

Temporal Self-Similarity for Appearance-Based Action Recognition in Multi-View Setups

Marco Körner and Joachim Denzler

Friedrich Schiller University of Jena, Computer Vision Group, Jena, Germany
{marco.koerner, joachim.denzler}@uni-jena.de, www.inf-cv.uni-jena.de

Abstract. We present a general data-driven method for multi-view action recognition relying on the appearance of dynamic systems captured from different viewpoints. Thus, we do not depend on 3d reconstruction, foreground segmentation, or accurate detections. We extend further earlier approaches based on *Temporal Self-Similarity Maps* by new low-level image features and similarity measures. *Gaussian Process* classification in combination with *Histogram Intersection Kernels* serve as powerful tools in our approach. Experiments performed on our new combined multi-view dataset as well as on the widely used IXMAS dataset show promising and competing results.

Keywords: Action Recognition, Multi-View, Temporal Self-Similarity, Gaussian Processes, Histogram-Intersection Kernel.

1 Introduction

The automatic recognition of actions from video streams states a very important problem in current computer vision research, as reflected by recent surveys[1]. A variety of possible applications—*e.g. Human-Machine Interaction*, surveillance, *Smart Environments*, entertainment, *etc.*—argues for the emerging relevance of this topic.

As monocular approaches rely on single-view images, they solely perceive 2d projections of the real world and discard important information. Hence, they are likely to suffer from occlusions and ambiguities. As a consequence, the majority of these methods use data-driven methods like *Space-Time Interest Points*[8] instead of model-based representations of the image content. In contrast, existing multi-view action recognition systems try to directly exploit 3d information, *e.g.* by reconstructing the scene or fitting anatomical models, resulting in a far higher complexity.

Having these observations in mind, we propose a method to recognize articulated actions, which meets the following demands: (i) it is designed to be general and not restricted to *human* action recognition, (ii) it avoids expensive dense 3d reconstruction, (iii) it is independent from the camera setup it was learned in, and (iv) it does not rely on foreground segmentation and exact localization.

The rest of this paper is structured as follows: in Sect. 2 we give a short introduction in theory of *Recurrence Plots* and *Temporal Self-Similarity Maps* and

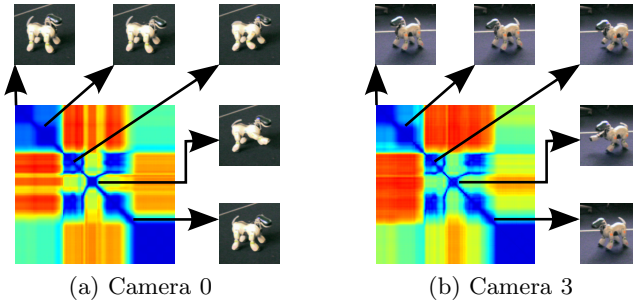


Fig. 1. Two SSMs obtained for a robot dog performing an `stand_kickright` action captured from different viewpoints. Action primitives induce similar local structures in the corresponding SSM even under changes of viewpoint, illumination, or image quality.

motivate their usage. We also suggest to extend the related approach of Junejo *et al.*[7] by new low-level features and distance metrics. Subsequently, Sect. 3 will present our approach to utilize SSMs for multi-view action recognition. Therefore we use a *Gaussian Process* classifier together with the *Histogram Intersection Kernel*, which has been shown to be more suitable for comparison of histograms. In Sect. 4, we show results of our approach on our own new multi-view action recognition dataset as well as on the widely used IXMAS dataset.

1.1 Related Work

Going through the related literature, methods for action recognition can be categorized into three groups: the first kind of approaches tries to reconstruct 3d information or trajectories from the scene[15] or augment these representations by a fourth time dimension[19,5]. Alternatively, relationships between action features obtained from different views are learned by applying transfer learning or knowledge transfer techniques[3,9]. The methods most related to our proposal try to directly model the dynamics of actions within a view-independent framework[7,2]. For a more extensive review about recent work on action recognition we refer to recent reviews[1,14].

2 Temporal Self-similarity Maps

To understand human actions and activities, observers take benefit of their prior knowledge of typical temporal and spatial recurrences in execution of actions. Besides all differences in execution, two actions can be perceived as being *semantically identical* if they share atomic *action primitives* in a similar frequency. Assuming those actions to be instances of deterministic dynamical systems—which can be modeled by differential equations—, Marwan *et al.*[11] presented an intensive discussion about their interpretation utilizing *Recurrence Plots* (RP). This work was further referenced for human gait analysis[2] and—due to their stability in the case of viewpoint changes—for cross-view action recognition[7].

Table 1. Semantic interpretations of patterns shown in SSMs introduced by recorded actions (interpretation of [11])

Pattern	Interpretation
Homogeneous areas	The corresponding atomic action represents a stationary process
Fading in corners	The recorded action represents a Non-stationary process
Periodic structures	The recorded action contains a cyclic/periodic motion
Isolated points	The recorded action contains an abrupt fluctuation
(Anti-) Diagonal straight lines	The recorded action contains different atomic actions with similar evolutionary characteristics in (reversed) time
Horiz. & vert. lines	No or slow change of states for a given period of time
Bow structures	The recorded action contains different atomic actions with similar evolutionary characteristics in reversed time with different velocities

Given a sequence $\mathbf{I}_{1:N} = \{I_1, \dots, I_N\}$ of images $I_i, 1 \leq i \leq N$, a *temporal Self-Similarity Map* (SSM) is generically defined as a square and symmetric matrix $\mathbf{S}_{f,d}^{\mathbf{I}_{1:N}} = [d(f(I_i), f(I_j))]_{i,j}$, $\mathbf{S}_{f,d}^{\mathbf{I}_{1:N}} \in \mathbb{R}^{N \times N}$ of pairwise similarities $d(\cdot, \cdot)$ between low-level image features $f(\cdot)$ computed independently for every sequence frame. In the literature, it has already been shown that SSMs preserve invariants of the dynamic systems they capture[12], they are stable wrt. different embedding dimensions[12,6], invariant under isometric transformations[12] and though not being invariant under projective or affine transformations, SSMs are heuristically shown to be stable under 3d view changes[7]. In Fig. 1, a robot dog performing a `stand_kickright` action was captured from two viewpoints with different illumination conditions. Apparently, atomic action primitives induce similar structures within the corresponding SSM. It can further be observed, that the local structure of these SSMs reflects the temporal relations between different system configurations over time, as summarized in Tab. 1.

2.1 Image Features

The choice for low-level image features $f(\cdot)$ is of inherent importance and has to suit the given scenario. In the following, we will discuss some possible alternatives.

Intensity Values. The simplest way to convert an image into a descriptive feature vector $f_{\text{int}}(I) \in \mathbb{R}^{M \cdot N}$ is to append its intensities, as proposed for human gait analysis[2]. While this is suitable for sequences with a single stationary actor, it yields large feature vectors and is very sensitive to noise and illumination changes.

Landmark Positions. Assuming to be able to track anatomical or artificial landmarks of the actor over time, their positions $f_{\text{pos}}(I) = (\mathbf{x}_0, \mathbf{x}_1, \dots), \mathbf{x}_i = (x_i, y_i, z_i)$, can be used to represent the current system configuration[7]. This is sufficient, as long as the tracked points are distributed over moving body parts, but it demands points to be able to be tracked continuously.

Table 2. Exemplary SSMs extracted from recordings of actions from the Aibo dataset using different low-level image features


























Feature	Action (performed in greeting pose)			
	greeting	scout right	stretch	dance1
Intensity				
HoG				
HoF				
Fourier				

Table 3. Exemplary SSMs extracted from the same `stand_dance1` action from the Aibo dataset using different similarity measures

Similarity Measure	Feature		
	HoG	HoF	Fourier
Euclidean Distance			
Normalized Cross-Correlation			
Histogram Intersection			

Histograms of Oriented Gradients have been shown to give good representations of shape for object detection. For this purpose, the image is subdivided into overlapping cells, where the distribution of gradient directions is approximated by a fixed-bin discretization. These certain local orientation histograms are normalized to the direction of the strongest gradient in order to obtain local rotation invariance. Appending those local gradient histograms gives the final descriptor $f_{\text{HoG}}(I) = (\mathbf{h}_0, \mathbf{h}_1, \dots)$, $\mathbf{h}_i = (n_i^0, n_i^1, \dots)$ [7].

Histograms of Optical Flows. When analyzing the displacements of each pixel between two succeeding frames, this *optical flow* field represents an early fusion of temporal dynamics. Building a global histogram over discretized flow orientations or appending histograms obtained from smaller subimages yield the HoF descriptor $f_{\text{HoF}}(I)$.

Fourier Coefficients. When computing the 2-dimensional discrete Fourier transform $\hat{a}_{k,l} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I_{m,n} \cdot e^{-2\pi i(\frac{mk}{M} + \frac{nl}{N})}$, $0 \leq k \leq M-1$, $0 \leq l \leq N-1$, $\hat{a}_{k,l} \in \mathbb{C}$ of an image patch I , the series of Fourier coefficients $[\hat{a}_{k,l}]$ contains spectral information up to a given cutoff frequency $0 \leq k \leq M_c-1$, $0 \leq l \leq N_c-1$ and inherently provides invariance against translation. Since the first Fourier coefficient $\hat{a}_{0,0}$ represents the mean intensity of the transformed image patch I , the Fourier coefficient descriptor $f_{\text{Fourier}} = (\hat{a}_{0,1}, \hat{a}_{0,2}, \dots, \hat{a}_{1,N_c-1}, \dots, \hat{a}_{M_c-1, N_c-1})$ is further invariant wrt. global illumination changes. By tuning the cutoff frequencies M_c, N_c , statistical noise can be suppressed as it is represented by higher-order frequencies. Since DFT can be implemented in parallel on modern GPU environments, these features can be computed very efficiently.

A qualitative comparison of these features extracted from different action classes is given in Tab.2. It can be seen that the HoF feature shows many abrupt changes, while the other SSMs contain more smooth transitions between the certain similarity values. The HoG feature seems to be more sensitive so temporal changes at small time scale, which could be explained by image noise

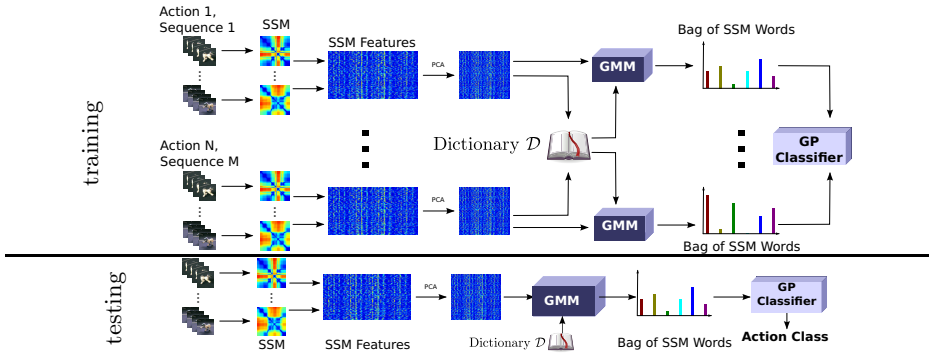


Fig. 2. Outline of the training and testing phase of our approach

and might harm the further processing. Hence, we further concentrate on using the proposed Fourier coefficients, since they are easily and fast to compute and provide some handy invariants by design.

2.2 Similarity Measures

Beside the choice for a suitable image representation $f(\cdot)$, the distance measure $d(\cdot, \cdot)$ plays an important role when computing self-similarities, as qualitatively compared in Tab. 3.

Euclidean Distances. The euclidean distance $d_{\text{eucl}}(\mathbf{f}_1, \mathbf{f}_2) = \|\mathbf{f}_1 - \mathbf{f}_2\|_2$ serves as a straightforward way to quantify the similarity between two image feature descriptors $\mathbf{f}_1 = f(I_1)$ and $\mathbf{f}_2 = f(I_2)$ of equal length, as proposed by [7]. While this is easy to compute, it might be unsuited for histogram data[10], since false bin assignments would cause large errors in the euclidean distance.

Normalized Cross-Correlation. From a signal-theoretical point of view, the image feature descriptors $\mathbf{f}_1, \mathbf{f}_2$ can be regarded as D -dimensional discrete signals of equal size. Then, the normalized cross-correlation coefficient $d_{\text{NCC}}(\mathbf{f}_1, \mathbf{f}_2) = \left\langle \frac{\mathbf{f}_1}{\|\mathbf{f}_1\|}, \frac{\mathbf{f}_2}{\|\mathbf{f}_2\|} \right\rangle \in [-1, 1]$ measures the cosine of the angle between the signal vectors \mathbf{f}_1 and \mathbf{f}_2 . Hence, this distance measure is independent from their lengths.

Histogram Intersection. The intersection $d_{\text{HI}}(\mathbf{h}_1, \mathbf{h}_2) = \sum_{i=0}^{D-1} \min(\mathbf{h}_{1,i}, \mathbf{h}_{2,i})$ of two histograms $\mathbf{h}_1, \mathbf{h}_2 \in \mathbb{R}^D$ was shown to perform better for codebook generation and image classification tasks[13]. In case of comparing normalized histograms, the histogram intersection distance is bounded by $[0, +1]$.

3 MVSSM Feature Extraction and Action Topic Model Learning and Classification

As mentioned before, SSMs obtained from videos capturing the identical action from different viewpoints share common patterns. Hence, local feature descriptors suitable for monitoring the structure of those patterns have to be developed

in order to use *Multi-View SSM* (MVSSM) representations for action recognition purposes in multi-view environments. Hooked on on the choice for features and the similarity measure used to create the SSM, self-similarity values are expected to become less reliable when moving away from the diagonal, as measuring the similarity gets more difficult. Junejo *et al.*[7] proposed to use a log-polar histogram of intensity gradients extracted on discrete positions at the main diagonal of the SSM to be analyzed, which yields a descriptor of dimension 88. The radius of this histogram, *i.e.* the temporal extend of interest, controls the amount of temporal information taken into account. As an extension, they constructed these histograms at different time scales to catch variations in executions.

Alternatively, we propose to extract 128-dimensional SIFT descriptors at keypoints equally distributed along the diagonal of fused multi-view SSMs. These are scale-invariant by design, as they examine and aggregate the image information on different scale spaces. To reduce the number of dimensions, we further apply PCA to the matrix of descriptor vectors.

Since the number of feature descriptors varies with the size of the SSM, *i.e.* the length of the sequence, and the density of keypoints used for extracting these features, we need to transform this set of features into a fixed-size representation. We used the widely popular *Bag of Visual Words* approach to assign the given action descriptors to representative prototypes identified by a custom cluster algorithm. Choosing an appropriate value for the number of prototypes, the obtained feature histograms are sparse and thus easy to distinguish. Fig. 2 outlines the training and testing phase of our system.

4 Experimental Evaluation

In order to evaluate our multi-view action recognition system, we firstly performed experiments on our own dataset. This dataset contains 10 sequences of each 56 predefined actions performed by SONY AIBO robot dogs simultaneously captured by six cameras.¹

In our general setup, the dimension of SIFT descriptors extracted along the SSM diagonal was reduced from 128 to 32 by applying PCA. Subsequently, all descriptors from all train sequences were clustered into a mixture of 512 Gaussians to create a *Bag of Self-Similarity Words* (BOSS Words). This is further used to represent each training sequence by a histogram of relative frequencies. These parameters heuristically show best results. While Junejo *et al.*[7] propose to employ a multiclass SVM, this yield a very high complexity in our case, as the AIBO dataset covers a relatively large number of classes to be distinguished. Hence, we use a *Gaussian Process* (GP) classifier combined with a *Histogram Intersection Kernel* $\kappa_{HIK}(\mathbf{h}, \mathbf{h}') = \sum_{i=0}^D \min(h_i, h'_i)$, $\mathbf{h}, \mathbf{h}' \in \mathbb{R}^D$, which can be evaluated efficiently, as recently shown by Rodner *et al.*[16] and Freytag *et al.*[4]. Recognition rates were obtained after 10-fold cross validation.

One of the most important questions concerning multi-view action recognition is the influence of the training and testing camera setup on the overall

¹ The complete dataset including labels, calibration data and background images is available at <http://www.inf-cv.uni-jena.de/JAR-Aibo>.

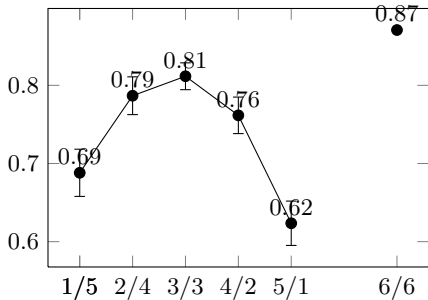


Fig. 3. Results obtained on Aibo dataset: average recognition rates for different $n_{\text{training}}/n_{\text{testing}}$ view partitions

Table 4. Results obtained on IXMAS dataset (cross-view evaluation)

Approach	Description	Rec.
our approach		79%
Junejo <i>et al.</i> [7]	HoG ¹	63%
Junejo <i>et al.</i> [7]	HoF ¹	67%
Junejo <i>et al.</i> [7]	HoG+HoF ¹	74%
Junejo <i>et al.</i> [7]	HoG+HoF ²	80%
Weinland <i>et al.</i> [17]	2d Silhouettes	58%
Farhadi <i>et al.</i> [3]	2d Silhouettes+OF	69%
Weinland <i>et al.</i> [18]	3d HoG ³	84%

¹Multi-Scale SSM, ²Space-Time Interest Points[8]
³all views used for training and testing

accuracy. In order to preserve generality, we evaluated our method on disjoint sets for training and testing views. Fig. 3 shows averaged results of experiments, where all 62 possible partitions of views for training and testing were used. As expected, the maximum performance was obtained when dividing the available views into equally-sized subsets. Confusions between semantically related classes only appeared occasionally. In general, we were even able to distinguish identical actions performed in different poses, which argues for the discriminativeness of our modeling scheme. For our experiments we used a standard desktop computer equipped with a Intel(R) Core(TM)2 Quad CPU 2.50 GHz and 8 GB of RAM. Some algorithms were parallelized, *e.g.* the Fourier Transform, SIFT extraction, or GMM modeling. While learning an action model for the whole dataset took about 3 hours, the SSM computation, feature extraction, and classification performed in real-time. Most of the approaches presented before concerning the recognition of actions in multi-view environments focus on cross-view setups, *i.e.* the system is trained on one single view and evaluated on another view. Hence, we adopted the evaluation method of Junejo *et al.*[7] in order to do a fair comparison. We did no further adaptations, especially we did not tune the process parameters to obtain optimal results for this scenario. Tab. 4 shows the resulting recognition rates compared to other not model-based approaches. While Junejo *et al.*[7] used a combination of HoF and HoG features, we can reach similar results using our proposed Fourier descriptors, which are assumed to be computed more efficiently. Furthermore, they enabled their approach to show time-scale invariance by extracting their SSM features on different scales, *i.e.* with distinct radii, while the SIFT features we used for representing SSMs are (time-) scale-invariant by design. By estimating 3d optical flow, Weinland *et al.*[18] obtained slightly higher recognition rates.

5 Summary and Outlook

We presented a framework for creating and evaluation temporal self-similarity maps to employ them for multi-view action recognition. It was pointed out, that

the invariance and stability properties of SSMs support our demands on a action recognition system.

We made three contributions: (i) we further extended the method originally presented in [7] by new low-level features and distance metrics, (ii) we applied a Gaussian Process (GP) classifier combined with histogram intersection kernel, which have been shown to be more suitable and efficient for comparing histograms[16,4], and (iii) we used a new extensive dataset for evaluating multi-view action recognition systems, which will be made publicly available.

It is straightforward to augment the *Bag of Self-Similarity Words* modeling scheme by histograms of co-occurrences of vocabulary words in order to improve the descriptive power of this representation. Another important aspect is the direct integration of calibration knowledge into our framework.

References

1. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. *ACM Comput. Surv.* 43(3), 16:1–16:43 (2011)
2. Cutler, R., Davis, L.S.: Robust real-time periodic motion detection, analysis, and applications. *TPAMI* 22(8), 781–796 (2000)
3. Farhadi, A., Tabrizi, M.K.: Learning to recognize activities from the wrong view point. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part I. LNCS*, vol. 5302, pp. 154–166. Springer, Heidelberg (2008)
4. Freytag, A., Rodner, E., Bodesheim, P., Denzler, J.: Rapid uncertainty computation with gaussian processes and histogram intersection kernels. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) *ACCV 2012, Part II. LNCS*, vol. 7725, pp. 511–524. Springer, Heidelberg (2013)
5. Holte, M.B., Chakraborty, B., Gonzalez, J., Moeslund, T.B.: A local 3-D motion descriptor for multi-view human action recognition from 4-D spatio-temporal interest points. *Selected Topics in Signal Processing* 6(5), 553–565 (2012)
6. Iwanski, J.S., Bradley, E.: Recurrence plots of experimental data: To embed or not to embed? *Chaos* 8(4), 861–871 (1998)
7. Junejo, I.N., Dexter, E., Laptev, I., Pérez, P.: View-independent action recognition from temporal self-similarities. *TPAMI* 33(1), 172–185 (2011)
8. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: *CVPR*, pp. 1–8 (2008)
9. Liu, J., Shah, M., Kuipers, B., Savarese, S.: Cross-view action recognition via view knowledge transfer. In: *CVPR*, pp. 3209–3216 (2011)
10. Maji, S., Berg, A.C., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: *CVPR*, pp. 1–8 (2008)
11. Marwan, N., Romano, M.C., Thiel, M., Kurths, J.: Recurrence plots for the analysis of complex systems. *Physics Reports* 438(5-6), 237–329 (2007)
12. McGuire, G., Azar, N.B., Shelhamer, M.: Recurrence matrices and the preservation of dynamical properties. *Physics Letters A* 237(1-2), 43–47 (1997)
13. Odone, F., Barla, A., Verri, A.: Building kernels from binary strings for image matching. *IP* 14(2), 169–180 (2005)
14. Poppe, R.: A survey on vision-based human action recognition. *IVC* 28(6), 976–990 (2010)
15. Rao, C., Yilmaz, A., Shah, M.: View-invariant representation and recognition of actions. *IJCV* 50(2), 203–226 (2002)

16. Rodner, E., Freytag, A., Bodesheim, P., Denzler, J.: Large-scale gaussian process classification with flexible adaptive histogram kernels. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 85–98. Springer, Heidelberg (2012)
17. Weinland, D., Boyer, E., Ronfard, R.: Action recognition from arbitrary views using 3D exemplars. In: ICCV, pp. 1–7 (2007)
18. Weinland, D., Özuysal, M., Fua, P.: Making action recognition robust to occlusions and viewpoint changes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 635–648. Springer, Heidelberg (2010)
19. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. CVIU 104(2), 249–257 (2006)