Learning Spatial Interest Regions from Videos to Inform Action Recognition in Still Images

A. Eweiwi[1], M.S. Cheema[1], and C. Bauckhage[1,2]

[1]B-IT, University of Bonn, Bonn, Germany
[2]Fraunhofer IAIS, Sankt Augustin, Germany

Abstract. Common approaches to human action recognition from images rely on local descriptors for classification. Typically, these descriptors are computed in the vicinity of key points which either result from running a key point detector or from dense or random sampling of pixel coordinates. Such key points are not a-priori related to human activities and thus of limited information with regard to action recognition. In this paper, we propose to identify action-specific key points in images using information available from videos. Our approach does not require manual segmentation or templates but applies non-negative matrix factorization to optical flow fields extracted from videos. The resulting basis flows are found to to be indicative of action specific image regions and therefore allow for an informed sampling of key points. We also present a generative model that allows for characterizing joint distributions of regions of interest and a human actions. In practical experiments, we determine correspondences between regions of interest that were automatically learned from videos and manually annotated locations of human body parts available from independent benchmark image data sets. We observe high correlations between learned interest regions and body parts most relevant for different actions.

Keywords: optical flow, non-negative matrix factorization.

1 Introduction

Research on recognizing human activities from still images is motivated by promising applications in automatic indexing of very large image repositories and also contributes to problems in automatic scene description, context dependent object recognition, or human pose estimation [4, 13, 25, 28].

Currently, approaches to action recognition can be categorized into two main classes: (a) pose-based and (b) bags-of-features (BoF) methods. Stirred by the idea of *poselets* [3], a notion of part-based templates, pose-based approaches have recently been met with rekindled interest [21, 25]. However, the construction of

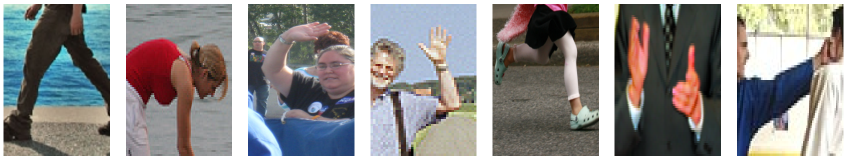


Fig. 1. Examples of still images in which we can easily recognize human activities even if neither image shows all of the human body.

poselets requires cumbersome manual annotations which impede their use in BIG DATA settings. BoF approaches are known for their good performance in object recognition and have therefore been adapted to action recognition [5]. Yet, local image descriptors are typically computed in the vicinity of key points that result from low-level signal analysis or dense or random sampling and are therefore uninformative or independent of the activity depicted in an image.

Most physical activities of people show characteristic articulations and movements of different body parts. Yet, although activities are inherently dynamic, the human visual system easily infers human activities from still images that show posture or limb configurations. Consider, for instance, the images in Fig. 1 which we can interpret even without a full view of the human body. This raises the question if it is possible to automatically learn or identify action-specific, informative, regions of interest in still images without having to rely on exhaustive mining of low-level image descriptors or labor-intensive annotations?

In an attempt to answer this question, we propose an efficient approach towards automatically learning of action specific regions of interest in still images. Considering the fact that activities are temporal phenomena, we make use of information available from videos. Given videos that show human activities, we compute optical flow fields and consider the magnitudes of flow vectors in each frame. Given a collection of frame-wise flow magnitudes, we apply non-negative matrix factorization (NMF) and obtain basis flows. These basis flows are indicative of the position and configuration of different limbs or body parts whose motion characterizes certain activities. Viewed as images, the basis flows indicate action specific regions of interest and therefore allow for an informed sampling of key points for subsequent feature extraction. We also devise a generative probabilistic model that characterizes joint distributions of regions of interest and human actions. To evaluate our approach, we consider correspondences between regions of interest that were automatically learned from videos and manually annotated locations of human body parts that are available from independent data sets of still images. Our empirical results reveal a high correlation between extracted interest regions and those body parts that are most relevant for different actions.

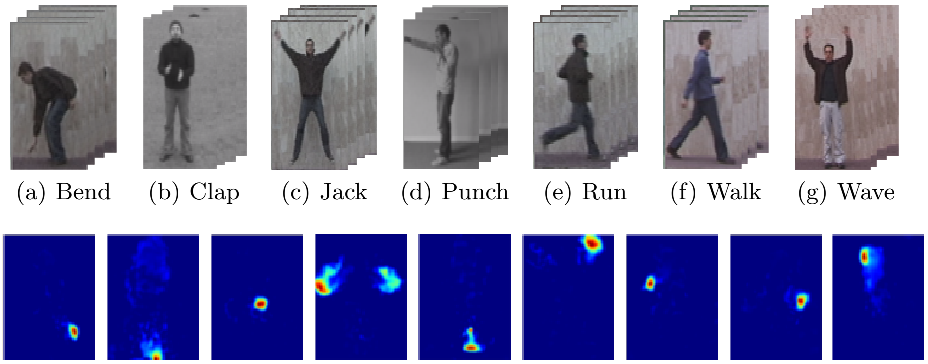
2 Related Work

Approaches that rely on the idea of bags of visual words (BoWs) are popular because BoWs are known for their simplicity, robustness, and good performance in content-based image or video classification. Corresponding work treats an image as a collection of independent visual descriptors computed at key point locations. Computing key points is crucial within the BoW framework since it preselects image patches subsequent classification. Naturally, one would like to focus only on those patches that are most discriminative.

So far, BoW approaches [18, 22] based on key points detection [2, 12, 20, 23], though generally discriminative, do not regard task specific objectives in key point localization. Rather, key point locations are determined from low-level properties of the image signal. Moreover, corresponding approaches typically assume key points to be independent and therefore fail to explain characteristics of spatial layouts. The work in [15] therefore proposes a representation that encodes spatial relationships among key points. The authors of [11] employ data mining to build high-level compound features from noisy and over-complete sets of low-level features and the work [24] uses a triangular lattice of grouped point features to encode spatial layouts. Still, these approaches, too, center around low-level signal properties which do not necessarily provide an accurate account of the characteristics of an activity.

Sampling techniques such as random sampling have shown good performances, too. The authors of [7] empirically demonstrate that random sampling provides equal or better activity classifiers than sophisticated multi-scale interest point detectors; yet, their work also illustrates that the most important aspect of sampling is the number of sample points extracted. The authors of [27] claim that dense sampling outperforms all point detectors in realistic scenarios and. Yet, at the same time, recent work in [10] shows that state-of-the-art performance in action recognition can also be obtained from only a few randomly sampled key points. It therefore appears that the jury is still out on whether to use dense or random sampling and methods which mark a middle ground, namely informed sampling, seem to merit closer investigation. It is, however, obvious that the success of dense sampling is bought at the expense of memory- and runtime efficiency whereas random sampling methods do not provide statistical guarantees as to the adequacy for the task at hand.

Part-based approaches, too, are popular in research on action recognition and have been shown to successfully cope with the *PASCAL* challenge. The authors of [9] describe a deformable model which achieves good performance on benchmark data sets [5]. The work in [3] introduces exemplar-based pose representation, or *poselets*, for human detection. This term denotes a set of patches with similar pose configurations. The work in [21] utilizes poselets for identifying human poses and actions in still images and the authors of [25] propose an articulated part-based model for human pose estimation and detection which adapts a hierarchical (coarse-to-fine) representation. Despite their recent success, it is still questionable if these methods can make use of the favorable statistics of


(a) Bend (b) Clap (c) Jack (d) Punch (e) Run (f) Walk (g) Wave

(h) Examples of basis vectors obtained from NMF

Fig. 2. (a–g) Examples of training videos from the Weizmann and KTH data sets; (h) examples of basis flows obtained from applying NMF to optical flow fields.

present day large scale data sets because the construction of suitable poselets requires extensive human intervention and manual labeling in the training phase.

The authors of [26] consider non-negative matrix factorization (NMF) for action recognition and apply it to learn pose- and background primitives. In [1], the authors estimate the human upper body pose through NMF and [16, 17] apply non-negative factor models to recognize activities from videos. The authors of [29] empirically evaluate human action recognition using pose- or appearance-based features and conclude that, even for rather coarse pose representations, pose-based features either match or outperform appearance-based features. However, they acknowledge that appearance-based features still represent an ideal resort for cases of considerable visual occlusion. Accordingly, it appears worthwhile to study methods that combine both approaches into a single framework.

Next, we discuss how the approach proposed in this paper indeed provides a method for the informed sampling of key points for appearance-based action recognition as well as an approach to learning descriptors of body poses.

3 Learning Action-specific Interest Regions from Videos

Our approach identifies discriminative regions in an image and subsequently learns the relative importance of those regions for different actions. In order to identify interesting spatial locations, we apply NMF to optical flow fields obtained from videos. Furthermore, we exploit NMF mixture coefficients to derive a generative probabilistic model that features joint distributions of regions of interest and human actions.

3.1 Learning NMF Bases

Given videos of different actions, we determine optical flow magnitudes at each pixel in a box of constant size surrounding a person visible in the video. Each frame can be transformed into an m dimensional non-negative vector $\mathbf{v}$. Let n_i represent the number of frames for an action $a_i \in \mathcal{A} = \{a_1, a_2, ..., a_r\}$ and let $n = \sum_{i=1}^{r} n_i$. We build an $m \times n$ data matrix $\mathbf{V}$ containing the flow magnitude vectors of all frames. Computing NMF yields k basis vectors, or *basis flows*, such that $\mathbf{V} \approx \mathbf{WH}$ where the columns of $\mathbf{W}_{m \times k}$ are non-negative basis elements and the columns of $\mathbf{H}_{k \times n}$ encode non-negative mixing coefficients.

In order to compute the factors $\mathbf{W}$ and $\mathbf{H}$, we apply the algorithm according to Lee and Seung [19]. This method is known to yield sparse basis elements for it converges to vectors that lie in the facets of the simplicial cone spanned by the data (see the discussions in [6, 14]). Accordingly, we can expect the resulting basis flows to be sparse in the sense that most elements of a basis element $\mathbf{w}_l$ will be (close to) zero and only a few entries will have noticeable values. Figure 2 (h) shows that this is indeed the case. It depicts pictorial representations of exemplary basis vectors $\mathbf{w}_l$ resulting from NMF. Note that for each basis element only a few pixels are larger than zero; in each case, these pixels apparently form distinct, more or less compact patches in the image plane.

3.2 Learning the Action-specific Importance of Basis Flows

Different actions are characterized by articulation and movements of different body parts. The NMF basis vectors determined through factorization of frame-wise optical flow magnitudes appear to indicate image regions of importance for different activities. Here, we propose to learn the relative importance of different basis elements with respect to different actions. To this end, we consider the matrix $\mathbf{H}$ since its entries encode linear mixing coefficients required to reconstruct the vectors in $\mathbf{V}$ from the basis flows in $\mathbf{W}$. Consequently, the columns of $\mathbf{H}$ encode the relevant importance of a basis for a given frame. Normalizing them to stochastic vectors allows us to estimate a joint probability distribution of actions and bases. The conditional probability of basis $\mathbf{w}_l$ given an action a_i is determined as

$$p(\mathbf{w}_l | a_i) = \frac{\sum\limits_{f} h_{lf}}{\sum\limits_{j,f} h_{jf}} \tag{1}$$

where the summation index f indicates all columns $\mathbf{v}_f$ in $\mathbf{V}$ that show activity a_i and index j ranges from 1 to k.

Figure 3 plots the resulting distribution. Note that this probability distribution, i.e. the set of weights of a basis element w.r.t. an action, again is sparse. The distribution in equation (1) immediately allows us to determine how characteristic a certain basis flow is for an activity.

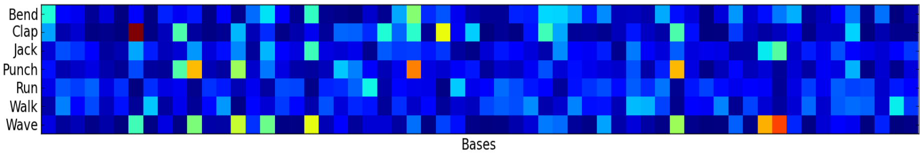


Fig. 3. Relative importance of bases w.r.t. different actions according to $p(\mathbf{w}_l|a_i)$. Note that actions flows can be approximated by a small number of basis vectors.

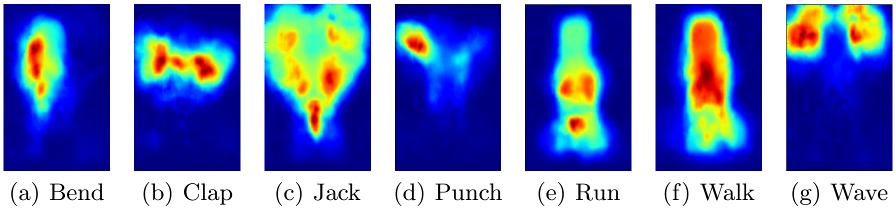

(a) Bend (b) Clap (c) Jack (d) Punch (e) Run (f) Walk (g) Wave

Fig. 4. Examples of action signatures resulting from equation (2).

The probability distribution $p(\mathbf{w}_l|a_i)$ in (1) also allows us to consider *action signatures* which we define to be the conditional expectations

$$\mathbf{s}_i = \sum_{l=1}^{k} p(\mathbf{w}_l|a_i)\,\mathbf{w}_l. \tag{2}$$

Computing and plotting action signatures s_i for different actions a_i, we find that characteristically different regions in the image plane are intensified for different actions. Figure 4 shows examples of action signatures which we obtained from basis flows extracted from the Weizmann[1] and KTH[2] data sets. Apparently, action signatures like these may serve two purposes. On the one hand, they provide us with a prior distribution for the sampling of interest points from still images showing people in order to compute action specific local features for activity classification. On the other hand, action signatures may be used as templates or filter masks for pose-based activity recognition.

3.3 Evaluation Methodology

To evaluate as to how far regions of interest extracted by our approach match the locations of human body parts in real images, we consider the manually annotated positions of limbs that are available in the H3D[3] and VOC2011[4] data sets. In particular, we determine the joint probability distribution of actions,

[1] www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

[2] www.nada.kth.se/cvap/actions/

[3] www.eecs.berkeley.edu/~lbourdev/h3d/

[4] pascallin.ecs.soton.ac.uk/challenges/VOC/voc2011/

interest regions, and body parts. Given the locations of a body part b_j in an image of action a_i, we assume the following conditional independence model

$$p(b_j, \mathbf{w}_l, a_i) = p(b_j|a_i)\, p(\mathbf{w}_l|a_i)\, p(a_i). \tag{3}$$

Using (1) and taking the prior $p(a_i)$ to be uniform, allows for solving for $p(b_j|a_i)$ which can be understood to encode the relative importance of different body parts for different actions a_i.

4 Experimental Results

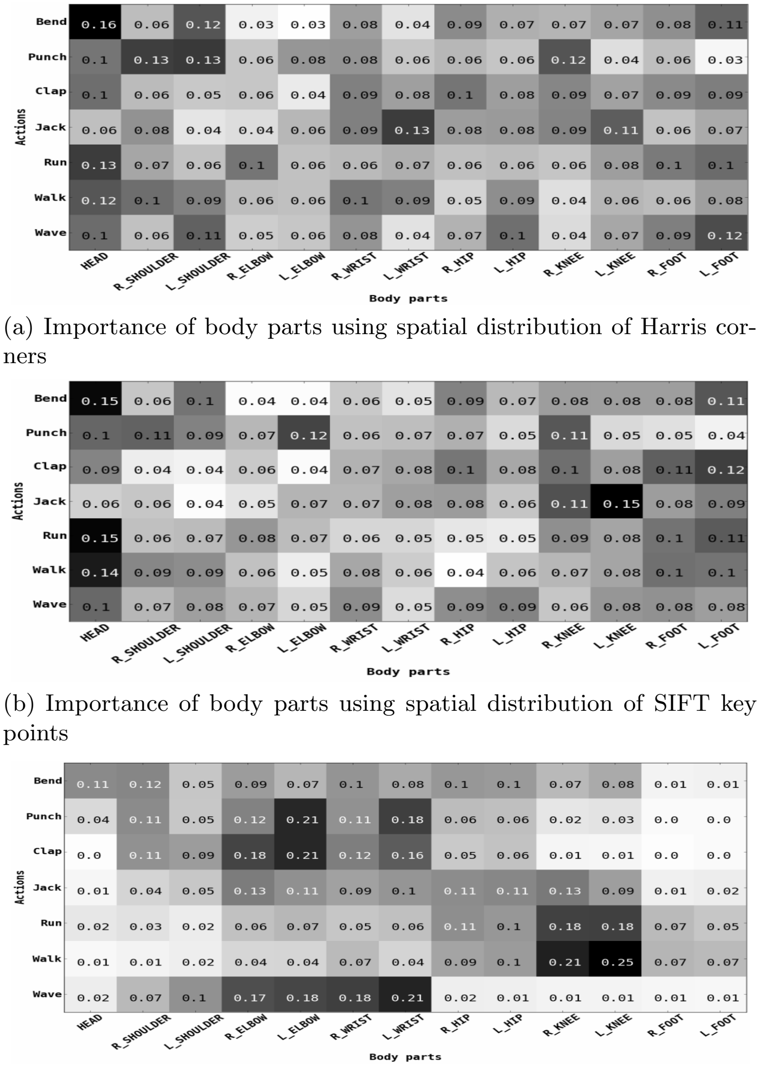
In order to learn action specific regions of interest, we considered the Weizmann and KTH data sets. As these video collections show little variations of background and view-point, they allow us to focus on estimating the importance of different body parts for different actions. In particular, we focused on the following actions *Bend*, *Clap*, *Jack*, *Punch*, *Run*, *Walk*, and *Wave*. We used the bounding boxes provided by [30] and resized them to size 88×64. To determine optical flows, we considered the method due to Farnebäck [8]. Finally, all of the results reported below were obtained using 200 basis flows $\mathbf{w}_i$.

To evaluate the suitability of the resulting interest regions for still image based action recognition, we considered limb or joint annotations available in the H3D and VOC2011 data sets. We used 240 annotated images and determined the joint distribution of actions, interest regions, and body parts. For each of the selected action classes, we considered the location of 13 body parts or joints including, for example, head, feet, knees, hips, shoulders, elbows, and hands.

We compared our interest regions to key points extracted by the popular Harris [12] and SIFT [20] key point detectors. In each case, we selected key points with the highest response in every image, assigned them to their nearest annotated body part, and normalized the resulting histogram. For each action, we obtained a stochastic vector by iterating over all images of that action thus representing the conditional distribution $p(b_j|a_i)$ discussed above.

Figure 5 compares results due to our approach of extracting interesting regions from video data to the ones obtained from using Harris and SIFT key points. It shows the relative importance of different body parts for different actions. In case of Harris and SIFT key points, head and feet dominate other limbs regardless of the action (Fig. 5 (a) and (b)). Furthermore, the probabilities for other body parts are almost uniform and do not convincingly relate to the different actions. For example, body parts naturally characterizing *Clap*, i.e. elbows, and hands, achieved rather low scores.

On the other hand, our approach exhibits logically coherent relationships between body parts and actions (Fig. 5(c)). Compare, for instance, the varying importance of different body parts for *clapping* and *running*. Clearly the lower body parts are dominant for the action of running while the arms are of higher importance for the action of clapping. From the perspective of body parts observe that, for instance, the head is less relevant for actions such as *Clap* or *Run* as compared to *Bend*. Figure 6 visualizes these results using stick figures where



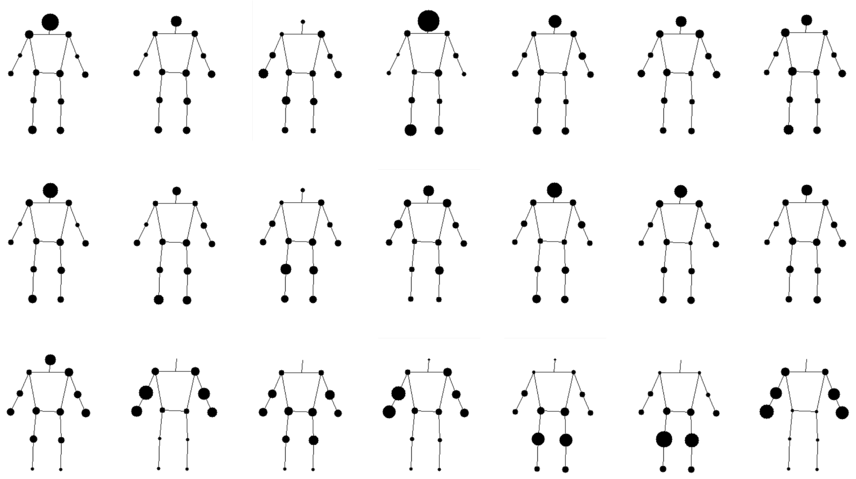
(a) Importance of body parts using spatial distribution of Harris corners

(b) Importance of body parts using spatial distribution of SIFT key points

(c) Importance of body parts using our approach

Fig. 5. Conditional probabilities of body parts w.r.t actions. Our approach (c) exhibits logically coherent relationships between body parts and actions as compared to appearance based sampling using Harris corners (a) and SIFT interest points (b).

the size of plotted body parts correspond to their relevance for an activity. In general these results suggest that the regions of interest which we obtain from factorizing flow fields are well correlated with the locations of action specific body parts available from independent sets of manually annotated images.



(a) Bend (b) Clap (c) Jack (d) Punch (e) Run (f) Walk (g) Wave

Fig. 6. Stick figures depicting the relevance of different body parts for different actions. Important key points computed using the Harris detector (first row) and SIFT detector (second row) hardly correlate to action-specific body parts; interest regions from our approach correlate better (third row).

5 Conclusion and Future Work

We presented an approach to the automatic detection of regions of interest for human action recognition in still images. Since human activities are inherently dynamic in nature, we proposed to learn interest regions from optical flow fields extracted from video sequences of human actions. Using non-negative matrix factorization, we obtained sets of basis flows which were found to be indicative of the location of different limbs or joints in different activities. Our approach fundamentally differs from existing pre-processing approaches for action recognition in still images. First, although we consider rather low-level properties of videos of activities, the characteristics of optical flow enable us to identify locations of body parts whose articulation define an action. Consequently, unlike common bag-of-features approaches, our approach facilitates informed sampling of key points in the image plane. Second, the proposed concept of action signatures provides probabilistic templates for pose-based recognition. Compared to common approaches based on distributed pose representations, our approach does not require meticulous manual annotation of images or frames and thus offers more scalability and convenience for large data sets. Also, compared to conventional part based approaches, our approach does not assume an underlying elastic model of body but provides priors even for cluttered or occluded images. This paper therefore established a baseline for video-based feature selection towards action recognition in still images.

The logical next step for future work is, of course, to build activity classifiers based on information available from action specific regions of interest. To this end, we currently consider standard descriptors(e.g. HOG, SIFT, SURF) which are computed at locations determined according to the probabilities encoded in action signatures.

References

1. Agarwal, A., Triggs, B.: A local basis representation for estimating human pose from cluttered images. In: ACCV (2006)
2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. CVIU 110(3), 346–359 (2008)
3. Bourdev, L., Malik, J.: Poselets: body part detectors trained using 3d human pose annotations. In: ICCV (2009)
4. Cheema, M., Eweiwi, A., Thurau, C., Bauckhage, C.: Action recognition by learning discriminative key poses. In: ICCV Workshop on Performance Evaluation of Recognition of Human Actions (2011)
5. Deltaire, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: A study of bag-of-features and part-based representations. In: BMVC (2010)
6. Donoho, D., Stodden, V.: When Does Non-negative Matrix Factorization Give a Correct Decomposition into Parts? In: NIPS (2004)
7. E.Nowak, Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: ECCV (2006)
8. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: SCIA (2003)
9. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. TPAMI 32, 1627 – 1645 (2010)
10. Gall, J., Yao, A., Razavi, N., Van Gool, L., Lempitsky, V.S.: Hough forests for object detection, tracking, and action recognition. TPAMI 33, 2188–2202 (2011)
11. Gilbert, A., Illingworth, J., Bowden, R.: Action recognition using mined hierarchical compound features. TPAMI 33, 883 – 897 (2011)
12. Harris, C., Stephens, M.: A combined corner and edge detection. In: In Alvey Vision Conference (1988)
13. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR (2011)
14. Klingenberg, B., Curry, J., Dougherty, A.: Non-negative matrix factorization: Ill-posedness and a geometric algorithm. PR 42(5), 918–928 (2008)
15. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: CVPR (2010)
16. Krausz, B., Bauckhage, C.: Action recognition in videos using nonnegative tensor factorization. In: ICPR (2010)
17. Krausz, B., Bauckhage, C.: Loveparade 2010: Automatic video analysis of a crowd disaster. CVIU 116(3), 307–319 (2012)
18. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR (2008)
19. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–799 (1999)
20. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)

21. Maji, S., Bourdev, L., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: CVPR (2011)
22. Matikainen, P., Hebert, M., Sukthankar, R.: Trajectons: action recognition through the motion analysis of tracked features. In: ICCV Workshop on Video-Oriented Object and Event Classification (2009)
23. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. IJCV 37, 151–172 (2000)
24. Song, Y., Goncalves, L., Perona, P.: Unserpervised learning of human motion. TPAMI 25, 814 – 827 (2003)
25. Sun, M., Savarese, S.: Articulated part-based model for joint object detection and pose estimation. In: ICCV (2011)
26. Thurau, C., Hlavac, V.: Pose primitive based human action recognition in videos or still images. In: CVPR (2008)
27. Wang, H., Ullah, M.M., Klaeser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC (2009)
28. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. CVIU 115, 224–241 (2011)
29. Yao, A., Gall, J., Fanelli, G., Van Gool, L.: Does human action recognition benefit from pose estimation? In: BMVC (2011)
30. Yao, A., Gall, J., Van Gool, L.: A hough transform-based voting framework for action recognition. In: CVPR (2010)