with complex non-stationary backgrounds.

recognition that gives high recognition performance compared to other relative approaches. For the purpose of evaluation we also introduced a novel video database that to the best of our knowledge is currently the largest database of human actions.

Representations of motion patterns in terms of local features have advantages of being robust to variations in the scale, the frequency and the velocity of the pattern. We also have indications that local features give robust recognition performance in scenes with complex non-stationary backgrounds and plan to investigate this matter in future work. Whereas local features have been treated independently in this work, the spatial and the temporal relations between features provide additional cues that could be used to improve the results of recognition. Finally, using the locality of features, we also plan to address situations with multiple actions in the same scene.

References

[1] J. Aggarwal and Q. Cai. Human motion analysis: A review. *CVIU*, 73(3):428–440, 1999.

[2] S. Belongie, C. Fowlkes, F. Chung, and J. Malik. Spectral partitioning with indefinite kernels using the nyström extension. In *Proc. ECCV*, volume 2352 of *LNCS*, page III:531 ff. Springer, 2002.

[3] M. Black and A. Jepson. Eigentracking: Robust matching and tracking of articulated objects using view-based representation. *IJCV*, 26(1):63–84, 1998.

[4] O. Chomat and J. Crowley. Probabilistic recognition of activity using local appearance. In *Proc. CVPR*, pages II:104–109, 1999.

[5] N. Cristianini and J. Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge UP, 2000.

[6] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. In *Proc. CVPR*, pages 928–934, 1997.

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

[8] 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.

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

[10] I. Laptev and T. Lindeberg. Velocity adaptation of space-time interest points. In *Proc. ICPR*, Cambridge, U.K., 2004.

[11] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *CVIU*, 81(3):231–268, March 2001.

[12] V. Vapnik. *Statistical Learning Theory*. Wiley, NY, 1998.

[13] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *Proc. ICCV*, pages 257–264, 2003.

[14] L. Wolf and A. Shashua. Kernel principal angles for classification machines with applications to image sequence interpretation. In *Proc. CVPR*, pages I:635–640, 2003.

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