3D Shape Context and Distance Transform for Action Recognition

Matthias Grundmann, Franziska Meier and Irfan Essa
Georgia Institute of Technology
{*grundman, fmeier3, irfan*}*@cc.gatech.edu*

Abstract

We propose the use of 3D (2D+time) Shape Context to recognize the spatial and temporal details inherent in human actions. We represent an action in a video sequence by a 3D point cloud extracted by sampling 2D silhouettes over time. A non-uniform sampling method is introduced that gives preference to fast moving body parts using a Euclidean 3D Distance Transform. Actions are then classified by matching the extracted point clouds. Our proposed approach is based on a global matching and does not require specific training to learn the model. We test the approach thoroughly on two publicly available datasets and compare to several state-of-the-art methods. The achieved classification accuracy is on par with or superior to the best results reported to date.

1 Introduction

Recognition of actions in video sequences is a challenging research goal for computer vision. Efforts in action recognition building on the success in static object recognition have shown some success recently. However, action recognition requires analysis of the temporal order of the action. A direct application of object recognition over a set of consecutive video frames assumes significant local structure over a small temporal window. This results in a limited discriminative power for capturing the variations in human actions. We propose a novel method that uses the entire temporal information of action in a video sequence.

Our primary contributions are

(1) the use of the 3D Shape Context on point-sampled Space-Time Shapes to classify actions. We extend the 2D Shape Context proposed by Belongie et al. [1] to 3D by including the temporal dimension. The proposed extension also improves over [11] by using a novel discretization. And

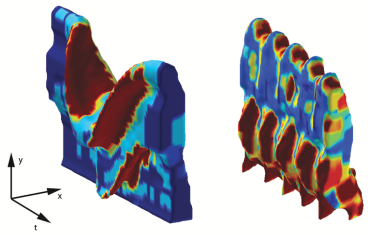

Figure 1: Response function R for the actions bend and skip from the Weizmann dataset. Red values indicate fast, blue values slow moving parts. (in color)

(2) we introduce a motion adaptive sampling technique by using the Euclidean 3D Distance Transform of Space-Time Shapes, that better models the time-variations.

In addition to the above technical contributions, we have attempted a detailed and thorough evaluation of our approach. It shows significant validity when tested on the Weizmann Dataset [2] and the KTH Dataset [18]. The obtained classification accuracy is on par with or superior to the best results reported to date. However, in contrast to other authors [4, 8, 15, 19] our proposed method does not employ any dataset specific learning stage.

Recent approaches for action recognition can be separated by their use of temporal information. *Limited temporal window approaches* rely on local features to capture local temporal information. Laptev and Lindeberg's [12] proposed scale-adapted space-time interest points, which are scale-adapted Harris features [13] extended along the temporal dimension. Using Support Vector Machines for classification, Schuldt *et al.* [18] utilized these features for action recognition. Dollar *et al.* [4] proposed separable linear filters in order to extend the feature response to constant and fluent motions. Niebles *et al.* [15] classified actions by applying these features to unsupervised learning in form of pLSA. Niebles and Fei-Fei [14] build a hierarchical

model for action recognition using Dollar *et al.*'s separable linear filters [4] as well as static edge features represented by the 2D Shape Context [1]. Most recently, Jhuang *et al.* [8] have presented a biologically inspired approach for action recognition. Among others, local temporal information is captured by space-time gradients, optical flow or complex features collected over 7 consecutive video frames. Their approach uses a learning stage to extract so called prototype features.

Global temporal information approaches rely on global features or mapping of frames that capture the whole time span of the action. Bobick and Davis [3] have represented an action by a motion energy image (MEI) and a motion history image (MHI) to capture *where* and *how* the motion happened. Some years before, Polana and Nelson [17] have described an action as one feature vector of summed optical flow magnitudes in a spatiotemporal grid. Efros *et al.* [5] have introduced the spatio-temporal motion descriptor consisting of 4 motion channels (optical flow separated by dimension and sign). The temporal order of the motion is established by taking the maximum frame by frame correlation accumulated over a temporal window. Recently, Wang *et al.* [19] have used a bag-of-words approach to represent an action sequence as a sequence of prototype frames and classify it by the use of a Semi-Latent Dirichlet Allocation (SLDA) model. The approach closest to ours was proposed by Blank *et al.* [2]. They obtain global features by integrating local measures (stickness and plateness) of Space-Time Shapes that are generated from silhouetted video data.

2 Representing Actions

We base our approach on the observation that the silhouette contains sufficient information to accurately identify the human action. We justify this assumption by noting that the object's pose in an image can be reliably determined by focusing on its edge representation [16] once the object's identity is known. Furthermore, silhouettes have advantages over local features, which become unstable in case of low-resolution or noisy, blurred video.

We extract the human's silhouette of each action sequence by applying Kim *et al.*'s [10] real-time background substraction method. We require a known background, however it is not limited to static cameras. In order to accommodate for the different speeds with which the actions are carried out, we discard frames which temporal variation (computed by a XOR mask) is less than the median variation of the sequence. We found that this technique performs as good as the usual sliding window approach.

Each action sequence is represented by a 3D point cloud obtained by uniformly sampling the silhouettes. To increase robustness w.r.t. segmentation errors we prior smooth the silhouettes along their spatial and temporal dimension.

2.1 3D-Shape Context

Our 3D-Shape Context is an extension of the 2D Shape Context proposed by Belongie *et al.* [1] by including the temporal dimension. It differs from Kortgen *et al.*'s [11] 3D Shape Context, which does not extend an important property of the 2D shape context to 3D: Bins of equal distance from the origin should have the same size. Otherwise implicitly more weight is given to the smaller bins. Although Huang and Trivedi [7] propose a 3D Shape Context for Gesture Analysis, their Shape Context is a global descriptor capturing the whole shape and differs significantly in its idea from our locally evaluated 3D shape context.

To obtain a correspondence between the 3D point clouds of two action sequences we associate a local feature descriptor with each point. Lets assume action sequence P is represented by N points p_i. For each p_i we compute the $N-1$ difference vectors $d_{i,j}$ to its neighbors p_j. This rich representation is converted to a histogram h_i collected over the discretization of difference vectors w.r.t. angles and magnitude.

Similar to [1] the magnitude R is logarithmical discretized to make the descriptor more sensitive to nearby points. A uniform discretization for the longitude angle $\phi \in [0, 2\pi]$ is used. However the transformation from 2D to 3D demands a non-uniform discretization for the latitude angle $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ so that each bin covers the same surface area. That means we want to determine $\theta_i,\ i = 1...N$ so that

$$\int_{\phi=0}^{2\pi} \int_{\theta=0}^{\theta_i} r^2 \cos\theta \, d\theta d\phi = \frac{i}{N} 2\pi r, \qquad (1)$$

with the right hand side being the i^{th} fraction of the surface area of the upper hemisphere. It follows that $\theta_i = \arcsin(\frac{i}{N})$.

Since the difference vectors $d_{i,j}$ are invariant under translation the obtained 3D Shape Context has the same property. Furthermore the histograms are very sparse, in average only 9.6% of a histogram's bins are non-zero, leading to a high discrimination.

As the histogram bins are independent normally distributed random variables it is appropriate to define the matching costs $c_{i,j}$ of two histograms h_i and h_j as the χ^2 test statistic, similar to [1]. The optimal correspondence between each point p_i of action sequence P

and each point q_j of action sequence Q is a bijective mapping π that minimizes the total matching costs of the corresponding histograms $\pi = \text{argmin}_\sigma \sum_i c_{i,\sigma(i)}$. This problem poses a linear assignment problem that can be solved efficiently by the method described in [9].

2.2 3D-Distance Transform

We propose a new motion-adaptive sampling method of the silhouettes that favors fast moving body parts to better discriminate actions that only differ by small dynamic parts, like running and jumping (see Fig. 3). We introduce the 3D Distance Transform of Space Time Shapes to compute these fast moving parts. For each voxel in the Space-Time volume we compute the closest distance to the boundary. We adopt the 2D Distance Transform described in [6] and apply it to 3D.

Let B be the set of all boundary points and G the discrete grid of the Space-Time volume. For each $\mathbf{g} \in G$ the squared distance to the boundary B is given by the Euclidean Distance Transform

$$D_B(\mathbf{g}) = \min_{\mathbf{b} \in B} \parallel \mathbf{b} - \mathbf{g} \parallel^2 . \tag{2}$$

As shown in [6] the above Eq. 2 can be transformed so that the minimum is taken over the whole grid G by introducing the indicator function

$$\mathcal{I}(\mathbf{g}) = \begin{cases} 0 & \text{if } \mathbf{g} \in B \\ \infty & \text{else} \end{cases} . \tag{3}$$

In our case G has three dimensions and the Distance Transform in Eq. 2 can then be expressed as follows:

$$\begin{aligned} & D_B[\mathcal{I}](\mathbf{g}) \\ &= \min_{\mathbf{h}_{x,y} \in G} (\parallel (\mathbf{g} - \mathbf{h})_{x,y} \parallel^2 + \min_{h_t \in G} (g_t - h_t)^2 + \mathcal{I}(\mathbf{h})) \\ &= \min_{\mathbf{h}_{x,y} \in G} (\parallel (\mathbf{g} - \mathbf{h})_{x,y} \parallel^2 + D_B[\mathcal{I}](\mathbf{g})|_{x,y}) \\ &= \min_{h_x \in G} ((g_x - h_x)^2 + \min_{h_y \in G} (g_y - h_y)^2 + D_t) \\ &\quad \text{with } D_t := D_B[\mathcal{I}](\mathbf{g})|_{x,y} \\ &= \min_{h_x \in G} ((g_x - h_x)^2 + D_B[D_t](\mathbf{g})|_x) \\ &= D_B[D_y](\mathbf{g}) \text{ with } D_y := D_B[D_t](\mathbf{g})|_x . \end{aligned} \tag{4}$$

The above Eq. 4 states, that the 3D Distance Transform can be computed by applying the one-dimensional version three successive times. The overall complexity is limited by $O(|G|)$.

Fast moving parts exhibit the property that their distance to the boundary changes rapidly over time. Therefore we use the squared derivative of the 3D Distance Transform w.r.t. time as a measure to identify these parts. We constrain the response function R to lie in the interval [0,1]:

$$R(\mathbf{g}) = \frac{\log(1 + \partial_t D_B[\mathcal{I}](\mathbf{g}))}{\max_{\bar{\mathbf{g}} \in G} \log(1 + \partial_t D_B[\mathcal{I}](\bar{\mathbf{g}}))} . \tag{5}$$

To propagate high response values to the boundary, R is maximum-filtered with a spatial radius of 3 and a temporal radius of 1. Fig. 1 shows an example of the response function R at the boundary for 2 actions from the Weizmann dataset.

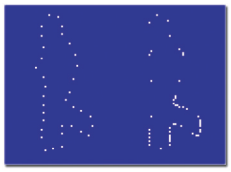

Figure 2: Uniform and adaptive sampling for action running.

To prefer fast moving body parts at the sampling stage, a smaller sample window is used for points with a higher response value. Fig. 2 shows the difference between both sampling techniques for a single frame.

3 Experimental Setup and Results

We classify actions by a leave-one-out method. The test sequence and all sequences showing the same action performed by the same actor are removed from the dataset. The test sequence is compared to the remaining sequences in the dataset based on the total matching costs.

Datasets: We tested our approach on two publicly available datasets:
Weizmann Dataset [2]: The dataset consists of 81 low resolution videos (180x144, 25 fps) showing 9 persons performing 9 different actions. There is another action "skip" that is usually not included in the dataset and makes classification more difficult. However we evaluate our approach also on this "extended" dataset.
KTH Dataset [18]: The dataset consists of 2391 low resolution videos (160x120, 25fps) showing six types of human actions each performed 4 times by 25 persons.

Results: In order to obtain a baseline we compare to Bobick and Davis' [3] Temporal Templates. As feature descriptor for each action sequence we compute the seven Hu moments (see appendix of [3]) for both the MHI and the MEI. We apply several machine learning algorithms to these descriptors: Bobick and Davis' Nearest Neighbor approach, Neural Network with prior PCA and Support Vector Machines(SVM) with RBF

Author	W. w/o skip	W. w/ skip	KTH
Our method	**96.39**	**94.6**	**93.52**
Blank *et al*. [2]	98.77	*	*
Jhuang *et al*. [8]	97.0	*	96.0
Wang *et al*. [19]	*	*	92.43
Dollar *et al*. [4]	*	*	88.2
Niebles *et al*. [15]	*	*	81.5
Niebles *et al*. [14]	72.8	*	*
SVM	90.36	84.9	79
Neural Net.	83.13	72.04	79.67
Nearest N.	74.7	72.04	74

Table 1: Left column shows normal, middle extended Weizmann dataset. Right column shows results for S1 subset. [18] tests on all sets while [19, 15] might test on all four subsets although example figures only show the S1 subset. * denotes that no results are available.

kernels. In case of the KTH dataset, similar to [8, 18] we perform 5 random splits of the data set into 9 test and 16 database persons. This is only done for comparison; the 16 other person are not used for a learning stage.

Table 1 shows the results for both datasets. Our achieved classification accuracy is on par with other efforts. Our proposed method scales well to the extended Weizmann dataset compared to the baseline.

We explored the advantage of the motion adaptive sampling on the Weizmann dataset. Classification accuracy increased from 92.77% to 96.39%. On the extended dataset, the impact is more significant. Accuracy increased from 86.02% to 94.6%. The right confusion matrix in Fig. 3 shows confusion clusters. Only actions that exhibit spatial similarity but differ along the temporal dimension are confused *e.g*. walk and side or jump, run and skip. By using motion adaptive sampling these actions can be very well discriminated as the left confusion matrix in Fig. 3 shows.

Running time of our system: Preprocessing consists of computing the local 3D Shape Contexts (4.4s) and the 3D Distance Transform (30 ms) for each action sequence. Classification on the Weizmann dataset takes 28.4s in average.

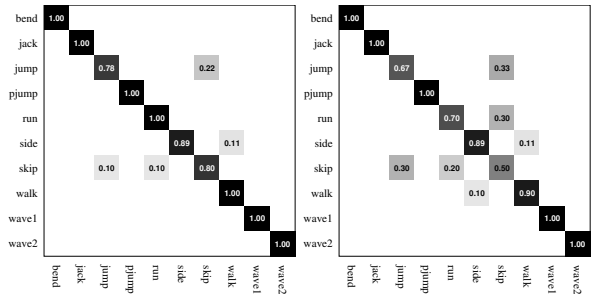

Figure 3: Confusion matrices for extended Weizmann Dataset with (left) and without (right) motion adaptive sampling

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE PAMI*, 24(4):509–522, 2002.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, pages 1395–1402, 2005.

[3] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE PAMI*, 23(3):257–267, 2001.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

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

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. Technical report, Cornell Computing and Information Science, 2004.

[7] K. S. Huang and M. M. Trivedi. 3d shape context based gesture analysis integrated with tracking using omni video array. In *CVPR*, page 80, 2005.

[8] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, 2007.

[9] R. Jonker and A. Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987.

[10] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. S. Davis. Real-time foreground-background segmentation using codebook model. *Real-Time Imaging*, 11(3), 2005.

[11] M. Koertgen, G. Parl, M. Novotni, and R. Klein. 3d shape matching with 3d shape contexts. *Seventh Central European Seminar on Computer Graphics*, 2003.

[12] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, page 432, 2003.

[13] T. Lindeberg. Feature detection with automatic scale selection. *Int. J. Comput. Vision*, 30(2):79–116, 1998.

[14] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *CVPR*, 2007.

[15] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

[16] C. F. Olson and D. P. Huttenlocher. Automatic target recognition by matching oriented edge pixels. *IEEE Transactions on Image Processing*, 6(1), 1997.

[17] R. Polana and R. Nelson. Low level recognition of human motion. *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 77–82, 1994.

[18] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004.

[19] Y. Wang, P. Sabzmeydani, and G. Mori. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *ICCV*, 2007.