Local Descriptors for Spatio-temporal Recognition*

Ivan Laptev and Tony Lindeberg

Computational Vision and Active Perception Laboratory (CVAP),
Dept. of Numerical Analysis and Computing Science,
KTH, S-100 44 Stockholm, Sweden

Abstract. This paper presents and investigates a set of local space-time descriptors for representing and recognizing motion patterns in video. Following the idea of local features in the spatial domain, we use the notion of space-time interest points and represent video data in terms of local space-time events. To describe such events, we define several types of image descriptors over local spatio-temporal neighborhoods and evaluate these descriptors in the context of recognizing human activities. In particular, we compare motion representations in terms of spatio-temporal jets, position dependent histograms, position independent histograms, and principal component analysis computed for either spatio-temporal gradients or optic flow. An experimental evaluation on a video database with human actions shows that high classification performance can be achieved, and that there is a clear advantage of using local position dependent histograms, consistent with previously reported findings regarding spatial recognition.

1 Introduction

When performing recognition from spatial or spatio-temporal images, the definition of the underlying image representation is of crucial importance for subsequent recognition. During recent years there has been a substantial progress on recognition schemes that are based on either local or global image features. In particular, the use of view-based approaches in terms of receptive field responses [10] has emerged as a highly promising approach for visual recognition.

When performing recognition, global methods are conceptually simple to implement. For complex scenes with occlusions and multiple moving objects, however, such methods require a complementary segmentation step, which may be non-trivial to achieve in practice. In this respect, local approaches have an interesting potential, while requiring a complementary matching step between the local features in the model and the data. For a recognition scheme to be invariant to size changes in the image domain as well as temporal phenomena

* 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.

W.J. MacLean (Ed.): SCVMA 2004, LNCS 3667, pp. 91–103, 2006.

that occur with different speed, it is natural to require the image descriptors to be invariant to spatial and temporal scale changes. Similarly, in order to handle unknown relative motions between the objects and the camera, invariance to local Galilean transformations can be expected to be a highly useful property.

In the area of motion-based recognition, a large number of different schemes have been developed based on various combinations of visual tasks and image descriptors; see e.g. the monograph by [24] and the survey paper by [7] for overviews of early works. Concerning more recent approaches, [1, 25] performed tracking and recognition using principal component analysis and parameterized models of optic flow. [8] presented a related approach using Zernike polynomial expansions of optic flow. [2] recognized human actions against a static background by computing templates of temporal differences and characterizing the resulting motion masks in terms of moments. [3, 26] recognized activities using probabilistic models of spatio-temporal receptive fields while [13] extended this approach to histograms of locally velocity-adapted receptive fields. Another statistical, non-parametric approach for motion recognition in terms of temporal multiscale Gibbs models was proposed by [5]. [4] presented a recognition scheme in terms of positive and negative components of stabilized optic flow in spatio-temporal volumes.

Space-time interest points [11] have recently been proposed to capture local events in video. Such points have stable locations in space-time and provide a potential basis for part-based representations of complex motions in video. The subject of this paper, is to study different ways of defining local space-time descriptors associated with such interest points and to use these descriptors for subsequent recognition of spatio-temporal events and activities. The approach can hence be seen as an extension of previous interest point based spatial recognition approaches [17, 19] into space-time.

In previous works in the spatial domain, it has been shown that the use of automatic scale selection allows for the computation of scale invariant image descriptors [14, 17, 19, 6], and that the SIFT descriptor [17], which can be seen as a scale-adapted position dependent histogram of spatial gradient vectors, is very powerful for spatial recognition [20]. Moreover, histograms of spatial or spatio-temporal derivatives have been shown to allow for spatial and spatio-temporal recognition [22, 26]. For handling perspective as well as Galilean image deformations, affine shape adaptation [16, 19] and velocity adaptation [21, 15, 13] have been demonstrated to be useful mechanisms.

In this paper, we shall combine and connect these types of mechanisms into new types of powerful spatio-temporal image descriptors. Specifically, we shall compare local space-time descriptors at interest points in terms of various combinations of N-jets, optic flow, principal component analysis as well as local histograms with or without spatial dependency. We will show that such local descriptors allow for matching of spatio-temporal events and activities between image sequences. The performance will be measured by evaluating classification rates on a video database with different types of human activities.

2 Spatio-temporal Interest Points

Following [11], let us adopt a local interest point approach for capturing spatio-temporal events in video data. Consider an image sequence f and construct a spatio-temporal scale-space representation L by convolution with a spatio-temporal Gaussian kernel $g(x, y, t;\ \sigma, \tau) = 1/(2\pi\sigma^2\sqrt{2\pi\tau})\exp(-(x^2+y^2)/2\sigma^2 - t^2/2\tau^2)$ with spatial and temporal scale parameters σ and τ. Then, at any point $p = (x, y, t)$ in space-time define a spatio-temporal second-moment matrix μ as

$$\mu(p) = \int_{q\in\mathbb{R}^3} (\nabla L(q))(\nabla L(q))^T\, g(p-q;\ \sigma_i, \tau_i)\, dq, \tag{1}$$

where $\nabla L = (L_x, L_y, L_t)^T$ denotes the spatio-temporal gradient vector and $(\sigma_i = \gamma\sigma, \tau_i = \gamma\tau)$ are spatial and temporal integration scales with $\gamma = \sqrt{2}$. 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 maxima of the interest point operator $H = \det\mu - k(\mathrm{trace}\,\mu)^2 = \lambda_1\lambda_2\lambda_3 - k(\lambda_1+\lambda_2+\lambda_3)^3$ over (x, y, t) subject to $H \geq 0$ with $k \approx 0.005$.

Scale selection. To estimate the spatial and the temporal extents (σ_0, τ_0) of events, we maximize the following normalized feature strength measure over spatial and temporal scales [14, 11] at each detected interest point $p_0 = (x_0, y_0, t_0)$

$$(\sigma_0, \tau_0) = \underset{\sigma,\tau}{\mathrm{argmax}}(\sigma^2\tau^{1/2}(L_{xx} + L_{yy}) + \sigma\tau^{3/2}L_{tt})^2. \tag{2}$$

Velocity adaptation. Moreover, to compensate for (generally unknown) relative motion between the camera and the moving pattern, we perform velocity adaptation [21, 15, 13, 12] by locally warping the neighborhoods of each interest point with a Galilean transformation using image velocity u estimated by computing optic flow [18] at the interest point.

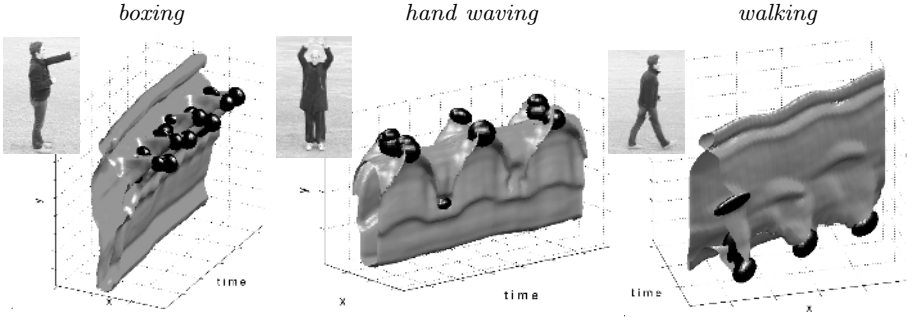

Fig. 1. Examples of scale and Galilean adapted spatio-temporal interest points. The illustrations show one image from the image sequence and a level surface of image brightness over space-time with the space-time interest points illustrated as dark ellipsoids.

Figure 1 shows a few examples of spatio-temporal interest points computed in this way from image sequences with human activities. As can be seen, the method allows us to extract scale-adaptive regions of interest around spatio-temporal events in a manner that is invariant to spatial and temporal scale changes as well as to local Galilean transformations.

3 Space-Time Image Descriptors at Interest Points

The subject of this section is to present a set of image descriptors to characterize the local space-time structure around interest points for subsequent recognition.

3.1 Image Measurements

As basis for defining spatio-temporal image descriptors, we shall make use of image measurements in terms of either:

- *Gaussian derivatives* up order four computed by applying scale normalized spatial and temporal derivatives [14] to the scale-space representation L

$$\mathcal{J}_{norm}\,(g(\cdot;\ \sigma_0, \tau_0) * f) = \{\sigma L_x, \sigma L_y, \tau L_t, \sigma^2 L_{xx}, \ldots, \sigma\tau^3 L_{ytttt}, \tau^4 L_{tttt}\} \quad (3)$$

  at locally adapted scale levels (σ_0, τ_0) as obtained from the scale selection step when detecting spatio-temporal interest points (see Fig. 2(left)). Specifically, we shall consider two types of Gaussian derivative descriptors; (i) the local (pointwise) N-jets [10] of order $N = 4$ evaluated at an interest point, and (ii) a multi-local gradient vector field obtained by evaluating the jet of order one at every point in a local neighborhood of an interest point.
- *Optic flow* computed from second-moment matrices around the space-time interest points, according to the method by Lukas and Kanade [18], and at locally adapted scale levels determined from the space-time interest points.

Due to the scale normalization in combination with the scale selection procedure, the N-jets will be scale invariant over both space and time [14, 11]. Scale invariance of the optic flow is achieved by computing the optic flow using scale-normalized Gaussian derivatives at locally adapted scale levels. For the purpose of Galilean invariance, both the N-jet and the optic flow are computed from locally warped space-time volumes as obtained from the velocity adaptation procedure.

There are a number of qualitative similarities as well as differences between these two types of image measurements: The N-jet contains a truncated encoding of the complete space-time image structure around the interest point, with an implicit encoding of the optic flow. By explicitly computing the optic flow, we obtain a representation that is invariant to local contrast in the image domain, at the cost of possible errors in the flow estimation step. In addition to the optic flow, the N-jet also encodes the local spatial structure, which may either help or distract the recognition scheme depending on the relation between the contents in the training and the testing data. Hence, it is of interest to investigate both types of image measurements.

3.2 Types of Image Descriptors

Then, we combine these measurements into image descriptors by considering:

- *Histograms* of either spatio-temporal gradients or optic flow computed at several scales. The histograms will be computed either for the entire neighborhood of an interest point, or over several ($M \times M \times M$) smaller neighborhoods around the interest point. For the latter case, here referred to as position dependent histograms, local coordinates are measured relative to the detected interest points and are used in the descriptors together with local image measurements (see Fig. 2(right)). Local measurements are weighted using Gaussian window function where we for simplicity marginalize the histograms and compute separable histograms over either the components of spatio-temporal gradients or the components of optic flow.
- *Principal component analysis* (PCA) of either optic flow or spatio-temporal gradient vectors (L_x, L_y, L_t) computed over local scale and velocity normalized spatio-temporal neighborhoods around the interest points. The principal components are computed from space-time interest points extracted from training data, and the data is then projected to a lower-dimensional space with D dimensions defined by the most significant eigenvectors (see Fig. 2(middle)).

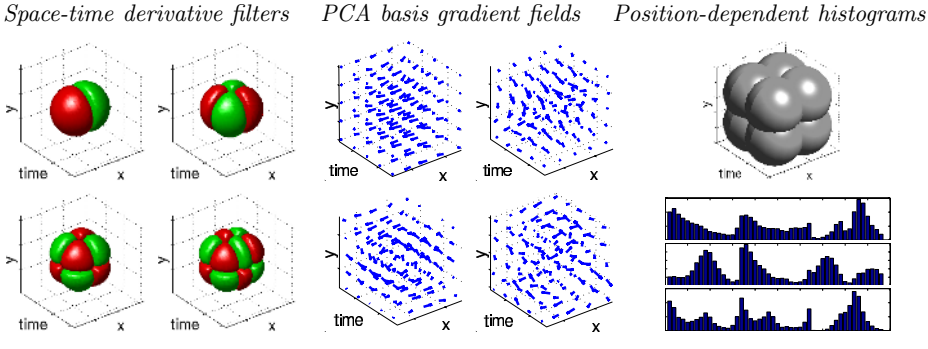

Fig. 2. (left) Examples N-jet components in terms of partial spatio-temporal derivative operators, here: $\partial_x, \partial_{xt}, \partial_{xyt}, \partial_{xxyt}$. (middle) Examples of basis vectors obtained by performing PCA on spatio-temporal gradients around space-time interest points. (right) Examples of position dependent histograms (bottom) computed using overlapping window functions (top).

3.3 Spatio-temporal Image Descriptors

By combining the abovementioned notions in different ways, we will consider the following types of space-time image descriptors:

1. N-jet of order 4 at a single scale, computed at (x_0, y_0, t_0) at scale (σ_0, τ_0).
2. Multi-scale N-jet of order 4, computed at all 9 combinations of 3 spatial scales $(\sigma_0/2, \sigma_0, 2\sigma_0)$ and 3 temporal scales $(\tau_0/2, \tau_0, 2\tau_0)$ at (x_0, y_0, t_0).

3. Local position dependent histograms of first-order partial derivatives.
4. Local position independent histograms of first-order partial derivatives.
5. Local position dependent histograms of optic flow.
6. Local position independent histograms of optic flow.
7. Local principal component analysis of optic flow.
8. Local principal component analysis of spatio-temporal gradients vectors.
9. Global histograms of first-order partial spatio-temporal derivatives computed over the entire image sequence using 9 combinations of 3 spatial scales and 3 temporal scales. This descriptor is closely related to [26] and is mainly considered here as a reference with respect to the previous global schemes for spatio-temporal recognition.

To obtain affine contrast invariance, the N-jets as well as the spatio-temporal gradient vectors are normalized to unit l_2-norm. For the principal component analysis of spatio-temporal gradient fields, the affine contrast normalization is performed at the level of scale normalized image volumes.

For an interest point detected at position (x_0, y_0, t_0) and scale (σ_0, τ_0), all histograms were computed at all 9 combinations of 3 spatial scales $(\sigma_0/2, \sigma_0, 2\sigma_0)$ and 3 temporal scales $(\tau_0/2, \tau_0, 2\tau_0)$. The global histograms were computed at combinations of spatial scales $\sigma \in \{1, 2, 4\}$ and temporal scales $\tau \in \{1, 2, 4\}$. When accumulating histograms of spatio-temporal gradients, only image points with L_t above a threshold were allowed to contribute. Moreover, all histograms were smoothed with a binomial filter and were normalized to unit l_1-norm. For the position dependent histograms (Descriptors 3 and 5), we initially consider $M = 2$ and evaluate the position dependent entities using Gaussian weighted window functions centered at $(x_0 \pm \alpha\sigma_0, y_0 \pm \alpha\sigma_0, t_0 \pm \beta\tau_0)$ with $\alpha = 1.5$ and $\beta = 1.5$. The spatial standard deviation of the Gaussian weighting function was 3σ and the temporal standard deviation 3τ. For the position dependent histograms, 16 bins were used for the components of the spatio-temporal gradients or the optic flow, while 32 bins were used for the position independent histograms. Thus, with $M = 2$ the position dependent histograms contain 9 scales $\times$ 8 positions $\times$ 3 derivatives $\times$ 16 bins = 3456 accumulator cells, and position independent histograms contain 9 scales $\times$ 3 derivatives $\times$ 32 bins = 864 cells. For the local principal component analysis, the gradient vectors and the optic flow were computed in windows of spatial extent $\pm 3\sigma$ and temporal extent $\pm 3\tau$ around the interest points. Prior to the computation of principal components using $D = 100$ dimensions, the gradient vectors and the optic flow were resampled to a $9 \times 9 \times 9$ grid using trilinear interpolation.

These descriptors build upon several previous works. The use of the N-jet for expressing visual operations was proposed by [10] and the first application to spatio-temporal recognition was presented in [3]. The use of histograms of receptive field responses goes back to [22, 26], and the use of PCA for optic flow was proposed by [1]. The use of complementary position information in histograms is closely related to the position dependency in the SIFT descriptor [17]. Recently, [9] added a local principal component analysis to the SIFT descriptor.

Hence, Descriptors 1, 2, 4, 6 and 7 can be seen as adaptations (and combinations) of previous approaches to space-time interest points, Descriptor 9 can

be seen as a variation of [26], while Descriptors 3, 5 and 8 are basically new, although with qualitative relations to some of the abovementioned works.

4 Matching

For recognizing spatio-temporal events and activities, we shall in this section explore the idea of matching space-time interest points attributed with image descriptors according to section 3.

Video database with human activities. For testing and evaluating our methods, we shall use a video database with 192 image sequences, with 8 people performing 6 types of actions ("walking", "jogging", "running", "boxing", "handclapping", "handwaving"). Each action is repeated four times by each subject, and for the cases of "walking", "jogging" and "running", there are two sequences where the subject is moving leftwards and two sequences with the subject moving rightwards (see figure 4 for a few sample image sequences for each type of activity).

Similarity/dissimilarity measures. For comparing descriptors h_1 and h_2 at different interest points, we consider the following similarity/dissimilarity measures:

- Normalized scalar product: $S(h_1, h_2) = \frac{\sum_i h_1(i) h_2(i)}{\sqrt{\sum_i h_1^2(i)} \sqrt{\sum_i h_2^2(i)}}$
- Euclidean distance: $E(h_1, h_2) = \sum_i (h_1(i) - h_2(i))^2$
- The χ^2-measure: $\chi^2(h_1, h_2) = \sum_i \frac{(h_1(i) - h_2(i))^2}{h_1(i) + h_2(i)}$

For descriptors in terms of N-jets, the feature vector h consists of Gaussian derivatives at an adaptively determined set of spatio-temporal scales. For the histogram descriptors, the feature vector is defined from the contents of all the accumulator cells. For PCA descriptors, the feature vector consists of projections of local image measurements onto D principal components.

Matching space-time interest points and image sequences. For matching local space-time features between image sequences, we will use a local greedy method. Given that the K strongest interest points have been computed in a training and a testing image, the similarity (dissimilarity) measure is evaluated for each pair of features. The pair with maximum similarity (or minimum dissimilarity) is matched and the corresponding features are removed from the training and testing sets. The procedure is repeated until no more feature pairs can be matched, either due to a threshold on similarity (dissimilarity) or lack of data.

Figure 3 shows a few examples of space-time interest points matched in this way for pairs of image sequences. As can be seen, many interest points identify the same type of events in different sequences disregarding variations in scale, cloth, lightning and complex backgrounds. To define similarity (dissimilarity) measures for pairs of sequences, we sum the individual similarities (dissimilarities) obtained from m best point matches. Of course, one could also consider

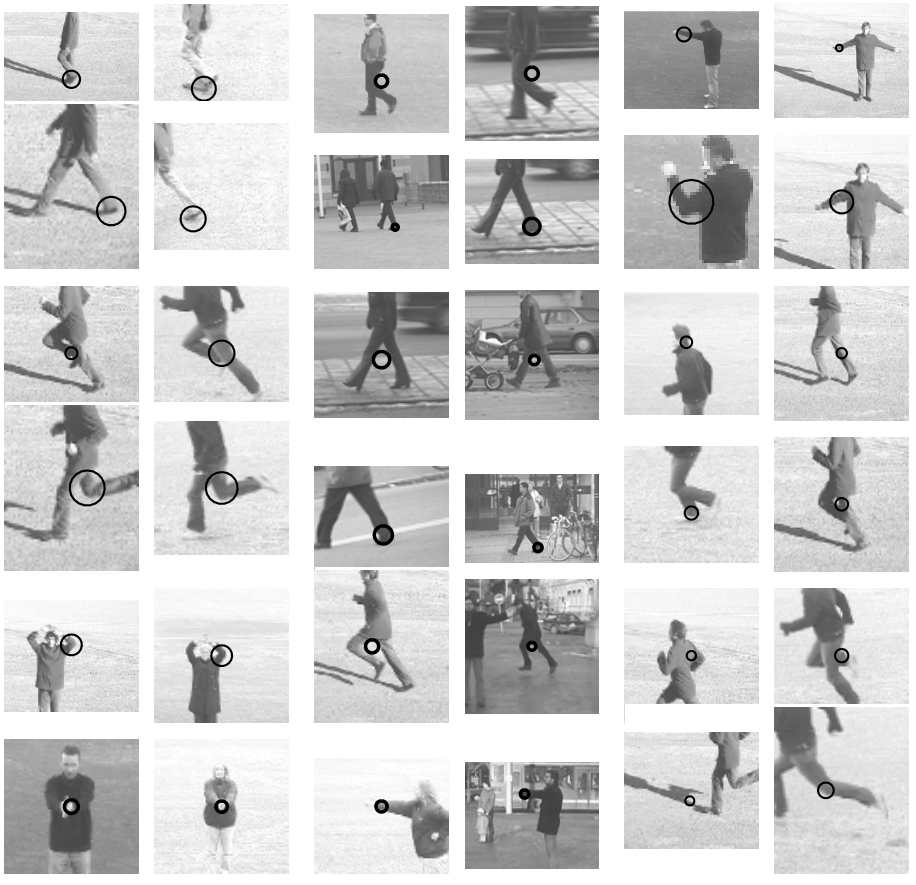

Fig. 3. Examples of point matches of space-time interest points using local image descriptors in terms of position dependent histograms of spatio-temporal gradient vectors

adding these measures transformed by a monotonically increasing function. Figure 4 shows a few examples of performing matching between image sequences in the database. As can be seen, the types of actions in the matched sequences (on the right) correspond to the actions in the test sequences (in the left column).

5 Experiments

To evaluate the performance of the different types of image descriptors, we will perform leave-X-out experiments for random perturbations of the database. In other words, the image sequences for X of the subjects will be removed from the database to be used as testing data, while the remaining image sequences will be used as training data. Then, for each image sequence in the test set, a best match is determined among all the image sequences in the training set. A match

Fig. 4. One sequence for each type of action in the database [23] with its best sorted matches (ordered from left to right). Here, all matches are correct except for the sequence with "running" for which the fourth best match is "jogging".

is regarded as correct if the activity of the best match in the training set agrees with the activity of the image sequence in the test set.

Figure 5 shows the result of computing classification rates in this way for the different types of image descriptors defined in section 3.3 using different types of similarity (dissimilarity) measures presented in section 4. For those descriptors that involve free parameters to be determined, we only show the results for the best parameters that were tested. The χ^2-measure is evaluated only for histogram-based descriptors.

As can be seen, for all three types of error metrics the position dependent histograms give the best results. Specifically, the position dependent histograms give better results than corresponding position independent histograms, both for spatio-temporal gradients and optic flow. Moreover, the position dependent histograms give better results than a principal component analysis of corresponding descriptors. In addition, most of our local methods give better re-

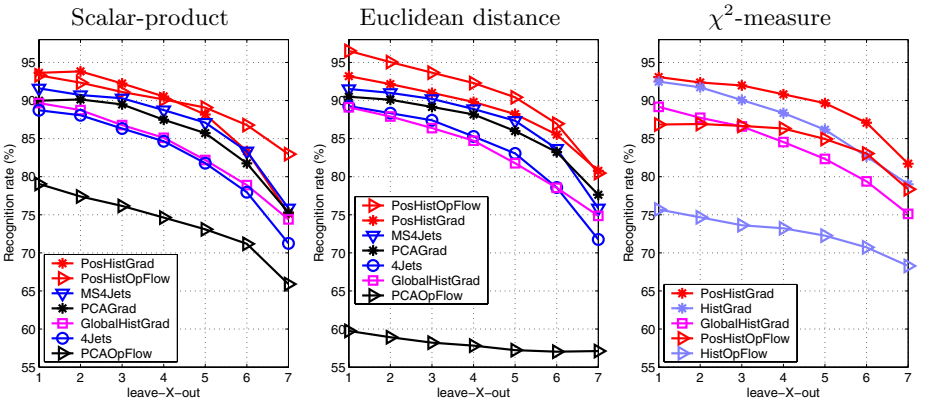

Fig. 5. Classification rates for different types of space-time image descriptors in leave-X-out experiments using either (a) normalized scalar product as similarity measure, (b) Euclidean distance as dissimilarity measure or (c) χ^2 dissimilarity measure. The results are averages over random permutations of the database. Specific comparison between position dependent histograms and position independent histograms for the χ^2 measure in (c) demonstrates the advantage of using position dependent histograms. Qualitatively similar results were obtained for the two other measures (left out here).

sults than the global histogram method. The multi-scale N-jet performs better than a principal component analysis of spatio-temporal gradient vectors or optic flow, and a multi-scale N-jet gives better results than a corresponding single-scale jet. We also evaluated recognition using N-jets of order two, but the performance of the forth order N-jets was slightly better. Position-dependent histograms with $M = 3$ were tested as well but did not give significant improvement.

Currently, the best results are obtained using position dependent histograms of optic flow in combination with a Euclidean distance measure. The second best method is a position dependent histogram of spatio-temporal gradients in combination with the normalized scalar product. The third best image descriptor out of these is the multi-scale N-jet, both for the case of using a normalized scalar

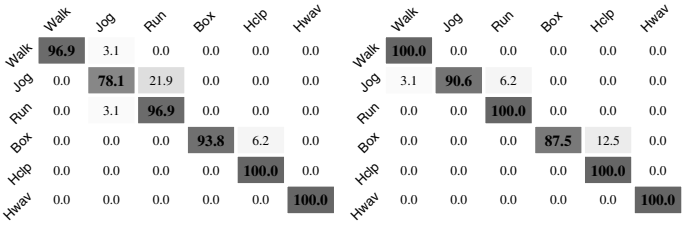

Fig. 6. Confusion matrices when classifying human activities with local descriptors in terms of position dependent histograms of spatio-temporal gradients (left) and optic flow (right)

product or the Euclidean distance as error metric. A conceptual advantage of the multi-scale N-jet is that it is essentially parameter free and gives a reasonable performance.

Figure 6 shows confusion matrices for the two best descriptors. As can be seen, most of the errors are due to mixing up the classes "jogging" and "walking" and mixing up the activities "boxing" and "handclapping", respectively. It is easy to explain why these types of misclassifications occur, since the activities "jogging/running" and "boxing/handclapping" contain similar types of local space-time events. For some of the subjects that were jogging and running in the video sequences, there is a somewhat fuzzy boundary between these two types of activities. If we merge "jogging" and "running" into a single class, the best overall recognition rate on this database increases from 96.4 % (position dependent histograms of optic flow) to 98.4 % (position dependent histograms of spatio-temporal gradients).

To conclude, these results show that it is possible to perform spatio-temporal recognition based on local space-time features. Moreover, considering that all these results have been computed using greedy matching of local image descriptors, there is potential for improvement by including spatio-temporal consistency constraints as well as overall motion descriptors into the recognition scheme.

6 Summary and Discussion

We have presented a set of image descriptors for representing local space-time image structures as well as a method for matching and recognizing spatio-temporal events and activities based on local space-time interest points.

By evaluating the proposed image descriptors on a video database with humans performing different types of actions, we have demonstrated that it is possible to obtain reasonably high recognition rates based on local space-time features. Specifically, we have shown that for this database two novel types of descriptors in terms of local position dependent histograms of either spatio-temporal gradients or optic flow give significantly better results than more traditional approaches of using global histograms, N-jets or principal component analysis of either optic flow or spatio-temporal gradients.

In on-going work, we are planning to extend the proposed histogram-based image descriptors to non-separable histograms as well as to evaluate Mahalanobis distances for matching. We will also perform evaluations on a larger database, including situations with multiple moving objects and cluttered backgrounds. Early results of recognizing human actions in scenes with complex and non-stationary backgrounds have been recently obtained and will be reported elsewhere. In this context, the locality of space-time features and of the proposed image descriptors is of key importance since it allows for matching of corresponding events in scenes with complex backgrounds as illustrated in figure 3.

Concerning other extensions, there is also potential for improving the current greedy point matching procedure to matching schemes which take the internal consistency of matching field as well as the overall motion patterns in the train-

ing data more explicitly into account. The replacement of the current nearest-neighbor classification scheme with the SVM classifier has recently been done in [23] and has shown additional increase in recognition performance.

References

1. M.J. Black and A.D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using view-based representation. *IJCV*, 26(1):63–84, 1998.
2. A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *IEEE-PAMI*, 23(3):257–267, 2001.
3. O. Chomat, J. Martin, and J.L. Crowley. A probabilistic sensor for the perception and recognition of activities. In *Proc. ECCV*, volume 1842 of *LNCS*, pages I:487–503. Springer, 2000.
4. A.A. Efros, A.C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc. ICCV*, pages 726–733, 2003.
5. R. Fablet and P. Bouthemy. Motion recognition using nonparametric image motion models estimated from temporal and multiscale co-occurrence statistics. *IEEE-PAMI*, 25(12):1619–1624, December 2003.
6. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, pages 264–271, Madison, Wisconsin, 2003.
7. D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.
8. J. Hoey and J.J. Little. Representation and recognition of complex human motion. In *Proc. CVPR*, pages I:752–759, 2000.
9. Y. Ke and R. Sukthankar. PCA-SIFT: A more disctinctive representation for local image descriptors. Technical Report IRP–TR–03–15, Intel, 2003.
10. J. J. Koenderink and A. J. van Doorn. Representation of local geometry in the visual system. *Biol. Cyb.*, 55:367–375, 1987.
11. I. Laptev and T. Lindeberg. Space-time interest points. In *Proc. ICCV*, pages 432–439, 2003.
12. I. Laptev and T. Lindeberg. Velocity adaptation of space-time interest points. In *Proc. of ICPR, to appear*, 2004.
13. I. Laptev and T. Lindeberg. Velocity-adapted spatio-temporal receptive fields for direct recognition of activities. *IVC*, 22(2):105–116, 2004.
14. T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):77–116, 1998.
15. T. Lindeberg. Time-recursive velocity-adapted spatio-temporal scale-space filters. In *Proc. ECCV*, volume 2350 of *LNCS*, pages I:52–67. Springer, 2002.
16. T. Lindeberg and J. Gårding. Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D structure. *IVC*, 15:415–434, 1997.
17. D.G. Lowe. Object recognition from local scale-invariant features. In *Proc. 7th Int. Conf. on Computer Vision*, pages 1150–1157, Corfu, Greece, 1999.
18. B. D. Lukas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Image Understanding Workshop*, 1981.
19. K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. ECCV*, volume 2350 of *LNCS*, pages I:128–142. Springer, 2002.
20. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proc. CVPR*, pages II: 257–263, 2003.

21. H.H. Nagel and A. Gehrke. Spatiotemporal adaptive filtering for estimation and segmentation of optical flow fields. In *ECCV'98*, pages 86–102, Freiburg, Germany, June 1998. Springer-Verlag.
22. B. Schiele and J. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *IJCV*, 36(1):31–50, 2000.
23. C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Proc. of ICPR, to appear*, 2004.
24. M. Shah and R. Jain, editors. *Motion-Based Recognition*. Kluwer, 1997.
25. Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities. *Computer Vision and Image Understanding*, 73(2):232–247, 1999.
26. L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *Proc. CVPR*, pages II:123–130, 2001.