Instant Action Recognition

Thomas Mauthner[1], Peter M. Roth[1], Horst Bischof[1]

Institute for Computer Graphics and Vision
Graz University of Technology
Inffeldgasse 16/II, 8010 Graz, Austria
{mauthner,pmroth,bischof}@icg.tugraz.at

Abstract. In this paper, we present an efficient system for action recognition from very short sequences. For action recognition typically appearance and/or motion information of an action is analyzed using a large number of frames. This is a limitation if very fast actions (e.g., in sport analysis) have to be analyzed. To overcome this limitation, we propose a method that uses a single-frame representation for actions based on appearance and motion information. In particular, we estimate Histograms of Oriented Gradients (HOGs) for the current frame as well as for the corresponding dense flow field. The thus obtained descriptors are efficiently represented by the coefficients of a Non-negative Matrix Factorization (NMF). Actions are classified using an one-vs-all Support Vector Machine. Since the flow can be estimated from two frames, in the evaluation stage only two consecutive frames are required for the action analysis. Both, the optical flow as well as the HOGs, can be computed very efficiently. In the experiments, we compare the proposed approach to state-of-the-art methods and show that it yields competitive results. In addition, we demonstrate action recognition for real-world beach-volleyball sequences.

1 Introduction

Recently, human action recognition has shown to be beneficial for a wide range of applications including scene understanding, visual surveillance, human computer interaction, video retrieval or sports analysis. Hence, there has been a growing interest in developing and improving methods for this rather hard task (see Section 2). In fact, a huge variety of actions at different time scales have to be handled – starting from waving with one hand for a few seconds to complex processes like unloading a lorry. Thus, the definition of an action is highly task dependent and for different actions different methods might be useful.

The objective of this work is to support the analysis of sports videos. Therefore, principle actions represent short time player activities such as running, kicking, jumping, playing, or receiving a ball. Due to the high dynamics in sport actions, we are looking for an action recognition method that can be applied to a minimal number of frames. Optimally, the recognition should be possible using only two frames. Thus, to incorporate the maximum information available per frame we want to use appearance and motion information. The benefit of this represenentation is motivated and illustrated in Figure 1. In particular, we apply Histograms of Oriented Gradients (HOG)

[1] to describe the appearance of a single-frame action. But as can be seen from Figure 1(a) different actions that share one specific mode can not be distinguished if only appearance-based information is available. In contrast, as shown in Figure 1(b), even if the appearance is very similar, additionally analyzing the corresponding motion information can help to discriminate between two actions; and vice versa. In particular, for that purpose we compute a dense optical-flow field, such that for frame t the appearance and the flow information is computed from frame $t-1$ and frame t only. Then the optical flow is represented similarly to the appearance features by (signed) orientation histograms.

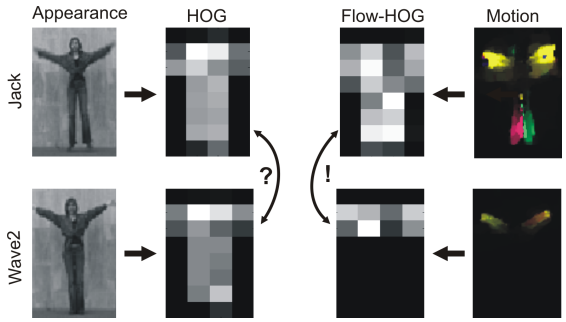


Fig. 1. Overview of the proposed ideas for single frame classification: By using only appearance-based information ambiguities complicate human action recognition (left). By including motion information (optical flow), additional crucial information can be acquired to avoid these confusions (right). Here, the optical flow is visualized using hue to indicate the direction and intensity for the magnitude; the HOG cells are visualized by their accumulated magnitudes.

Since the thus obtained HOG descriptors for both, appearance and motion, can be described by a small number of additive modes, similar to [2, 3], we apply Non-negative Matrix Factorization (NMF) [4] to estimate a robust and compact representation. Finally, the motion and the appearance features (i.e., their NMF coefficients) are concatenated to one vector and linear one-vs-all SVMs are applied to learn a discriminative model. To compare our method with state-of-the-art approaches we evaluated it on a standard action recognition database. In addition, we show results on beach-volleyball videos, where we use very different data for training and testing to emphasize the applicability of our method.

The remainder of this paper is organized as follows. Section 2 gives an overview of related work and explains the differences to the proposed approach. In Section 3 our new action recognition system is introduced in detail. Experimental results for a typical benchmark dataset and a challenging real-world task are shown in Section 4. Finally, conclusion and outlook are given in Section 5.

2 Related work

In the past, many researchers have tackled the problem of human action recognition. Especially for recognizing actions performed by a single person various methods exist that yield very good classification results. Many classification methods are based on the analysis of a temporal window around a specific frame. Bobick and Davis [5] used motion history images to describe an action by accumulating human silhouettes over time. Blank et al. [6] created 3-dimensional space-time shapes to describe actions. Weinland and Boyer [7] used a set of discriminative static key-pose exemplars without any spatial order. Thurau and Hlaváč [2] used pose-primitives based on HOGs and represented actions as histograms of such pose-primitives. Even though these approaches show that shape or silhouettes over time are well discriminating features for action recognition, the use of temporal windows or even of a whole sequence implies that actions are recognized with a specific delay.

Having the spatio-temporal information, the use of optical flow is an obvious extension. Efros et al. [8] introduced a motion descriptor based on spatio-temporal optical flow measurements. An interest point detector in spatio-temporal domain based on the idea of Harris point detector was proposed by Laptev and Lindeberg [9]. They described the detected volumes with several methods such as histograms of gradients or optical flow as well as PCA projections. Dollár et al. [10] proposed an interest point detector searching in space-time volumes for regions with sudden or periodic changes. In addition, optical flow was used as a descriptor for the 3D region of interest. Niebles et al. [11] used a constellation model of bag-of-features containing spatial and spatio-temporal [10] interest points. Moreover, single-frame classification methods were proposed. For instance, Mikolajczyk and Uemura [12] trained a vocabulary forest on feature points and their associated motion vectors.

Recent results in the cognitive sciences have led to biologically inspired vision systems for action recognition. Jhuang et al. [13] proposed an approach using a hierarchy of spatio-temporal features with increasing complexity. Input data is processed by units sensitive to motion-directions and the responses are pooled locally and fed into a higher level. But only recognition results for whole sequences have been reported, where the required computational effort is approximatly 2 minutes for a sequence consisting of 50 frames. Inspired by [13] a more sophisticated (and thus more efficient approach) was proposed by Schindler and van Gool [14]. They additionally use appearance information, but both, appearance and motion, are processed in similar pipelines using scale and orientation filters. In both pipelines the filter responses are max-pooled and compared to templates. The final action classification is done by using multiple one-vs-all SVMs.

The approaches most similar to our work are [2] and [14]. Similar to [2] we use HOG descriptors and NMF to represent the appearance. But in contrast to [2], we do not not need to model the background, which makes our approach more general. Instead, similar to [14], we incorporate motion information to increase the robustness and apply one-vs-all SVMs for classification. But in contrast to [14], in our approach the computation of feature vectors is less complex and thus more efficient. Due to a GPU-based flow estimation and an efficient data structure for HOGs our system computation time is optimized, and therefore can be performed in real time. Moreover, since we can

estimate the motion information using a pair of subsequent frames we require only two frames to analyze an action.

3 Instant Action Recognition System

In this section, we introduce our action recognition system, which is schematically illustrated in Figure 2. In particular, we combine appearance and motion information to enable a frame-wise action analysis. To represent the appearance, we use histograms of oriented gradients (HOGs) [1]. HOG descriptors are locally normalized gradient histograms, which have shown their capability for human detection and can also be estimated efficiently by using integral histograms [15]. To estimate the motion information, a dense optical flow field is computed between consecutive frames using an efficient GPU-based implementation [16]. The optical flow information can also be described using orientation histograms without dismissing the information about the gradient direction. Following the ideas presented in [2] and [17], we reduce the dimensionality of the extracted histograms by applying sub-space methods. As stated in [3, 2] articulated poses, as they appear during human actions, can be well described using NMF basis vectors. We extend this ideas by building a set of NMF basis vectors for appearance and the optical flow in parallel. Hence the human action is described in every frame by NMF coefficient vectors for appearance and flow, respectively. The final classification on per-frame basis is realized by using multiple SVMs trained on the concatenations of the appearance and flow coefficient vectors of the training samples.

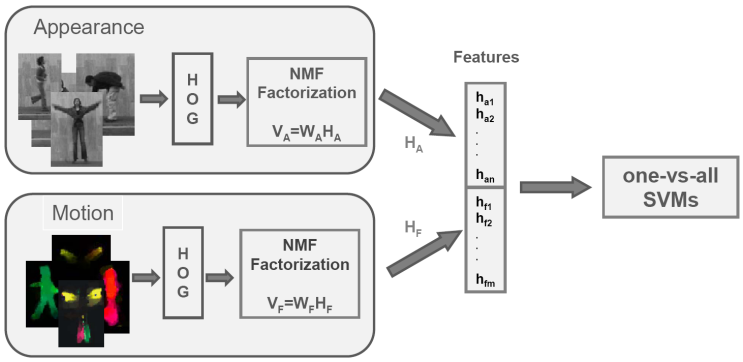


Fig. 2. Overview of the proposed approach: Two representations for appearance and flow are estimated in parallel. Both are described by HOGs and represented by NMF coefficients, which are concatenated to a single feature vector. These vectors are then learned using one-vs-all SVMs.

3.1 Appearance Features

Given an image $\mathbf{I_t} \in \mathbb{R}^{\mathbf{m}\times\mathbf{n}}$ at time step t. To compute the gradient components $g_x(x, y)$ and $g_y(x, y)$ for every position (x, y), the image is filtered by 1-dimensional

masks $[-1, 0, 1]$ in x and y direction [1]. The magnitude $m(x, y)$ and the signed orientation $\Theta_S(x, y)$ are computed by

$$m(x, y) = \sqrt{g_x(x, y)^2 + g_y(x, y)^2} \tag{1}$$

$$\Theta_S(x, y) = \tan^{-1}\left(g_y(x, y)/g_x(x, y)\right) \ . \tag{2}$$

To make the orientation insensitive to the order of intensity changes, only unsigned orientations Θ_U are used for appearance:

$$\Theta_U(x, y) = \begin{cases} \Theta_S(x, y) + \pi & \theta_S(x, y) < 0 \\ \Theta_S(x, y) & \text{otherwise} \end{cases} . \tag{3}$$

To create the HOG descriptor, the patch is divided into non-overlapping 10×10 cells. For each cell, the orientations are quantized into 9 bins and weighted by their magnitude. Groups of 2×2 cells are combined in so called overlapping blocks and the histogram of each cell is normalized using the L2-norm of the block. The final descriptor is built by concatenation of all normalized blocks. The parameters for cell-size, block-size, and the number of bins may be different in literature.

3.2 Motion Features

In addition to appearance we use optical flow. Thus, for frame t the appearance features are computed from frame t, and the flow features are extracted from frames t and $t-1$. In particular, to estimate the dense optical flow field, we apply the method proposed in [16], which is publicly available: *OFLib*[1]. In fact, the GPU-based implementation allows a real-time computation of motion features.

Given $\mathbf{I_t}, \mathbf{I_{t-1}} \in \mathbb{R}^{\mathbf{m}\times\mathbf{n}}$, the optical flow describes the shift from frame $t-1$ to t with the disparity $\mathbf{D_t} \in \mathbb{R}^{\mathbf{m}\times\mathbf{n}}$, where $d_x(x, y)$ and $d_y(x, y)$ denote the disparity components in x and y direction at location (x, y). Similar to the appearance features, orientation and magnitude are computed and represented with HOG descriptors. In contrast to appearance, we use signed orientation Θ_S to capture different motion directions for same poses. The orientation is quantized into 8 bins only, while we keep the same cell/block combination as described above.

3.3 NMF

If the underlying data can be described by distinctive local information (such as the HOGs of appearance and flow) the representation is typically very sparse. In contrast to other sub-space methods, Non-negative Matrix Factorization (NMF) [4] does not allow negative entries, neither in the basis nor in the encoding. Formally, NMF can be described as follows. Given a non-negative matrix (i.e., a matrix containing vectorized images) $\mathbf{V} \in \mathbb{R}^{m\times n}$, the goal of NMF is to find non-negative factors $\mathbf{W} \in \mathbb{R}^{n\times r}$ and $\mathbf{H} \in \mathbb{R}^{r\times m}$ that approximate the original data:

[1] http://gpu4vision.icg.tugraz.at/

$$\mathbf{V} \approx \mathbf{WH} \; . \tag{4}$$

Since there is no closed-form solution, both matrices, $\mathbf{W}$ and $\mathbf{H}$, have to be estimated in an iterative way. Therefore, we consider the optimization problem

$$\begin{aligned} &\min ||\mathbf{V} - \mathbf{WH}||^2 \\ &\text{s.t. } \mathbf{W}, \mathbf{H} > 0 \; , \end{aligned} \tag{5}$$

where $||.||^2$ denotes the squared Euclidean Distance. The optimization problem (5) can be iteratively solved by the following update rules:

$$\mathbf{H}_{a,j} \leftarrow \mathbf{H}_{a,j} \frac{\left[\mathbf{W}^T\mathbf{V}\right]_{a,j}}{\left[\mathbf{W}^T\mathbf{WH}\right]_{a,j}} \quad and \quad \mathbf{W}_{i,a} \leftarrow \mathbf{W}_{i,a} \frac{\left[\mathbf{VH}^T\right]_{i,a}}{\left[\mathbf{WHH}^T\right]_{i,a}} \; , \tag{6}$$

where $[.]$ denote that the multiplications and divisions are performed element by element.

3.4 Classification via SVM

For the final classification the NMF-coefficients obtained for appearance and motion are concatenated to a final feature vector. As we will show in Section 4, less than 100 basis vectors are sufficient for our tasks. Therefore, compared to [14] the dimension of the feature vector is rather small, which drastically reduces the computational costs. Finally, a linear one-vs-all SVM is trained for each action class using *LIBSVM*[2]. In particular, no weighting of appearance or motion cue was performed. Thus, the only tuning parameter is the number of basis vectors for each cue.

4 Experimental Results

To show the benefits of the proposed approach, we split the experiments into two main parts. First, we evaluated our approach on a publicly available benchmark dataset (i.e., *Weizmann Human Action Dataset* [6]). Second, we demonstrate the method for a real-world application (i.e., action recognition for beach-volleyball).

4.1 Weizmann Human Action Dataset

The *Weizmann Human Action Dataset* [6] is a publicly available[3] dataset, that contains 90 low resolution videos (180×144) of nine subjects performing ten different actions: running, jumping in place, jumping forward, bending, waving with one hand, jumping jack, jumping sideways, jumping on one leg, walking, and waving with two hands. Illustrative examples for each of these actions are shown in Figure 3. Similar to, e.g., [2, 14] all experiments on this dataset were carried out using a leave-one-out strategy (i.e., we used 8 individuals for training and evaluated the learned model for the missing one.

[2] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

[3] http://www.wisdom.weizmann.ac.il/ vision/SpaceTimeActions.html

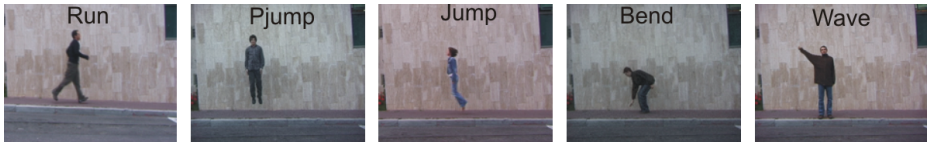
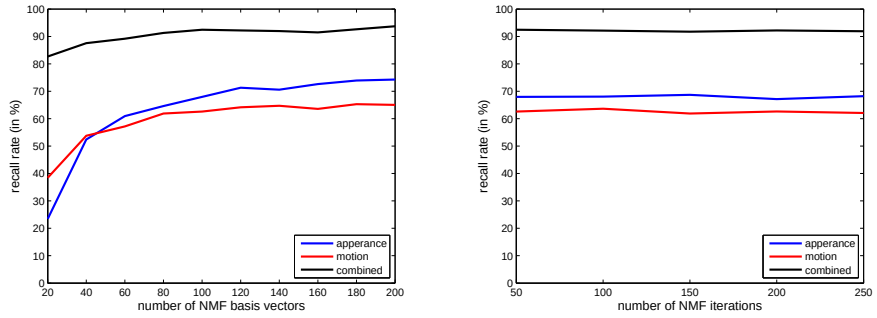


Fig. 3. Examples from the Weizmann human action dataset.

In general, Figure 4 shows the benefits of the proposed approach. It can be seen that neither the appearance-based nor the motion-based representation solve the task satisfactorily. But if both representations are combined, we get a significant improvement of the recognition performance! To analyzed the importance of the NMF parameters used for estimating the feature vectors that are learned by SVMs, we ran the leave-one-out experiments varying the NMF parameters, i.e., the number of basis vectors and the number of iterations. The number of basis vectors was varied in the range from 20 to 200 and the number of iterations from 50 to 250. The other parameter was kept fixed, respectively. It can be seen from Figure 4(a) that increasing the number of basis vectors to a level of 80-100 steadily increases the recognition performance, but that further increasing this parameter has no significant effect. Thus using 80-100 basis vectors is sufficient for our task. In contrast, it can be seen from Figure 4(b) that the number of iterations has no big influence on the performance. In fact, a representation that was estimated using 50 iterations yields the same results as one that was estimated using 250 iterations!



(a) (b)

Fig. 4. Importance of NMF parameters for action recognition performance: recognition rate depending (a) on the number of basis vectors using 100 iterations and (b) on the number of NMF iterations for 200 basis vectors.

In the following, we present the results for the leave-one-out experiment for each action in Table 1. Due to the results discussed above, we show the results obtained by using 80 NMF coefficients obtained by 50 iterations. It can be seen that with exception of “run” and “skip”, which are very similar in both appearance and motion on a short

frame basis, the recognition rate is always near 90% or higher, see confusion matrix in Table 3.

action	bend	run	side	wave2	wave1	skip	walk	pjump	jump	jack
rec.-rate	95.79	78.03	99.73	96.74	95.67	75.56	94.20	95.48	88.50	93.10

Table 1. Recognition rate for the leave-one-out experiment for the different actions.

Estimating the overall recognition rate we get a correct classification rate of 91.28%. In fact, this average is highly influenced by the results on the "run" and "skip" dataset. Without these classes, the overall performance would be significantly higher than 90%. By averaging the recognition results in a temporal window (i.e., we used a window size of 6 frames) we can boost the recognition results to 94.25%. This improvement is mainly reached by incorporating more temporal information. Further extenting the temporal window size has not shown additional significant improvements. In the following, we compare this result with state-of-the-art methods considering the reported recognition rate and the number of frames that were used to calculate the response. The results are summarized in Table 2.

method	rec.-rate	# frames
proposed	91.28%	2
	94.25%	6
Thurau & Hlaváč [2]	70.4%	1
	94.40%	all
Niebles et al. [11]	55.0%	1
	72.8%	all
Schindler & v. Gool [14]	93.5%	2
	96.6%	3
	99.6%	10
Blank et al. [6]	99.6%	all
Jhuang et al. [13]	98.9%	all
Ali et al. [18]	89.7	all

Table 2. Recognition rates and number of required frames for different approaches.

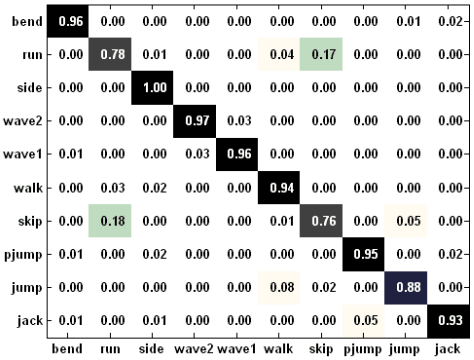


Table 3. Confusion matrix for 80 basis vectors and 50 iterations.

It can be seen that most of the reported approaches that use longer sequences to analyze the actions clearly outperform the proposed approach. But among those methods using only two frames our results are competitive.

4.2 Beach-Volleyball

In this experiment we show that the proposed approach can be applied in practice to analyze events in beach-volleyball. For that purpose, we generated indoor training sequences showing different actions including *digging*, *running*, *overhead passing*, and *running sideways*. Illustrative frames used for training are shown in Figure 5. From these sequences we learned the different actions as described in Section 3.

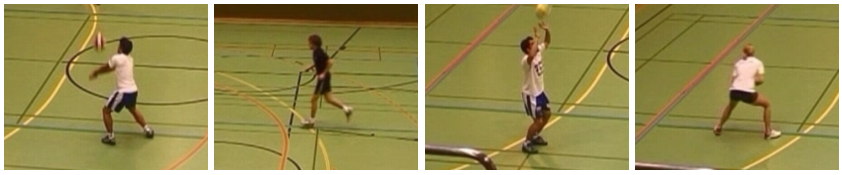


Fig. 5. Volleyball – training set: (a) *digging*, (b) *run*, (c) *overhead passing*, and (d) *run sideway*.

The thus obtained models are then applied for action analysis in outdoor beach-volleyball sequences. Please note the considerable difference between the training and the testing scenes. From the analyzed patch the required features (appearance NMF-HOGs and flow NMF-HOGs) are extracted and tested if they are consistent with one of the previously learned SVM models. Illustrative examples are depicted in Figure 6, where both tested actions, *digging* (yellow bounding box in (a)) and *overhead passing* (red bounding box in (b)) are detected correctly in the shown sequences!

Fig. 6. Volleyball – test set: (left) *action digging* (yellow bounding box) and (right) action *overhead passing* (red bounding box) are detected correctly.

5 Conclusion

We presented an efficient action recognition system based on a single-frame representation combining appearance-based and motion-based (optical flow) description of the data. Since in the evaluation stage only two consecutive frames are required (for estimating the flow), the methods can also be applied for very short sequences. In particular, we propose to use HOG descriptors for both, appearance and motion. The thus obtained feature vectors are represented by NMF coefficients and are concatenated to learn action models using SVMs. Since we apply a GPU-based implementation for optical flow and an efficient estimation of the HOGs, the method is highly applicable for tasks where quick and short actions (e.g., in sports analysis) have to be analyzed. The experiments showed that even using this short-time analysis competitive results can be obtained on a standard benchmark dataset. In addition, we demonstrated that the proposed method can be applied for a real-world task such as action detection in volleyball. Future work will mainly concern the training stage by considering a more sophisticated learning method (e.g., an weighted SVM) and improving the NMF implementation. In fact extensions such as sparsity constraints or convex formulation (e.g.,[19, 20]) have shown to be beneficial in practice.

Acknowledgment

This work was supported be the Austrian Science Fund (FWF P18600), by the FFG project AUTOVISTA (813395) under the FIT-IT programme, and by the Austrian Joint Research Project Cognitive Vision under projects S9103-N04 and S9104-N04.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2005)
2. Thurau, C., Hlaváč, V.: Pose primitive based human action recognition in videos or still images. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2008)
3. Agarwal, A., Triggs, B.: A local basis representation for estimating human pose from cluttered images. In: Proc. Asian Conf. on Computer Vision. (2006)
4. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401** (1999) 788–791
5. Bobick, A.F., Davis, J.W.: The representation and recognition of action using temporal templates. IEEE Trans. on Pattern Analysis and Machine Intelligence **23**(3) (2001) 257–267
6. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Proc. IEEE Intern. Conf. on Computer Vision. (2005) 1395–1402
7. Weinland, D., Boyer, E.: Action recognition using exemplar-based embedding. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2008)
8. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: Proc. European Conf. on Computer Vision. (2003)
9. Laptev, I., Lindeberg, T.: Local descriptors for spatio-temporal recognition. In: Proc. IEEE Intern. Conf. on Computer Vision. (2003)
10. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Proc. IEEE Workshop on PETS. (2005) 65–72
11. Niebles, J.C., Fei-Fei, L.: A hierarchical model of shape and appearance for human action classification. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2007)
12. Mikolajczyk, K., Uemura, H.: Action recognition with motion-appearance vocabulary forest. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2008)
13. Jhuang, H., Serre, T., L.Wolf, Poggio, T.: A biologically inspired system for action recognition. In: Proc. IEEE Intern. Conf. on Computer Vision. (2007)
14. Schindler, K., van Gool, L.: Action snippets: How many frames does human action recognition require? In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2008)
15. Porikli, F.: Integral histogram: A fast way to extract histograms in cartesian spaces. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Volume 1. (2005) 829–836
16. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l1 optical flow. In: Proc. DAGM Symposium. (2007)
17. Lu, W.L., Little, J.J.: Tracking and recognizing actions at a distance. In: CVBASE, Workshop at ECCV. (2006)
18. Ali, S., Basharat, A., Shah, M.: Chaotic invariants for human action recognition. In: Proc. IEEE Intern. Conf. on Computer Vision. (2007)
19. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research **5** (2004) 1457–1469
20. Heiler, M., Schnörr, C.: Learning non-negative sparse image codes by convex programming. In: Proc. IEEE Intern. Conf. on Computer Vision. Volume II. (2005) 1667–1674