

# Activity Recognition via Autoregressive Prediction of Velocity Distribution

Miha Peternel and Aleš Leonardis  
University of Ljubljana  
Faculty of Computer and Information Science  
Tržaška 25, SI-1001 Ljubljana, Slovenia  
{miha.peternel, ales.leonardis}@fri.uni-lj.si

## Abstract

We present a novel approach for view-based learning and recognition of motion patterns of articulated objects. We formulate the intervals of motion as a predictive model of local spatio-temporal receptive field activation. We compute local velocity distribution using a Bayesian approach, and then approximate the local velocity distribution in space and time using a set of Gaussian receptive fields. The activation sequence of receptive fields over time is modeled in a PCA subspace using linear auto-regression to arrive at a model of the motion pattern. Recognition is performed using the MDL principle. We test the approach on a number of human motion patterns to demonstrate the applicability of the proposed approach to simple action recognition and identification.

## 1 Introduction

The ability to learn and later recognize articulated activities with few assumptions about the object geometry, appearance, and the nature of the activity would have a number of applications in monitoring, video interpretation, video indexing, smart environments, and human-robot interactions.

Several models have been developed to enable recognition of motion patterns, trajectories, and consequently activities [7]. Standard methods either use a predefined geometrical model and try to estimate its parameters [8, 1, 2, 11], or attempt to model the observed motion directly [3, 16, 4, 5]. Recently, attempts have been made to summarize motion using local spatio-temporal features [14, 12]. It has been shown [12] that distribution density of local trajectories can be used for learning and recognition of cyclic human motion, however such representation is scale dependent and applies to cyclic motion only. Fablet *et al.* [6] have used causal probabilistic models to represent video dynamics. Jebara and Pentland [11] have used Gaussian probabilistic models to predict reaction from an interval of action. Agarwal and Triggs [1] have demonstrated that a mixture of regression models can be used as a predictor with a geometrical model. Bissacco *et al.* [2] have used subspace angles between autoregressive models of skeletal angles to recognize gait.

There have been few attempts of motion-based learning of activities without assuming a specific geometric model. Most of these approaches are based on modeling of optical flow fields [3, 16, 5], motion history [4] or on the modeling of manifolds of local features [14, 12]. We propose that instead of modeling the motion directly, prediction of local motion features can be exploited as a cue for activity recognition.

In this paper we present a novel view-based model of articulated motion based on the premise that we can recognize a motion pattern if we can predict it. On top of a robust local velocity estimate we build an auto-regressive predictor of the change in the future velocity distribution. We demonstrate that the proposed model can be used to learn and recognize individual patterns of locomotion as well as short non-cyclic actions.

The key novelties of our approach compared to prior works are: we model motion patterns based on the changes in their velocity distribution instead of using estimated optical flow; we apply a predictive model of a large number of local receptive fields based on co-activation in their spatio-temporal neighborhood to enable approximate modeling of arbitrary geometry in motion; we predict local co-activation patterns in a PCA subspace to decrease the dimensionality of the predictive function and thus enable learning from short action sequences; and finally, we recognize the motion patterns based on the utility of the predictor as measured by the MDL criterion. The approach is summarized in Figure 1.

The rest of the paper is organized as follows. In the next section we introduce Bayesian estimate of local velocity distribution. In Section 3 we describe how the velocity distribution can be approximated by Gaussian receptive fields. In Section 4 we outline autoregressive learning of receptive field activation patterns. In Section 5 we detail the proposed activity recognition method. In Section 6 we present experimental results. In Section 7 we conclude with a summary and outline work in progress.

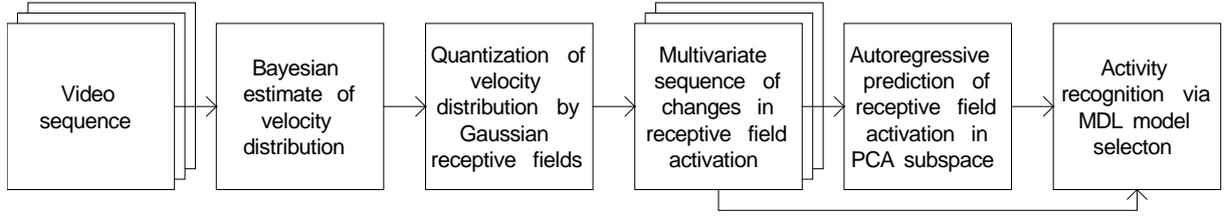


Figure 1: Activity recognition via autoregressive prediction of changes in the quantized velocity distribution.

## 2 Bayesian estimate of local velocity distribution

A complex motion pattern will undoubtedly appear as local velocity in a point of space-time. We compute a Bayesian estimate of local velocity distribution in a point of the view space based on a model proposed by Weiss and Fleet [15]:

$$P(v|I(x,t)) = \alpha P(v) * P(I(x,t)|v), \quad (1)$$

where  $v$  is velocity,  $x$  is view space position,  $t$  is time,  $I$  is image patch, and  $\alpha$  is an undetermined constant. The prior velocity likelihood  $P(v)$  is assumed to be Gaussian. Such a model results in a number of properties that are similar with how people perceive local motion [15], moreover we treat the aperture problem by estimating the ambiguity. A useful estimate of image likelihood  $P(I(x,t)|v)$  can be computed by expressing the probability as a function of the energy between two image patches  $I(x,t)$  and  $I(x-v,t-1)$  given a velocity estimate  $v$ . We will compute the energy as sum of square differences between the luminance values:

$$SSD(v) = \sum_{x \in W} (I(x,t) - I(x-v,t-1))^2, \quad (2)$$

where  $x$  runs over the image window  $W$  for which we estimate the probability. Assuming Gaussian noise with variance  $\sigma^2$ , we can then estimate the local image likelihood as:

$$P(I(x,t)|v) = \alpha \exp\left(-\frac{SSD(v)}{4\sigma^2}\right). \quad (3)$$

We quantize and constrain the velocity space to be in the expected range, and normalize the distribution by choosing  $\alpha$ , so that  $\sum_{v \in \mathcal{V}} P(v|I(x,t)) = 1$ . The observed distribution space  $\mathcal{D}$  is a compositum of the view space  $\mathcal{X}$  and the local velocity space  $\mathcal{V}$  over time, where each point  $(x,t)$  in space-time contributes an equal share of probability  $P(x,v;t)$  dispersed over the local velocity parameter  $v$ . See Figure 2(a) for an illustration. We have to compute (1) for each  $(x,v) \in \mathcal{D}$ .

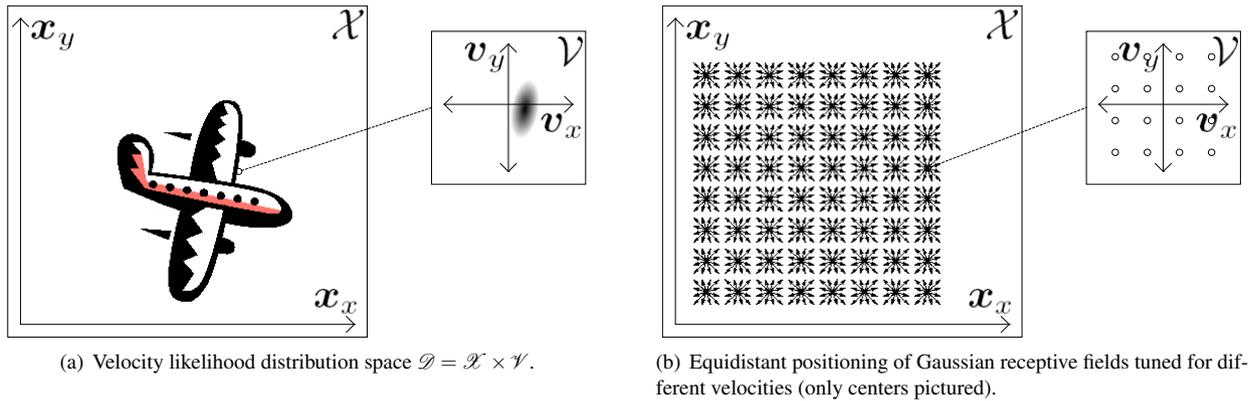


Figure 2: Gaussian receptive fields in the velocity likelihood distribution space

## 3 Approximating probability distribution using local receptive fields

Let  $\Phi_t$  represent the velocity probability distribution over  $\mathcal{D}$  at time  $t$ . Inferring  $\Phi_t$  from  $\Phi_{t-1}, \Phi_{t-2}, \dots$  is statistically an ill-posed problem in general due to a large number of parameters that have to be learned from short intervals. We propose  $\Phi_t$  be approximated using a large number of Gaussian bases  $\mathcal{N}_i$ , such that  $\sum_{i=1}^K k_i \mathcal{N}_i \approx \Phi_t$ . The evolution of coefficients  $k_i$  of the bases can then be modeled statistically using multivariate sequence prediction methods. In order to generalize the approximation we apply a four-dimensional grid of bases

$$\mathcal{N}_i(x,v;\mu_x,\mu_v) = \frac{1}{4\pi^2\sigma_x^2\sigma_v^2} \exp\left(-\frac{1}{2}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(v-\mu_v)^2}{\sigma_v^2}\right)\right) \quad (4)$$

placed equidistantly in  $\mathcal{X}$  and  $\mathcal{V}$  (see Figure 2(b) for an illustration).  $\Phi_t$  is approximated by a vector  $\phi_t$ , where

$$\Phi_t \approx \beta \sum_{i=1}^K \phi_{t,i} \mathcal{N}_i, \quad \phi_{t,i} = \Phi_t * \mathcal{N}_i. \quad (5)$$

Kernels  $\mathcal{N}_i$  thus correspond to local receptive fields in the velocity distribution, each tuned for a specific velocity and a specific view space position. To observe changes in the distribution, an exact decomposition is not strictly necessary, however, the kernels are spaced apart and dimensioned for a small overlap in order to make them more orthogonal which approximately preserves (5) to a constant factor  $\beta$ .

## 4 Learning spatio-temporal receptive field activation patterns

An articulated motion sequence can be represented by approximating the distributions through all the frames:  $\Phi_t \mapsto \phi_t$ . Thus we can model a video sequence by a multivariate sequence of vectors  $\phi_1, \phi_2, \dots, \phi_T$ , each vector representing the activation of Gaussian receptive fields in each frame. Alternatively we can represent an activity by approximating the change in the distribution by defining  $\delta_t = \phi_t - \phi_{t-1}$ . Such a sequence can be predicted by assuming a linear autoregressive model of order  $\tau$ :

$$\hat{\delta}_t = \sum_{i=1}^{\tau} A_i \delta_{t-i} + b. \quad (6)$$

Parameters  $A_i$  and  $b$  are chosen to minimize the error between  $\delta_t$  and  $\hat{\delta}_t$ . We use ARfit [13] algorithm to compute the parameters. Figure 3 illustrates autoregressive prediction of 30 active receptive fields in a locomotion sequence.

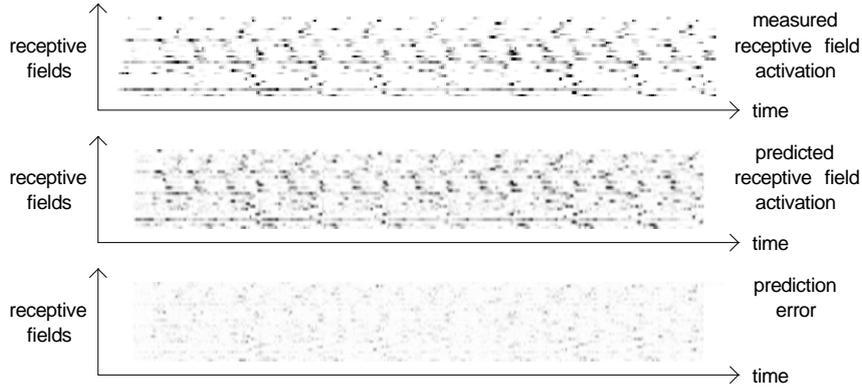


Figure 3: Measured receptive field activation sequence, autoregressive prediction and prediction error of 30 active receptive fields in a locomotion sequence.

If the number of receptive fields is large, it may be beneficial to compute a PCA subspace of  $\delta_t$  to constrain the dimensionality of the predictive model as well as to constrain the prediction to the active subset of the receptive fields and exploit their co-activation statistics. Given a matrix of principal eigenbases  $U$ , we predict the subspace sequence  $U \delta_t$  and compute the prediction:

$$\hat{\delta}_t = \sum_{i=1}^{\tau} U^T A_i U \delta_{t-i} + U^T b. \quad (7)$$

PCA bases define complex receptive fields (*i. e.* a linear combination of Gaussian receptive fields) which together predict local receptive field activation.

## 5 Recognition as prediction vs. measurement

Given an input video sequence (or an interval from the sequence) and a set of models  $\mathcal{M}_c = \{A_1^c, \dots, A_\tau^c, b^c, U^c\}$  we apply the Minimum Description Length [10] principle to choose the model in the following way. First, we compute the local velocity probability distribution in each frame. Secondly, the probability distribution in each timepoint is decomposed using a set of kernels corresponding to  $\mathcal{N}_i$  to produce a sequence of activation measurements  $\delta_1, \dots, \delta_T$ . We define the prediction error sequence as follows:

$$\epsilon_t^c = \delta_t - \hat{\delta}_t^c = \delta_t - U^{cT} b_i^c - \sum_{i=1}^{\tau} U^{cT} A_i^c U^c \delta_{t-i}. \quad (8)$$

We can formulate maximum *a posteriori* model selection probabilistically:

$$\arg \max_c P(\mathcal{M}_c | \delta) = \frac{P(\delta | \mathcal{M}_c) P(\mathcal{M}_c)}{P(\delta)}. \quad (9)$$

Assuming  $P(\delta)$  is the same for all models, and noting that  $P(\delta | \mathcal{M}_c) = P(\varepsilon^c)$ , (9) simplifies to:

$$\arg \max_c P(\varepsilon^c) P(\mathcal{M}_c) = \arg \max_c \log P(\varepsilon^c) + \log P(\mathcal{M}_c). \quad (10)$$

Let  $\mathcal{L}(\delta)$  represent the description length of the measurement and  $\mathcal{L}(\varepsilon | \mathcal{M}_c)$  the description length of the residual  $\varepsilon^c$  given a model  $\mathcal{M}_c$  of length  $\mathcal{L}(\mathcal{M}_c)$ . We relate  $\mathcal{L}$  to probability using the Shannon information measure:  $\mathcal{L}(d) = -\log_2 P(d)$ . Note that optimizing (10) equals optimizing:

$$\arg \min_c \mathcal{L}(\varepsilon | \mathcal{M}_c) + \mathcal{L}(\mathcal{M}_c). \quad (11)$$

We can choose the most efficient model to explain the data by minimizing MDL based objective function:

$$\arg \min_c \frac{\mathcal{L}(\varepsilon | \mathcal{M}_c) + \mathcal{L}(\mathcal{M}_c)}{\mathcal{L}(\delta)}. \quad (12)$$

Note that  $\mathcal{L}(\delta)$  is independent of  $c$ , we reintroduce it in (12) only to be able to interpret the results as relative bit savings. By assuming that the sequence vectors are conditionally independent, we can compute:

$$\mathcal{L}(\varepsilon | \mathcal{M}_c) = \sum_{t=\tau+1}^T -\log_2 P(\varepsilon_t^c), \quad \mathcal{L}(\delta) = \sum_{t=\tau+1}^T -\log_2 P(\delta_t). \quad (13)$$

Under the assumption that both the measurement and prediction error sequences are zero mean and normally distributed, and all model descriptions are equal in length Eq. (12) further simplifies to:

$$\arg \min_c \frac{\sum_{t=\tau+1}^T (\varepsilon_t^c)^2}{\sum_{t=\tau+1}^T (\delta_t)^2}. \quad (14)$$

Due to high dimensionality of both  $\delta_t$  and  $\varepsilon_t^c$  it is not feasible to model their distribution directly. Instead, we observe and model the distribution of their magnitude  $|\delta_t|$  and  $|\varepsilon_t^c|$  by computing their histograms from the test sequences. The resulting probabilities are used as priors for  $P(\delta_t)$  and  $P(\varepsilon_t^c)$ .

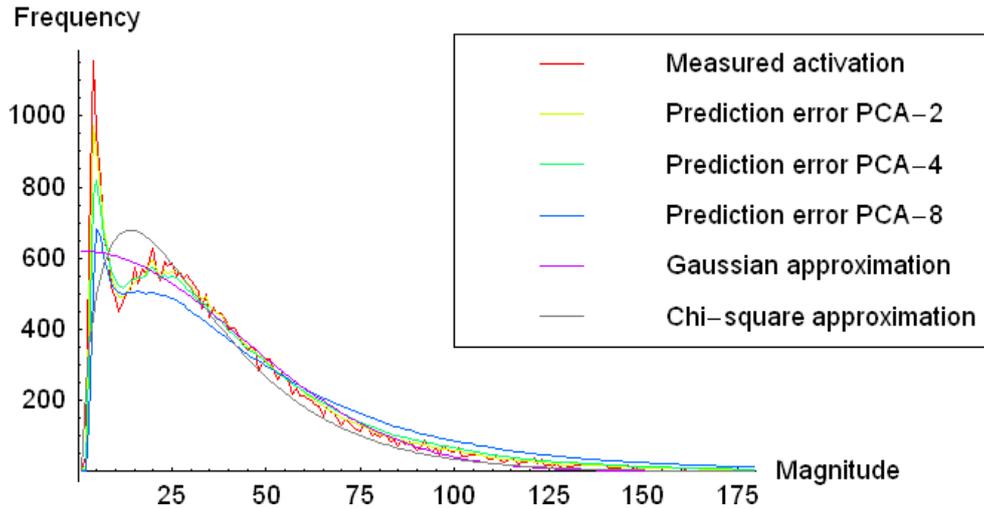


Figure 4: Distributions of activation vector magnitude and error vector magnitudes compared to Gaussian and Chi-square approximation.

Figure 4 presents actual distributions (as measured on 300 short action sequences) of activation vector magnitude  $|\delta_t|$ , error vector magnitudes  $|\varepsilon_t^c|$  using different numbers of PCA bases, a Gaussian approximation, and a Chi-square approximation. Note that the measured distributions are remarkably similar when the number of PCA bases is small. As the number of the bases increases the distribution flattens and the tail extends further right. The first (largest) mode is slightly offset from zero due to measurement noise, and the modes further right are likely the contributions of errors from the parts of sequences that are not accurately predicted by linear autoregression. Considering the nature of the valleys around the largest mode, a unimodal Gaussian approximation seems reasonable.

## 6 Experiments

We perform all the tests on MJPEG sequences downsampled 1 : 2 and the velocity estimated at every 4 \* 4 pixels using an 8 \* 8 patch and prior velocity likelihood distribution assuming standard deviation of 1.5 pixels. We use  $K = 1024$  receptive fields in a 8 \* 8 configuration with a distance of 28 pixels and standard deviation 8 in the view space  $\mathcal{X}$  and a 4 \* 4 configuration with a distance of 4 pixels/frame and standard deviation 1 in the velocity space  $\mathcal{V}$ . We use AR models of order  $\tau = 3$ .

### 6.1 Locomotion based identification

We perform two experiments on the CMU MoBo database [9]. We use 25 sequences of people walking on the treadmill. We divide each sequence in a learning half and a testing half. We learn 25 models of locomotion. In the first experiment, we classify the testing sequences using the proposed approach. The results are in Figure 5. As the number of PCA bases increases, the recognition performance initially increases towards 100%. The average of objective function Eq. (12) reaches its minimum at 14 PCA bases, where the average description length of the residuals is minimal. Note that the recognition performance around this point is 100%. With more bases, the models start to overfit, and the recognition performance finally deteriorates at more than 31 PCA bases. The identification performance is comparable to prior work by Peternel and Leonardis [12], however our model requires neither prolonged local tracking nor extraction of walking cycles.

In the second experiment, we vary the position and scale of the receptive field structure, to test the local monotonicity of the recognition function. We notice a clear extreme in  $\mathcal{X}$  for all cases, however changes in scale introduce slight fluctuations.

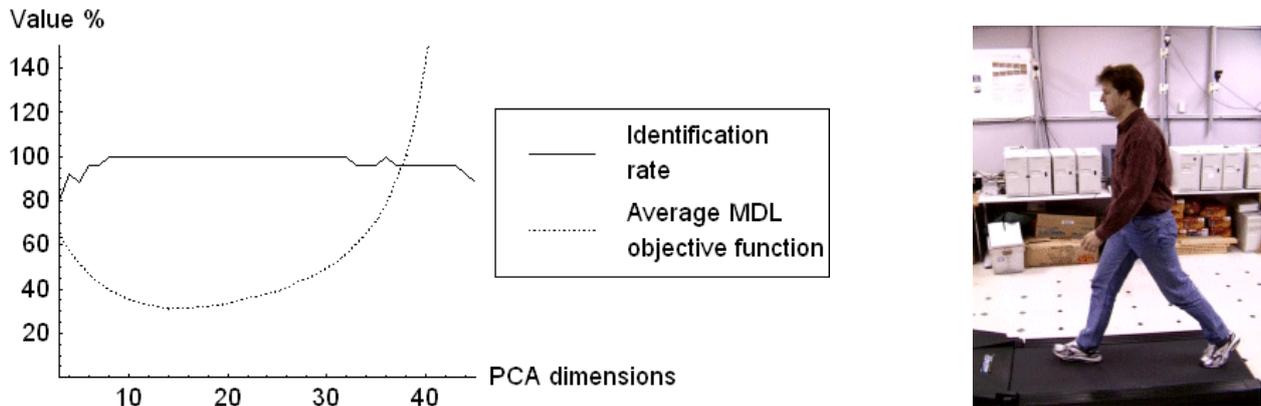


Figure 5: Locomotion pattern identification results on 25 fast-walk sequences.

### 6.2 Stand-sit disambiguation

For the third experiment we recorded a sequence of 5 people standing up and sitting down on the same chair 4-5 times. We learn models of all the instances of both actions. We perform leave-one-out classification to test the categorization and identification with two simple actions. The results are in Figure 6. The method correctly categorizes all the sequences when the number of PCA bases used is lower than 12.

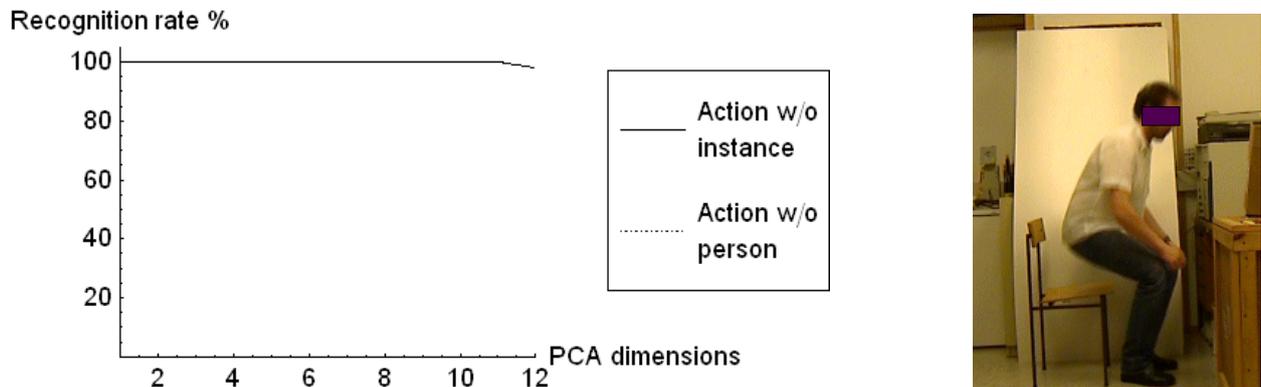


Figure 6: Categorization and identification on 48 stand-up & sit-down sequences of 5 people (recognition curves overlap).

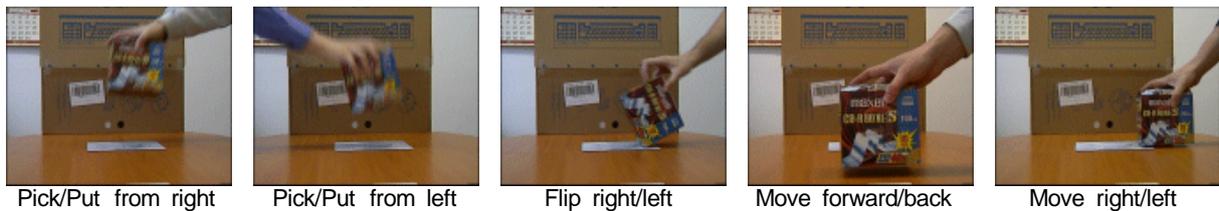


Figure 7: Actions performed by manipulating a box of CDs.

Method description	Error Likelihood Prior	Action Recognition	
		Leave-one-instance-out	Gaussian
Leave-one-person-out	Gaussian	271	90.3%
Leave-one-instance-out	Chi-square	280	93.3%
Leave-one-person-out	Chi-square	264	88.0%
Leave-one-instance-out	average histogram	263	87.7%
Leave-one-person-out	average histogram	246	82.0%

Table 1: Summary of action categorization on 300 short action sequences of 6 people.

ID	Action	1	2	3	4	5	6	7	8	9	10	Recog. Rate
1	pick from right	30	0	0	0	0	0	0	0	0	0	100.0%
2	put from right	0	28	0	2	0	0	0	0	0	0	93.3%
3	pick from left	1	0	29	0	0	0	0	0	0	0	96.7%
4	put from left	0	1	0	29	0	0	0	0	0	0	96.7%
5	flip right	0	0	0	0	28	0	0	0	1	1	93.3%
6	flip left	0	0	0	0	0	29	0	0	0	1	96.7%
7	move forward	0	0	0	0	0	0	29	1	0	0	96.7%
8	move backwards	0	1	0	0	0	0	1	28	0	0	93.3%
9	move right	1	0	0	0	1	0	0	0	28	0	93.3%
10	move left	0	1	0	0	0	2	0	0	0	27	90.0%
												Avg. 95.0%

Table 2: Confusion matrix for categorization of 10 actions from 300 sequences of 6 people (Leave-one-instance-out).

ID	Action	1	2	3	4	5	6	7	8	9	10	Recog. Rate
1	pick from right	30	0	0	0	0	0	0	0	0	0	100.0%
2	put from right	0	27	0	2	0	0	0	0	0	1	90.0%
3	pick from left	2	0	28	0	0	0	0	0	0	0	93.3%
4	put from left	0	4	0	26	0	0	0	0	0	0	86.7%
5	flip right	0	0	0	0	28	0	0	0	1	1	93.3%
6	flip left	0	0	0	0	0	29	0	0	0	1	96.7%
7	move forward	0	0	0	0	0	0	24	5	0	1	80.0%
8	move backwards	0	1	0	0	0	0	0	29	0	0	96.7%
9	move right	1	0	0	0	2	0	0	0	27	0	90.0%
10	move left	0	1	0	0	0	4	0	2	0	23	76.7%
												Avg. 90.3%

Table 3: Confusion matrix for categorization of 10 actions from 300 sequences of 6 people (Leave-one-person-out).

### 6.3 Manipulative action categorization

For the last set of experiments we recorded over 300 sequences of 10 simple actions performed by 6 people with a box of CDs (see Figure 7 for an illustration). The volunteers were instructed to perform each action at least 5 times in a natural way, and they did not see how the others performed the same actions. The approaches varied as did the exact locus of each action. Moreover, the volunteers were instructed to ignore the placement markers rather than to change their natural moves. The field of view contained a flat table area where the actions were performed, a box object, a pad and mostly one or two arms of the person performing the action. The background was only partially textured to introduce different levels of ambiguity in the velocity estimation.

We evaluate the performance of the proposed approach for the problem of learning and categorization of 10 short actions from 300 short video sequences. We trimmed each sequence to keep five activity-free frames before and after each action. We learn autoregressive models of order 3 in a PCA subspace dimensioned from 2–8. We use objective function Eq. (12) to select among models with different number of PCA bases.

We evaluate the action categorization in two ways, both with three error likelihood priors: Gaussian, Chi-square

and data-driven histogram based. We summarize the results in Table 1.

In the first experiment we categorize an instance of an action by ignoring its model while preserving the other instances of actions from the same person. The results for the Gaussian prior are in Table 2. To summarize the results, exactly 285 out of 300 video sequences are categorized correctly. In the second experiment we categorize an instance of an action by ignoring all the instances of all actions from the same person. The results for the Gaussian prior are in Table 3. Still, 271 of 300 sequences are categorized correctly, showing that the method is able to generalize actions from the exemplars of other people.

We repeat the experiments with the Chi-square prior and with the histogram based prior computed by accumulating distribution histograms over all available sequences. We already presented the distributions in Figure 4. Despite a more accurate modeling of error distribution by a histogram, recognition results are not improved, meaning that the average prior is not particularly useful for recognition of different activities. If the number of instances of each action were larger it might be possible to provide enough statistics for activity specific modeling of distribution.

## 7 Summary and Conclusions

We proposed a model for view-based learning and recognition of articulated motion sequences without specific assumptions about the geometry or cyclicity, and no requirement for local tracking.

The main contributions of this paper are a novel representation of articulated motion and the associated methods for learning and recognition of activity patterns from short video sequences, and a new database of short manipulative actions.

We have applied the proposed approach to learning and recognition of several types of articulated motion patterns: cyclic and non-cyclic motion of humans, as well as short actions performed by manipulating a box.

We demonstrated that even a rough linear prediction of the change in velocity distribution can be used with a MDL criterion to identify individual locomotion patterns and categorize simple actions. Experiments show that different levels of prediction accuracy are required for the problems of identification and categorization. Sufficiently accurate modeling of co-activation subspace is required for individual cyclic locomotion pattern identification, however weaker predictors attained with smaller dimensional subspaces are most useful for action categorization, whereas higher dimensional subspaces only help at disambiguation of some slightly more complex actions.

Current research is directed towards extending the method to activity localization in space and time, and other remaining problems in articulated motion recognition.

## Acknowledgements

This research has been supported in part by the following funds: Research program Computer Vision P2-0214 (RS), EU FP6-004250-IP project CoSy, EU FP6-511051-2 project MOBVIS, CONEX project, SI-A project.

## References

- [1] A. Agarwal and B. Triggs. Tracking articulated motion using a mixture of autoregressive models. In *ECCV*, volume 3, pages 54–65, 2004.
- [2] A. Bissacco, A. Chiuso, Yi Ma, and S. Soatto. Recognition of human gaits. In *CVPR 2001*, volume 2, pages 52–58, 2001.
- [3] M. J. Black, Y. Yacoob, and X. S. Ju. Recognizing human motion using parameterized models of optical flow. In *Motion-Based Recognition*, pages 245–269. Kluwer Academic Publishers, 1997.
- [4] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.
- [5] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV 2003*, volume 2, pages 726–733, October 2003.
- [6] R. Fablet, P. Bouthemy, and P. Pérez. Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval. *IEEE Transactions on Image Processing*, 11(4):393–407, 2002.
- [7] D. M. Gavrila. The visual analysis of human movement: A survey. *CVIU*, 73(1):82–98, 1999.
- [8] M. A. Giese and T. Poggio. Morphable models for the analysis and synthesis of complex motion patterns. *IJCV*, 38(1):59–73, 2000.
- [9] R. Gross and J. Shi. The CMU Motion of Body (MoBo) Database. Technical Report CMU-RI-TR-01-18, Carnegie Mellon University, June 2001.

- [10] P. Grünwald. *The Minimum Description Length Principle and Reasoning under Uncertainty, ILLC Dissertation Series DS 1998-03*. PhD thesis, CWI, the Netherlands, 1998.
- [11] T. Jebara and A. Pentland. Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour. In *ICVS '99*, pages 273–292, 1999.
- [12] M. Peternel and A. Leonardis. Visual learning and recognition of a probabilistic spatio-temporal model of cyclic human locomotion. In *ICPR*, volume 4, pages 146–149, 2004.
- [13] T. Schneider and A. Neumaier. ARFIT — a Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Trans. on Math. Software*, 27(1):58–65, 2001.
- [14] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, volume 3, pages 32–36, 2004.
- [15] Y. Weiss and D. J. Fleet. Velocity likelihoods in biological and machine vision. In *Probabilistic Models of the Brain: Perception and Neural Function*, pages 81–100. MIT, 2001.
- [16] Y. Yacoob and M. J. Black. Parameterized Modeling and Recognition of Activities. *CVIU*, 73(2):232–247, 1999.