Thurau C., Hlavac V.: Pose primitive based human action recognition in videos or still images, Proceedings of the Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA, June 22-24, 2008. IEEE Computer Society Press.

Pose primitive based human action recognition in videos or still images

Christian Thurau
Technical University Dortmund
Department of Computer Science
christian.thurau@udo.edu

Václav Hlaváč
Czech Technical University
Center for Machine Perception
hlavac@cmp.felk.cvut.cz

Abstract

This paper presents a method for recognizing human actions based on pose primitives. In learning mode, the parameters representing poses and activities are estimated from videos. In run mode, the method can be used both for videos or still images. For recognizing pose primitives, we extend a Histogram of Oriented Gradient (HOG) based descriptor to better cope with articulated poses and cluttered background. Action classes are represented by histograms of poses primitives. For sequences, we incorporate the local temporal context by means of n-gram expressions. Action recognition is based on a simple histogram comparison. Unlike the mainstream video surveillance approaches, the proposed method does not rely on background subtraction or dynamic features and thus allows for action recognition in still images.

1. Introduction

Human action recognition aims at automatically telling the activity of a person, i.e. to identify if someone is walking, dancing, or performing other types of activities. It is a crucial prerequisite for a number of applications, including surveillance, content-based image retrieval, or human-robot interaction. The task is challenging due to changes in the appearance of persons, articulation in poses, changing backgrounds, and camera movements.

In this work, we concentrate on pose based activity recognition. We infer action classes based on a single recognized *pose primitive*, or based on a sequence of recognized poses. The action classes considered are often referred to as *primitive actions*, whereas more complex activities can be understood as a sequencing of these primitive actions. Ideally, we would be able to assign a suitable action class label to arbitrary long or short sequences. In contrast to other contributions, we do not use dynamic features. While ignoring dynamic features makes the task of behavior recognition more demanding, it allows for action recognition in still images which we find too important to be left out.

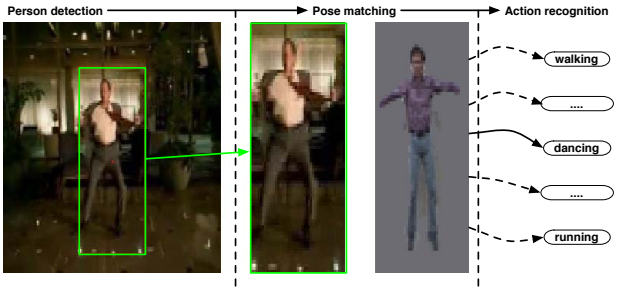

Figure 1. Given an image or an image sequence (left picture), we detect a person, match the detection to a pose prototype, and finally classify the corresponding activity.

For pose based action recognition we have to target three disjoint problems. We have to (a) detect a person in the image, (b) recognize the expressed pose, and (c) assign the pose to a suitable action category, see Figure 1 for an illustrative example. While we target all three problems in this work, the focus is on (b) and (c).

The contribution of this paper is threefold: (i) we present a complete approach for recognizing activities from single images and image sequences, (ii) we extend a *Histogram of Oriented Gradient* (HOG) [5] based pedestrian descriptor to account for articulated poses in cluttered images, (iii) we develop a histogram based action recognition approach that incorporates a weighting scheme for more distinctive poses.

The paper is organized as follows. Related work will be discussed in Section 2. In Section 3 we describe the concept of pose primitives. In Section 4, we introduce a pose based action recognition method. Finally, we present experimental results in Section 5.

2. Related work

The topic of action recognition from image sequences and still images gained increasing interest throughout the last years. Since it is beyond the scope of this paper to give a complete overview, we focus in **(a)** on contributions related to the idea of view/pose based action recognition. Besides,

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

we briefly summarize contributions related to the proposed pose representation/estimation in **(b)**.

(a) Having a closer look at the underlying features representing activities, we can spot two main classes, *dynamic features* [4, 11, 10, 15, 16] and *static pose based features* [6, 9, 14, 23, 20]. For image sequences, dynamic features are arguably more successful. Unfortunately, they can not be extracted from still images.

Regarding dynamic features, Blank et al. [4] use three dimensional space-time shapes extracted from silhouette images of humans for classifying activities. In [11], shape and motion cues are used for action recognition of two different actions in the movie "Coffee and Cigarettes". [10] introduces a biologically inspired action recognition approach which uses hierarchically ordered spatio-temporal feature detectors. Niebles et al. [15, 16] extend the *bag-of-features* concept to account for activity recognition. In [16] human action-categories are represented and learned using space-time interest points, in [15] a hierarchical bag of features approach is applied to image sequences.

Besides motion features, we can find static pose based representations where poses are often described by silhouettes [6, 9]. In [9], a *bag-of-rectangles* method is used for action recognition, effectively modeling human poses for individual frames and thereby recognizing various action categories. Goldenberg et al. [6] use *Principal Component Analysis* (PCA) to extract eigenshapes from silhouette images for behavior classification. Since silhouettes usually require background subtraction and are therefore rather restrictive, other static pose descriptors were suggested [14, 23, 20]. Lu et al. [14] represent actions of hockey players as PCA-HOG descriptors, action recognition is based on a set of predefined poses. Zhang et al. [23] recognize different actions by finding articulated poses in infrared images. They cluster a set of primitive poses based on HOG descriptors. In [20], we introduced a first approach towards action recognition using histograms of clustered HOG-descriptors of human poses.

(b) Regarding pose estimation, the methods mostly related to the presented paper are [23, 3, 1]. In [23], individual object detectors are trained for a number of poses. For dealing with varying backgrounds, a pixel weighting weights foreground (pose) and background pixels during training. Bissacco et al. [3] use a Latent Dirichlet Allocation (LDA) based segmentation of human poses represented by HOG descriptors. Agarwal et al. [1] estimate 3D human poses using a local basis representation extracted by means of *non-negative matrix factorization* (NMF). Similar to this paper, they apply NMF to a set of clean (no background clutter) human poses, and use the NMF weights to reconstruct novel poses. In contrast to [1], we include a set of background bases to further alleviate the influence of background clutter for pose estimation.

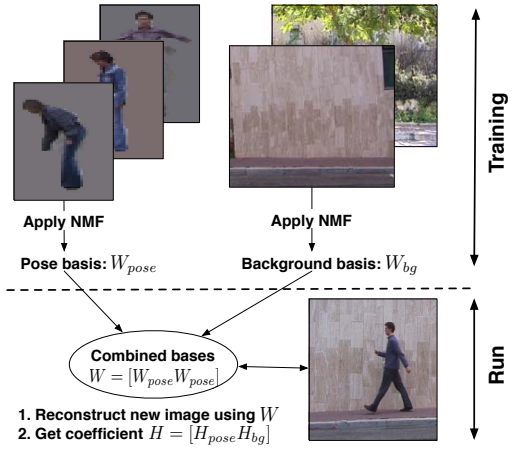

Figure 2. We extract a set of NMF bases W from clean pose images and background images. Novel images are reconstructed by these bases.

3. Pose representations

For pose based action recognition, a reliable representation and recognition of individual poses is crucial. Most difficulties in pose matching arise from cluttered background and pose articulations. Often, background objects are falsely recognized as limbs or parts of a pose. Following previously introduced ideas [14, 23], we recognize poses by matching them to a set of learned *pose primitives*. In the following, we will denote an individual pose primitive by $\mathbf{a_i}$, and a sequence of pose primitives by $\mathbf{A} = [\mathbf{a_1}, \ldots, \mathbf{a_n}]$.

As a basic feature for describing poses we use HOG descriptors [5]. HOG descriptors are locally normalized gradient histograms that showed to be a robust descriptor for pedestrian detection. Recent contributions also showed applicability for recognition of more articulated poses [3, 23]. Taking into account the mentioned problems for pose matching, we extend a standard HOG based pose descriptor to better cope with background clutter and articulated poses by exploiting a NMF basis representation of gradient histograms.

3.1. A NMF basis pose representation

Following [1], we learn a set of basis representations from human pose images with clean background by application of non-negative matrix factorization. See Figure 2 for examples of the training data and for an overview of the proposed pose representation. Applying NMF to a non-negative data matrix $\mathbf{V}$ (in our case the HOG descriptors for a set of training poses or backgrounds) leads to a factorization $\mathbf{V} \approx \mathbf{WH}$ where both $\mathbf{W}$ and $\mathbf{H}$ are constrained to be non-negative. In the following, we interpret $\mathbf{W}$ as a set of

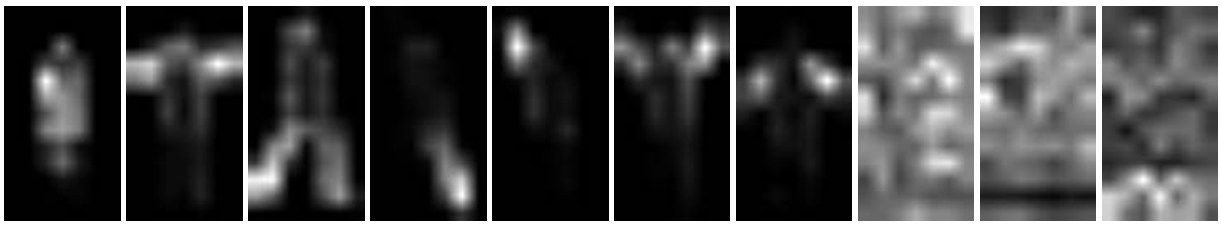

Figure 3. Visualization of basis vectors as the outcome of non-negative matrix factorization. The first seven images correspond to weights learned from human poses, the remaining 3 were learned from background images.

basis vectors and $\mathbf{H}$ as coefficients. Consequently, the original data matrix $\mathbf{V}$ can be reconstructed using $\mathbf{W}$ and $\mathbf{H}$. For image processing, unlike PCA or similar techniques, NMF can result in part based representations of objects. For example, in [12] face images were decomposed into a set of meaningful parts, i.e. ears, eyes, etc. Note that NMF will not always result in part-based representations. While this was not an issue in the presented work, future work might as well include the recently introduced NMF with sparseness constraint that is able to enforce part-based representations [8].To find an approximate factorization $\mathbf{V} \approx \mathbf{WH}$ we use the standard multiplicative update rule [13].

In contrast to [1], we aim at finding a larger variety of articulated poses. If reconstructing a gradient image solely from pose bases $\mathbf{W_{pose}}$, we found that also background clutter would often be reconstructed which is obviously not intended. Therefore, we apply two modifications: we compute the bases $\mathbf{W}$ for the whole image containing the pose, and we add a second set of bases $\mathbf{W_{bg}}$ that is computed from background images as well (see also Figure 2). This leads to the following reconstruction $\mathbf{V} = [\mathbf{W_{pose}}\mathbf{W_{bg}}][\mathbf{H_{pose}}\mathbf{H_{bg}}]$.

During training, pose $(\mathbf{W_{pose}}, \mathbf{H_{pose}})$ and background $(\mathbf{W_{bg}}, \mathbf{H_{bg}})$ parameters are learned independently of each other. Figure 3 visualizes some resulting bases. It can be seen that the $\mathbf{W_{pose}}$ represent meaningful parts of a pose. We found $40 - 80$ basis vectors sufficient for achieving reliable pose matching and accurate activity recognition.

For estimating the pose from a novel image $\mathbf{V^{new}}$, we first compute the coefficient $\mathbf{H^{new}}$ corresponding to $\mathbf{W} = [\mathbf{W_{bg}}\mathbf{W_{pose}}]$. For this, we use the standard iterative algorithm while holding $\mathbf{W}$ fixed. Thereby, the weights are combined so that they give the best possible explanation for $\mathbf{V^{new}}$ under the usage of $\mathbf{W_{bg}}$ and $\mathbf{W_{pose}}$.

Due to the strictly additive combination of weights in NMF, we usually have either a usage of background weights $\mathbf{W_{bg}}$ or pose weights $\mathbf{W_{pose}}$ for certain descriptor parts. The outcome are the coefficients $\mathbf{H^{new}}$ composed of separate coefficients for pose and background weights $\mathbf{H^{new}} = [\mathbf{H^{new}_{bg}}\mathbf{H^{new}_{pose}}]$. Effectively, we decouple the background appearance from the foreground by means of NMF basis reconstruction.

Note that the proposed extraction of background bases $\mathbf{W_{bg}}$ can not result in accurate modeling of arbitrary background appearances, due to a lack of training samples and insufficient number of bases. However, including $\mathbf{W_{bg}}$ considerably reduces the effect of falsely combining pose bases $\mathbf{W_{pose}}$ for modeling the background. While the background bases $\mathbf{W_{bg}}$ can only result in an approximate modeling of background appearances, the pose bases $\mathbf{W_{pose}}$ are sufficient for providing an accurate modeling of poses.

The resulting coefficients $\mathbf{H_{pose}}$ can now be compared against example poses or pose primitives. Pose primitives are extracted from a set of training sequences and their corresponding coefficients $\mathbf{H_{pose}}$. For clustering, we use the Euclidean distance and a standard agglomerative clustering method [22]. Usually, $30 - 80$ pose primitives lead to sufficient action recognition results, see also Section 5 for details on parameter selection. Finally, we get an index for the best matching pose primitive according to the minimum Euclidean distance, i.e. we express a novel image using the most similar training pose.

3.2. Human detection

In the following, we introduce a generative method for human detection. Since the focus of this paper is on action recognition, we by no means intend to compete with recent discriminative approaches for human detection (see e.g. [5]). However, it is still interesting to explore how the proposed NMF basis representation is usable for human detection. Note that the later presented experiments make use of the following method.

For detection we use a standard sliding window approach, i.e. we move a sliding window across the whole image and decide for each location whether it shows a human. In order to cope with scale changes we varied the detector window size and combined individual detector results. The NMF bases $[\mathbf{W_{pose}}\mathbf{W_{bg}}]$ represent two alternative models for reconstructing a gradient image. Thus, similar to [3], we can base human detection on the following likelihood ratio:

$$L = \frac{P(\mathbf{I}|bg)}{P(\mathbf{I}|pose)}, \tag{1}$$

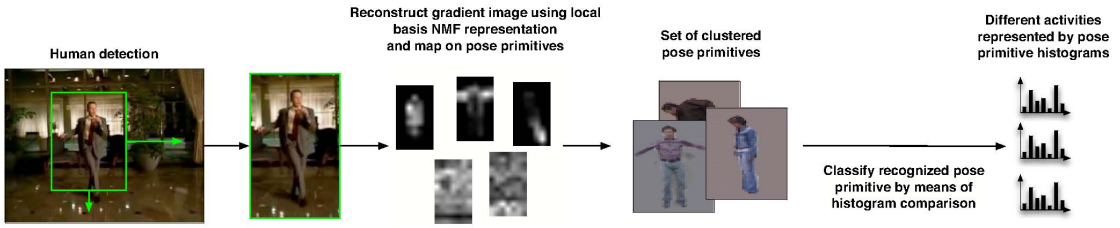

Figure 4. Action recognition using histograms of pose primitives.

where $P(\mathbf{I}|pose)$ and $P(\mathbf{I}|bg)$ denote the likelihood of $\mathbf{I}$ being modeled by $\mathbf{W_{pose}}$ or $\mathbf{W_{bg}}$ respectively. NMF reconstruction of $\mathbf{I}$ is given by $\mathbf{I} \approx \mathbf{WH}$, for the combined bases $\mathbf{W} = [\mathbf{W_{pose}}\mathbf{W_{bg}}]$, where $\mathbf{H}$ decomposes into $\mathbf{H} = [\mathbf{H_{pose}}\mathbf{H_{bg}}]$, thus $\mathbf{V} = \mathbf{V_{pose}} + \mathbf{V_{bg}} = \mathbf{W_{pose}}\mathbf{H_{pose}} + \mathbf{W_{bg}}\mathbf{H_{bg}}$. Since $\mathbf{V_{pose}}, \mathbf{V_{bg}} \leq \mathbf{V}$, and $\mathbf{V_{bg}} + \mathbf{V_{pose}} = \mathbf{V}$, we have $|\mathbf{V} - \mathbf{V_{bg}}|/|\mathbf{V}|, |\mathbf{V} - \mathbf{V_{pose}}|/|\mathbf{V}| \in [0, 1]$ and $|\mathbf{V} - \mathbf{V_{bg}}|/|\mathbf{V}| + |\mathbf{V} - \mathbf{V_{pose}}|/|\mathbf{V}| = 1$. Thus, we can express Equation (1) by the estimated coefficients $\mathbf{V_{pose}}\mathbf{V_{bg}}$

$$L = \frac{P(\mathbf{I}|bg)}{P(\mathbf{I}|pose)} \sim \frac{1 - |\mathbf{V} - \mathbf{V_{pose}}|/|\mathbf{V}|}{1 - |\mathbf{V} - \mathbf{V_{bg}}|/|\mathbf{V}|}. \quad (2)$$

The basic assumption for detecting humans via this likelihood ratio is that, in case there is a human in the image, the model accounting for human poses is contributing more to the overall reconstructed gradient image than the background model. Obviously, every image that contains a human also contains background. Setting a threshold of 1 which implies that the gradient image is equally constructed by pose and background bases worked surprisingly well in our experiments.

4. Action recognition

The task for action recognition is to infer an action class based on a single recognized pose primitive, or a sequence of recognized poses. For this, we propose a histogram based action recognition approach inspired by [7]. In contrast to [7], we focus on more primitive actions and consider clustered poses instead of predefined events. The underlying assumption is that we can construct arbitrarily complex actions by sequencing pose primitives. Besides this basic idea, we found that it is important to account for the degree of information per pose (some poses might contain more information about the underlying behavior than others), and to incorporate the local temporal context of pose primitives by providing a sub-sequencing of primitives by means of n-grams.

4.1. Pose histogram classification

Instead of directly analyzing the sequential ordering of pose primitives, we concentrate on the number of occurrences of specific poses, i.e. we classify by means of histogram comparison. Given a normalized histogram $\phi(\mathbf{A})$ for a sequence of pose primitives $\mathbf{A}$, we classify based on the minimum *Kullback-Leibler* (KL) divergence to a set of normalized training histograms $\phi(\mathbf{T}_i)$ of pose primitive sequences $\mathbf{T}_i$.

$$j = \underset{i}{\operatorname{argmin}} \, \mathrm{D_{KL}}(\phi(\mathbf{A}) \;||\; \phi(\mathbf{T}_i)), \quad (3)$$

where $i = 1, \ldots, n$ and

$$\mathrm{D_{KL}}(\phi(\mathbf{A}) \;||\; \phi(\mathbf{T}_i)) = \sum_k \mathbf{a}_k \log \left(\frac{\mathbf{a}_k}{\mathbf{a}_k^{T_i}} \right). \quad (4)$$

The histograms $\phi(T_i)$ are acquired during training where each histogram corresponds to one specific activity performed by one subject. Due to a limited supply of training histograms we usually used a simple 1-NN classifier. However, for future research and larger amounts of training data we might as well consider histogram aggregations [18].

For histogram comparison, the KL-divergence does not penalize zero bins of the query histogram [18]. This is important for the recognition of variable length pose primitive sequences. Here, each histogram bin corresponds to one recognized pose of a complete image sequence. Interestingly, the KL-divergence intuitively extends to the recognition of action classes from still images. For still images, the sequence $\mathbf{A}$ reduces to $\mathbf{A} = [\mathbf{a}_l]$ where l denotes the recognized pose primitive index, thus

$$j = \underset{i}{\operatorname{argmin}} \, \mathrm{D_{KL}}(\phi(\mathbf{A}) \;||\; \phi(\mathbf{T}_i)) = \underset{i}{\operatorname{argmin}} \log \frac{1}{\mathbf{a}_l^{\mathbf{T}_i}}, \quad (5)$$

where $\mathbf{a}_l^{\mathbf{T}_i}$ corresponds to the histogram entry for pose primitive $\mathbf{a}_l$ in the training histogram $\mathbf{T}_i$. Effectively, the same simple framework can be used for action recognition in still images as well as image sequences. Figure 4 summarizes the idea of histogram based activity recognition.

4.2. Pose primitive weightings

Intuitively, certain poses contain more information about the underlying behavior than others. For example, seeing a

person in a simple upright position could indicate almost any behavior, whereas we can immediately spot the distinctive poses of someone waving his arm.

In [21], classification of faces in a bag of features approach could been improved by weighting more informative fragment features. Since pose primitives are fragment features of a complete behavior, an adaption of that approach is straightforward. We reweight occurrences of pose primitives $\mathbf{a}_i$ in a histogram $\phi(\mathbf{A})$ of an action primitive sequence $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ by the summed likelihood ratio R of pose primitive $\mathbf{a}_i$ found in a specific behavior class $\mathbf{B}$, where

$$\mathrm{R}(\mathbf{a}_i) = \frac{P(\mathbf{a}_i|\mathbf{B})}{P(\mathbf{a}_i|\overline{\mathbf{B}})}, \tag{6}$$

where $\mathbf{B}_k$ corresponds to the combined histogram of all training sequences of a certain behavior k, and $\overline{\mathbf{B}_k}$ corresponds to the combined histogram of all training behavior histograms except $\mathbf{B}_k$. As a weight w_i for the bin corresponding to $\mathbf{a}_i$, we use

$$\mathrm{w}_i = \log\left(\sum_k R(\mathbf{a}_i)_{\mathbf{B}_k}\right) = \log\left(\sum_k \frac{P(\mathbf{a}_i|\mathbf{B}_k)}{P(\mathbf{a}_i|\overline{\mathbf{B}_k})}\right). \tag{7}$$

The weights are directly applied to pose primitive histograms

$$\phi(\mathbf{A}) = \mathbf{w}\phi(\mathbf{A}) = [\mathrm{w}_1\mathbf{a}_1, \ldots, \mathrm{w}_n\mathbf{a}_n]. \tag{8}$$

Reweighting individual bins results in an intuitive relevance weighting for pose primitive histograms.

4.3. Local temporal context

Activities are generally not defined by their content alone [7]. Ordering of events or specific poses allow to introduce context. An activity description solely based on unordered poses (or in our case histograms of poses) might lead to the wrong conclusion about the underlying behavior. Therefore, we found it beneficial to include the local temporal context, i.e. instead of having a look at individual pose primitives, we provide a subsequencing by means of n-gram expressions (see also [7] for a more comprehensive introduction to n-grams in the context of activity recognition).

n-grams are a common technique known from text mining or speech recognition. Essentially, they provide a subsequencing of length n, where the subsequences are, in our case, overlapping. For example, the bi-gram ($n = 2$) expression of a sequence of pose primitives $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]$ could be simply written as $\mathbf{A}_{bi-\mathrm{gram}} = [\{\mathbf{a}_1\mathbf{a}_2\}, \{\mathbf{a}_2\mathbf{a}_3\}, \ldots, \{\mathbf{a}_{n-1}\mathbf{a}_n\}]$. The maximum number of instances corresponding to all possible combinations of pose primitives is k^n, where k denotes to the number of individual pose primitives. Since we only consider subsequences that were observed during training, the actual number tends to be much lower and only represents a fraction of all possible n-gram instances.

Converting pose primitive sequences to sequences of n-gram instances does not require any modifications to the presented approach. Instead of computing histograms of pose primitives, we now compute histograms of n-gram instances. The underlying pose primitives as a descriptor for each frame stay the same. Concerning the subsequencing length, we usually got the best results for bi or tri grams, the later presented experiments use bi-grams. Note that it is impossible to extract n-grams from still images. Therefore, the in the next Section presented experiments for action recognition from still images use the conventional pose primitive histograms.

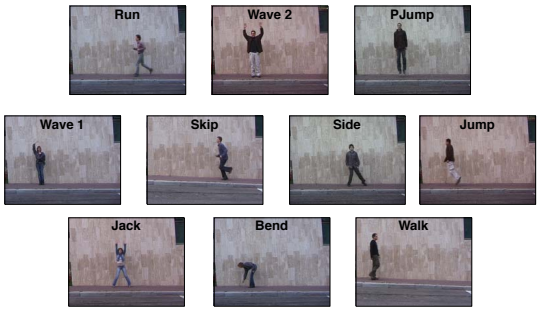

Figure 5. The Weizmann action-recognition data set includes 10 different behaviors performed by 9 subjects.

# pose primitives	precision
30	65.6 (80.0)
50	65.3 (81.0)
70	67.0 (78.8)
90	68.2 (84.4)
110	**70.4** (78.8)
130	70.3 (83.3)

Table 2. Average precision for action recognition from still images taken from the Weizmann data set. The values in brackets denote the average precision for a majority voting over all single frame results for one sequence.

5. Experiments

To verify the presented approach, we carried out a series of experiments. In **(a)** we test for action recognition precision for image sequences and still images. In **(b)** we explore how well our approach is able to deal with novel action categories, and in **(c)** we exploit the limitations of the proposed pose matching and present qualitative results. Regarding parameterization, for the HOG-descriptors we use a cellwidth/cellheight of 6, a detector window size of 78×42

Number of pose prototypes	10	30	50	70
1. NMF Clean pose images	50 (48.2)	67.7 (65.1)	72.2 (70.0)	67.7 (69.07)
2. NMF Clean pose images, relevance weighting	50 (50.7)	68.8 (67.2)	72.2 (70.0)	67.7 (68.83)
3. NMF Clean pose images & backgrounds	62.2 (58.14)	93.3 (91.23)	86.6 (88.67)	86.6 (87.16)
4. NMF Clean pose images & backgrounds, rel. weighting	61.1 (57.50)	**94.4 (92.2)**	87.7 (90.0)	86.6 (87.9)
5. NMF	52.2 (54.3)	86.6 (84.4)	91.1 (86.3)	83.3 (83.5)
6. NMF, relevance weighting	53.3 (55.9)	86.6 (84.6)	91.1 (86.2)	85.5 (83.6)

Table 1. The numbers correspond to the average precision for a leave one (subject) out cross validation scheme. We compared various methods of utilizing a NMF basis representation of poses. Computing separate bases for poses (from clean images) and backgrounds and the proposed relevance weighting gave the best performance. Interestingly, simply extracting pose primitives from the images with cluttered background came second (no. 6). This might be due to the rather simple background in the data set used.

Methods	(%)	(%) still images
This paper	94.40	70.4
Niebles et al. [15]	72.8	55.0
Thurau [20]	86.66	57.45
Blank et al. [4]	99.61	-
Ikizler et al. [9]	100.00	-
Jhuang et al. [10]	98.8	-
Ali et al. [2]	89.70	-

Table 3. Comparison of different approach evaluated on the Weizmann action recognition benchmark set [4]. Note that the cited papers all use slightly different evaluation schemes with variations in image sequence lengths and separation of test and training set. The most common evaluation method is leave one out cross validation and recognition of complete sequences. Besides [15], the approach presented is arguably the most flexible, since it can be applied to still images and does not require background subtraction or similar techniques.

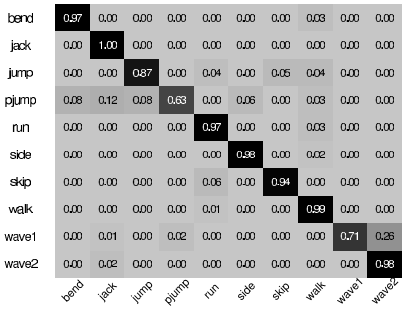

Figure 6. Confusion matrix showing per action class recognition results for subsequences of length 30.

pixel, and 3×3 cells per block where blocks are overlapping. For NMF we use 40 bases, and $10 - 20$ iterations for the multiplicative update rule in run-mode. In a first basic Matlab implementation we achieve about 1 FPS (dependent on the image resolution) in run-mod using integral histograms [17] to speed up HOG computation (without NMF computation this could be increased up to 5 FPS).

(a) In a first series of experiments, we used the well known Weizmann data set [4] for benchmarking the action

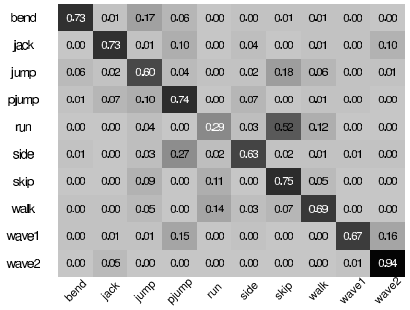

Figure 7. Detailed results for classifying still images taken from the Weizmann data set.

recognition approach. The data set contains 10 different behaviors performed by 9 subjects, see also Figure 5. We extracted a 78×42 pixel wide region around the center location of each subject. To focus on action recognition we did not apply the proposed human detection approach (it will be used later on). For the training phase, we also acquired each sequence without the background. For testing, we used the original 78×42 wide sequences including the background. We measured the average precision of leave one (subject) out cross validation series.

Table 1 summarizes the results for action recognition for image sequence for various variations in the used method. We used subsequences of length 20 (i.e. 20 frames). Using only 20 frame long sequences allows for a continuous recognition of primitive actions, which might be afterwards used to recognize more complex or higher level action classes. Here, classification of a whole sequence is based on a majority voting scheme applied to action classes recognized for each subsequence. The best precision of 94% was achieved for 30 pose primitives (*bi*-grams) and 40 NMF basis vectors for poses and backgrounds. Resulting confusion matrices are shown in Figure 6 and Figure 7. Interestingly, for individual activities the worst recognition occurs for the "jumping in one place" action. This might be due to the rather unspecific pose which simply shows a person standing and does not change significantly. Table 2 shows the results for classifying still images from the Weiz-

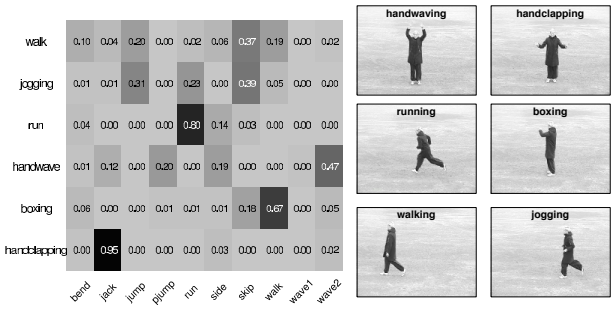

Figure 8. Results for classifying 6 image sequences taken from the KTH action data set against models trained on the Weizmann set.

mann data set, i.e. every single frame is separately classified. Here, we varied among the number of pose primitives where we achieved the best recognition rates of 70% for 110 pose primitives.

Table 3 compares results of recent approaches. It can be seen that recognition results are highly dependent on the used features. As already mentioned in Section 2, silhouettes or dynamic features usually perform best [4, 9, 9, 10]. However, our results of up to 94% come very close and exceed approaches more comparable to ours [15, 20]. For still images, a recognition accuracy of up to 70% outperforms previous approaches.

(b) In a second experimental setup, we tested the learned histograms/pose primitives against the KTH data set [19]. However, we decided to randomly select 6 samples sequences, one for each category [1]. We learned the models based on the Weizmann data set, and classified against the unknown action categories from the KTH set. The test videos were slightly scaled, for human detection the proposed likelihood-ratio was used.

Here, we were interested in how our approach is dealing with completely novel action categories. Again, we classified overlapping subsequences of length 20. The results are summarized in Figure 8. Three out of six behaviors (handclapping, handwaving, run) were associated with the intuitively best matching action class from the Weizmann set. The remaining 3 are associated with varying classes, e.g. jogging is confused with jumping on one leg (skip) and running. However, a closer look at the used sequences shows a larger difference than what could be expected from the action class labeling.

(c) In a third experimental setup, we solely concentrated on pose matching. Here, we wanted to exploit the limits of the proposed exemplar based human pose estimation. For training, we used the same data as in the last experiment. As test data we used sequences from two music clips, "A Road to Nowhere" (*The Talking Heads*) and "Weapon of Choice" (*Fatboy Slim*). In both clips, the proposed human detection scheme was applied. For "A Road to Nowhere", due to the original image size, we cut out a region centered around a specific interesting area which contains a running person in front of an ever changing background.

Since the music clips provide a challenging data set, we did not expect a perfect matching onto pose primitives. In fact, most of the poses in the clips do not have an appropriate match in the Weizmann data set. Also the backgrounds used for extracting NMF background bases are completely different. Interestingly, we could observe not only sufficient human detection using the proposed approach, but also reasonable matching to pose primitives. Figure 9 shows some exemplary matches. Despite the heavily cluttered or changing backgrounds, pose estimations appear reasonable for most cases. It also showed that the approach is limited by the availability of pose primitives, this shows especially for the complex dancing scene with Christopher Walken. Effectively, we can only sufficiently recognize poses that were learned during the training phase.

[1] We picked the example sequences presented on `http://www.nada.kth.se/cvap/actions/`

6. Conclusion

We presented a pose based approach for action recognition from still images and image sequences. The approach does not require background subtraction or a still camera, and can be easily extended to multiple persons. Experimental results on publicly available benchmark data shows a high accuracy for action recognition.

The experiments presented indicate that the pose of a human already contains sufficient information about the underlying activity. While we believe that additional information gained by dynamic features could result in a better precision for activity recognition, it is interesting to see that we can neglect motion features to a certain extent.

7. Acknowledgments

The authors would like to thank Andrew Zisserman for helpful discussions. Christian Thurau was supported by a grant from the European Community under the EST Marie Curie Project VISIONTRAIN MRTN-CT-2004-005439. The second author was supported by the project MSM6840770013 and by the EC project DIPLECS No. 215078.

References

[1] A. Agarwal and B. Triggs. A Local Basis Representation for Estimating Human Pose from Cluttered Images. In *ACCV'06*, 2006.

[2] S. Ali, A. Basharat, and M. Shah. Chaotic Invariants for Human Action Recognition. In *ICCV'07*, 2007.

[3] A. Bissacco, M. H. Yang, and S. Soatto. Detecting Humans via Their Pose. In *NIPS'06*, 2006.

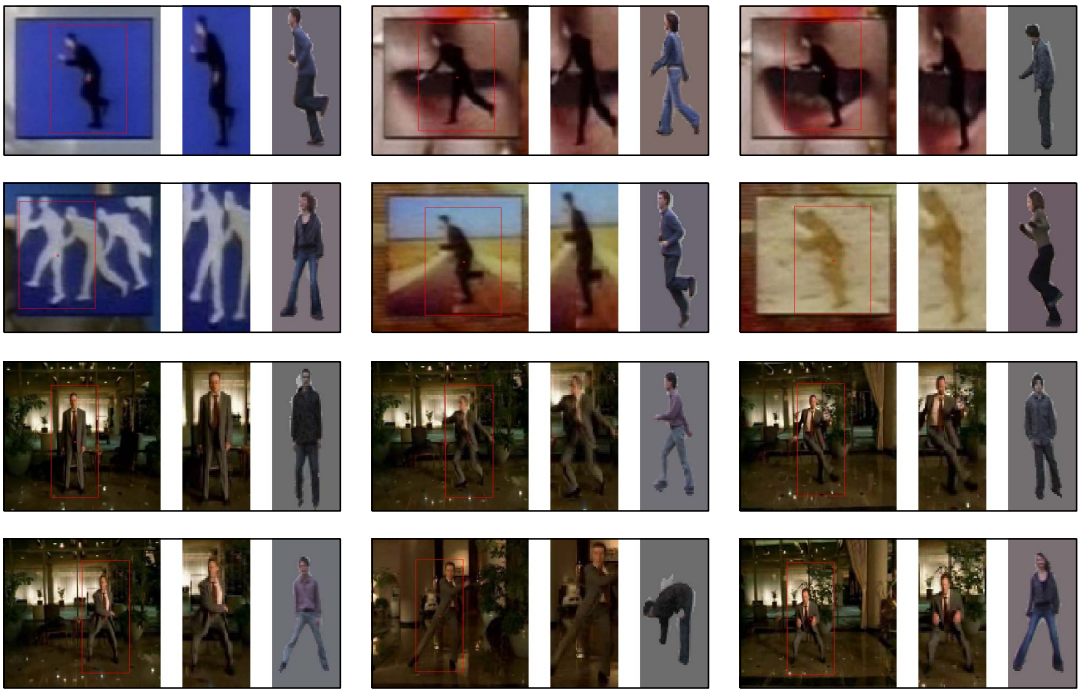

Figure 9. Pose matching results for the music clips "Road to nowhere" by *The Talking Heads* (upper two rows) and "Weapon of Choice" by *Fatboy Slim* (lower two rows). Pose primitives were extracted from the Weizmann data set [4]. Human detection is achieved using a simple ratio of NMF reconstructions. The first clip is difficult due to the ever changing background and the changing human appearances. The second clip is challenging due to pose articulations, camera movement, and the similar colors of Christopher Walken's suit and the background.

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. In *ICCV'05*, 2005.

[5] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR'05*, 2005.

[6] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky. Behavior classification by eigendecomposition of periodic motions. *Pattern Recognition*, 38:1033–1043, 2005.

[7] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman. Detection and Explanation of Anomalous Activities: Representing Activities as Bags of Event n-Grams. In *CVPR'05*, 2005.

[8] P. O. Hoyer. Non-negative Matrix Factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[9] N. Ikizler and P. Duygulu. Human Action Recognition Using Distribution of Oriented Rectangular Patches. In *Human Motion ICCV'07*, 2007.

[10] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A Biologically Inspired System for Action Recognition. In *ICCV'07*, 2007.

[11] I. Laptev and P. Perez. Retrieving actions in movies. In *ICCV'07*, 2007.

[12] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–799, 1999.

[13] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorizationi. In *NIPS'01*, 2001.

[14] W. L. Lu and J. J. Little. Simultaneous Tracking and Action Recognition using the PCA-HOG Descriptor. In *CRV'06*, 2006.

[15] J. C. Niebles and L. Fei-Fei. A Hierarchical Model of Shape and Appearance for Human Action Classification. In *CVPR'07*, 2007.

[16] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. In *BMVC'06*, 2006.

[17] F. Porikli. Integral histogram: A fast way to extract higtograms in cartesian spaces. In *CVPR'05*, 2005.

[18] F. Schroff, A. .Criminisi, and A. Zisserman. Single-Histogram Class Models for Image Segmentation. In *ICVGIP'06*, 2006.

[19] C. Schuldt, I. Laptev, and B. Caputo. Recognizing Human Actions: A Local SVM Approach. In *ICPR'04*, 2004.

[20] C. Thurau. Behavior Histograms for Action Recognition and Human Detection. In *Human Motion ICCV'07*, 2007.

[21] S. Ullman, M. Vidal-Naquet, and E. Sali. Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, 5:682–687, June 2002.

[22] J. H. J. Ward. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, 58:236–244, 1963.

[23] L. Zhang, B. Wu, and R. Nevatia. Detection and Tracking of Multiple Humans with Extensive Pose Articulation. In *ICCV'07*, 2007.