Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Action recognition in still images by learning spatial interest regions from videos[☆]

Abdalrahman Eweiwi[a,*], Muhammad Shahzad Cheema[a], Christian Bauckhage[a,b]

[a] Bonn Aachen International Center for IT, University of Bonn, Germany
[b] Fraunhofer Institute for Intelligent Analysis and Information Systems, Sankt Augustin, Germany

ARTICLE INFO

Article history:
Received 18 September 2013
Available online 14 August 2014

Keywords:
Action recognition
Non-negative matrix factorization
Pose estimation
Optical flow

ABSTRACT

A common approach to human action recognition from still images consists in computing 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 sampling of pixel coordinates. Such key points are not a priorly related to human activities and thus might not be very informative with regard to action recognition. Several recent approaches, on the other hand, are based on learning person–object interactions and saliency maps in images. In this article, we investigate the possibility and applicability of identifying action-specific points or regions of interest in still images based on information extracted from video data. In particular, we propose a novel method for extracting spatial interest regions where we apply non-negative matrix factorization to optical flow fields extracted from videos. The resulting basis flows are found to indicate image regions that are specific to certain actions and therefore allow for an informed sampling of key points for feature extraction. We thus present a generative model for action recognition in still images that allows for characterizing joint distributions of regions of interest, local image features (visual words), and human actions. Experimental evaluation shows that (a) our approach is able to extract interest regions that are highly correlated to those body parts most relevant for different actions and (b) our generative model achieves high accuracy in action classification.

1. Introduction

Throughout the last decade, the problem recognizing human activities from still images has found considerable attention. Corresponding research is motivated by promising applications in areas such as automatic indexing of very large image repositories but is also expected to contribute to the solution of problems in automatic scene description, context dependent object recognition, or pose estimation [33,19,38].

Based on the underlying problem formulation, approaches to action recognition can be categorized into two main classes: (a) pose-based and (b) bag-of-features (BoF) approaches. Motivated by the idea of *poselets* [4], a notion of distributed part-based templates, pose-based approaches have recently been met with rekindled interest [40,27,43]. But the construction of poselets still requires a cumbersome procedure of manual annotation which impedes their use on large training sets. BoF approaches based on local descriptors are known for their state-of-the-art performance in object recognition and therefore have been adapted to action recognition [9]. However, local image descriptors are typically computed in the vicinity of key points that result from low-level signal analysis or from dense or random sampling and are therefore uninformative and independent of the activity depicted in an image. Several recent approaches are based on the idea of learning interaction between people and objects using saliency maps in images [31] or videos [6,5]. In this article, we presents a novel approach in this direction.

Most physical activities of people are characterized by articulation and movement of different body parts. And although activities are inherently dynamic, most people can easily infer human activities in still images just by looking at posture or configuration of particular body parts. Consider, for instance, the images shown in Fig. 1 which we can interpret even without having 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 yet effective approach towards automatic learning of action

[☆] This paper has been recommended for acceptance by S. Wang.

[*] Corresponding author. Tel.: +49 228 2699131.

E-mail addresses: eweiwi@bit.uni-bonn.de (A. Eweiwi), cheema@bit.uni-bonn.de (Muhammad Shahzad Cheema), christian.bauckhage@iais.fraunhofer.de (Christian Bauckhage).

http://dx.doi.org/10.1016/j.patrec.2014.07.017
0167-8655/© 2014 Elsevier B.V. All rights reserved.

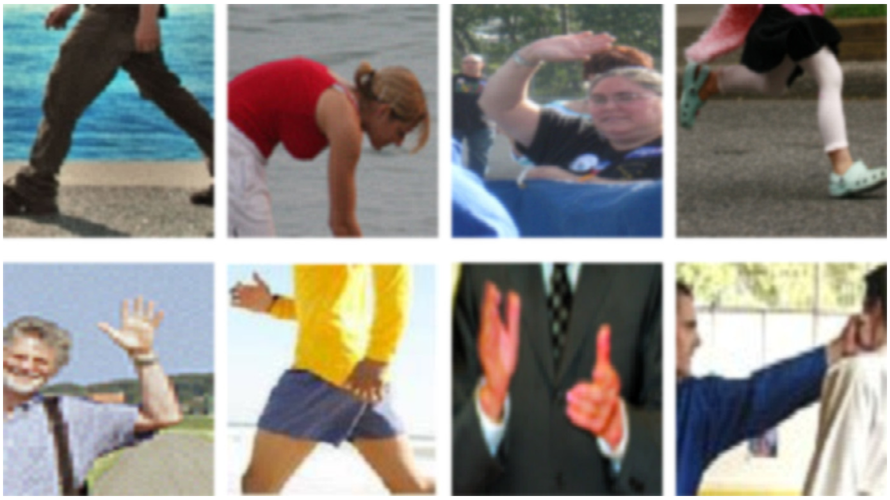

Fig. 1. Examples of image patches in which we can recognize human activities even though a view of the whole body is not available.

specific regions of interest in still images. Based on the observation that activities are temporal phenomena, we make use of information that is available from video analysis. Fig. 2, shows a diagram of the components of our approach towards determining action-specific regions of interest regions and subsequent image classification.

Given videos that show human activities, we compute optical flow fields and consider the magnitudes of flow vectors in each frame of a video. Given a collection of frames of flow magnitudes, we then 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 part 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 interest points or regions for subsequent feature extraction. For action classification in still images, we devise a generative probabilistic model that characterizes joint distributions of regions of interest, local image features (visual words) and human actions.

To evaluate the usefulness of regions of interest contained in basis flows, 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 an independent set 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. Below, we show that, even in the absence of any annotation of joints or body parts, our generative model achieves a high accuracy in action classification.

The major contributions of this article are the following: (i) we propose a novel scheme for extracting discriminative spatial regions for action recognition in still images using simple videos; (ii) we apply NMF to determine action specific regions of interest from motion flows; (iii) we incorporate action saliency maps based on videos and local spatial features of action images in a Bayesian framework for human action classification.

Our presentation proceeds as follows. In Section 2, we review related work on human action recognition. Section 3 describes our method of learning action specific interest regions from videos. In Section 4, we present a generative model for action classification. Section 5 provides an experimental evaluation with respect to the location of human body parts. Finally, Section 6 summarizes our work and results.

2. Related work

As an exhaustive review of work on visual activity recognition is beyond the scope of this article, we restrict our discussion to the

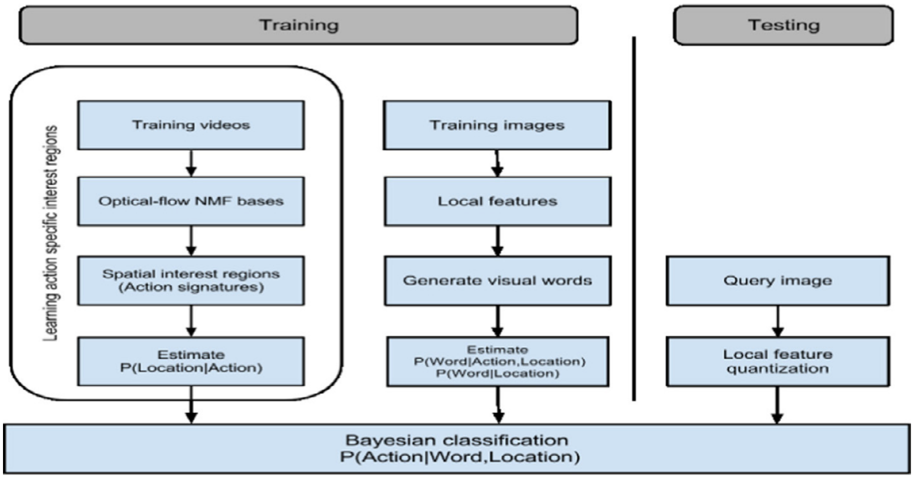
Training
Testing
Learning action specific interest regions
Training videos
Optical-flow NMF bases
Spatial interest regions (Action signatures)
Estimate P(Location|Action)
Training images
Local features
Generate visual words
Estimate P(Word|Action,Location) P(Word|Location)
Query image
Local feature quantization
Bayesian classification P(Action|Word,Location)

Fig. 2. General diagram of our approach. In the training phase, we learn the actions' priors P(Location | Action) from training videos, and codebook prior P(Word | Action, Location) from training images in order to perform human action recognition for new test queries in a fully Bayesian framework.

two arguably most popular approaches in the recent literature. In addition, we briefly review related matrix factorization methods. Approaches that rely on the idea of bags of visual words (BoW) are popular for their simplicity, robustness, and good performance in content-based image or video classification. Corresponding work treats an image or a video as a collection of independent visual descriptors computed at certain 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.

BoW approaches such as Matikainen et al., Laptev et al. and Cheema et al. [29,23,3] based on key points detection [2,17,22,10,18,30], though generally discriminative, do not regard task specific objectives in key point localization. Rather, key points are determined from low-level properties of the image or video-signal. Moreover, corresponding approaches typically assume key points to be independent and therefore fail to explain characteristics spatial and temporal layouts. Liu et al. [45] address this limitation and propose extracting reliable and informative features from the unconstrained videos by mining image low-level features. Alternatively, Kovashka and Grauman [21] suggest learning mid-level representations that encodes spatial and temporal relationships among key points. Gilbert et al. and Liu et al. [16,25] employ data mining to build high-level compound features from noisy and over-complete sets of low-level spatio-temporal features, Song et al. [32] use a triangular lattice of grouped point features to encode spatial layouts. Coates and Ng, Malinowski and Fritz, Sharma et al., [8,28,31] propose weighting local features while pooling in a way that regards the classification task in hand. Still, these approaches, too, center around low-level signal properties which do not necessarily provide an accurate account of the characteristics of an activity and are often time and resource expensive.

Some recent approaches proposed human-based fixation for sampling features [5–7]. Mathe and Sminchisescu [6] propose a saliency map learned from eye movements, Vig et al. [5] propose a saliency-based descriptor for action recognition. These approaches indicate that using saliency maps learned from human fixation locations enhance the performance in comparison to other sampling techniques while using an order of magnitude less descriptors. As opposed to the previous work, our approach automatically learns the saliency maps from training videos (without human intervention) by analyzing their motion fields using NMF.

Sampling techniques such as random sampling have shown state-of-the-art performance, too. Nowak et al. [12] 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. Wang et al. [37] state that dense sampling outperform all point detectors in realistic scenarios and, correspondingly, Wang et al. [36] utilize motion trajectories to compute space–time features. However, at the same time, recent work in Gall et al. [15] 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 human action recognition and were indeed shown to successfully cope with the PASCAL visual object recognition challenge. Felzenszwalb et al. [14] describe a deformable model for human detection which was used to achieve state-of-the-art performance in action recognition on benchmark data sets [9]. The work in Bourdev and Malik [4] introduces exemplar-based pose representation, or *poselets*, for human detection. This term denotes a set of patches with similar pose configurations. The work in Maji et al. [27] utilizes poselets to identify human poses as well as actions in still images while [33] propose an articulated part-based model for human pose estimation and detection which adapts a hierarchical (coarse-to-fine poselet-like) representation. Yang et al. [40] exploits poselets as a coarse representation of the human pose and treats them as latent variables for action recognition. Despite their recent success, it is still questionable if these methods can make use of the favorable statistics of present day large scale data sets because the construction of suitable poselets requires extensive human intervention and manual labeling in the training phase.

Thurau and Hlavac, Eweiwi et al. [34,39] consider non-negative matrix factorization (NMF) for action recognition and apply it to learn a set of pose and background bases. In Agarwal and Triggs [1], the authors estimate the human upper body pose through NMF. The work in Yao et al. [41] empirically evaluates the problem of human action recognition using pose or appearance-based features. The authors 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 allow for integrating both approaches into a single framework. Next, we discuss how the approach proposed in this article indeed provides a method for the informed sampling of key points for appearance-based action recognition as well as an approach towards learning descriptors of body poses.

3. Learning action-specific interest regions from videos

Our approach identifies discriminative regions in the image plane and subsequently learns the relative importance of these 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 in order to derive a generative probabilistic model that features joint distributions of regions of interest and human actions.

3.1. Learning basis flows using NMF

Given a set of videos of different actions, we determine optical flow magnitudes at each pixel within a bounding bounding box of constant size surrounding a person visible in the video. Each frame can thus be transformed into a d dimensional non-negative vector $\mathbf{u}$. Let n_i represent the number of frames for an action $a_i \in \mathcal{A} = \{a_1, a_2, \ldots, a_r\}$ and let $N = \sum_{i=1}^{r} n_i$. We build a data matrix $\mathbf{U}$ of dimension $d \times N$ containing the flow magnitude vectors of all frames. Computing NMF yields K basis vectors, or *basis flows*, such that $\mathbf{U} \approx \mathbf{WH}$ where the columns of $\mathbf{W}_{d\times K}$ 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 gradient descent algorithm according to Lee and Seung [24]. 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 Donoho and Stodden, Klingenberg et al., Thurau et al. [11,20,35]. Accordingly, we can expect the resulting basis flows to be sparse in the sense that most entries of a basis element $\mathbf{w}_k$ will be (close to) zero and only a few entries will have noticeable values. Fig. 3(h) shows that this is indeed the case. It depicts pictorial representations of exemplary basis vectors $\mathbf{w}_k$ resulting from our NMF step. Note that for each basis element only

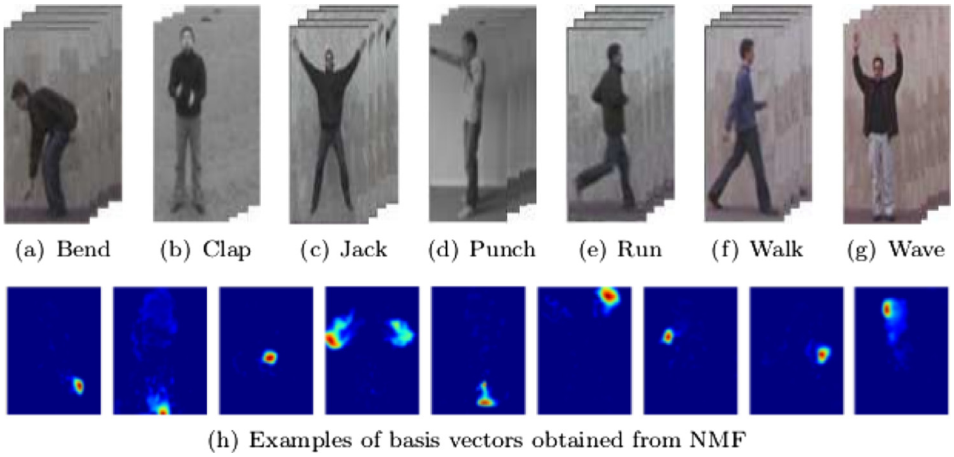

Fig. 3. (a–g) Examples of training videos from the Weizmann and KTH data sets; (h) examples of basis flows obtained from applying NMF on optical flow fields.

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 **H** since its entries encode linear mixing coefficients required to reconstruct the vectors in **U** from the basis flows in **W**. Consequently, the columns of **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 $\mathbf{w}_k$ given an action a_i is determined as:

$$P(\mathbf{w}_k|a_i) = \frac{\sum_{f\in a_i} h_{kf}}{\sum_{j=1}^{K}\sum_{f\in a_i} h_{jf}} \tag{1}$$

Note in Fig. 4(a) that the resulting probability distribution, i.e. the set of weights of a basis element w.r.t. an action, again is sparse. Therefore, the distribution in Eq. (1) immediately allows us to determine how characteristic a certain basis flow is for an activity.

3.3. Action signatures and salient regions

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

$$\mathbf{s}_i = \sum_{k=1}^{K} P(\mathbf{w}_k|a_i)\mathbf{w}_k. \tag{2}$$

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. Fig. 4(b–h) shows examples of action signatures which we obtained from basis flows extracted from the Weizmann and KTH 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. Regarding the former, each action signature s_i, i.e. a d-dimensional vector in the image space, can be used to derive an action-specific spatial saliency for an image region l_k, namely

$$P(l_k|a_i) = \sum_{j\in l_k} s_i. \tag{3}$$

4. Action classification in still images using spatial interest regions

In this section, we describe a Bayesian framework for action classification that combines bags of visual words approach used for still images with action signatures learned from videos. For a given set of training images $\mathbf{F} = \{(\mathbf{f}_i, y_i),\ i = 1, 2, \dots M\}$ where $y_i \in \mathcal{A}$, each image is first divided into a set L of cells (or locations) and local histogram of oriented gradient are extracted for each of the locations. A vocabulary of visual words $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$ is then learned using k-means clustering based on the L_2 norm. Thus, each training image is represented as vector of $|L|$ visual words.

Our classification approach considers the likelihood of an action given spatial locations (with their relative importance) and visual words for those locations. This likelihood $P(a_i|\mathbf{v}_j, l_k)$ is estimated as

$$P(a_i|\mathbf{v}_j, l_k) = \frac{P(\mathbf{v}_j|a_i, l_k)P(a_i|l_k)}{P(\mathbf{v}_j|l_k)} \tag{4}$$

$$= \frac{P(\mathbf{v}_j|a_i, l_k)P(l_k|a_i)P(a_i)}{P(\mathbf{v}_j|l_k)P(l_k)} \tag{5}$$

$$= \alpha\frac{P(\mathbf{v}_j|a_i, l_k)P(l_k|a_i)}{P(\mathbf{v}_j|l_k)} \tag{6}$$

where α is the normalization factor and priors $P(a_i)$ and $P(l_k)$ are assumed to be uniformly distributed. The three conditional probabilities on the right hand side of Eq. (6) are estimated from the training data (**U** and **F**).

$P(l_k|a_i)$ is the action-specific importance of each region and is derived from action signatures that were learned from action videos (see Eq. (3)).

The term $P(\mathbf{v}_j|l_k)$ denotes the likelihood of visual word v_j given location l_k and is determined by

$$P(\mathbf{v}_j|l_k) = \frac{\sum_{\mathbf{f}\in\mathbf{F}}\chi(\mathbf{v}_j, l_k)}{|\mathbf{F}|} \tag{7}$$

where χ is an indicator function that has value 1 only if v_j is assigned to l_k.

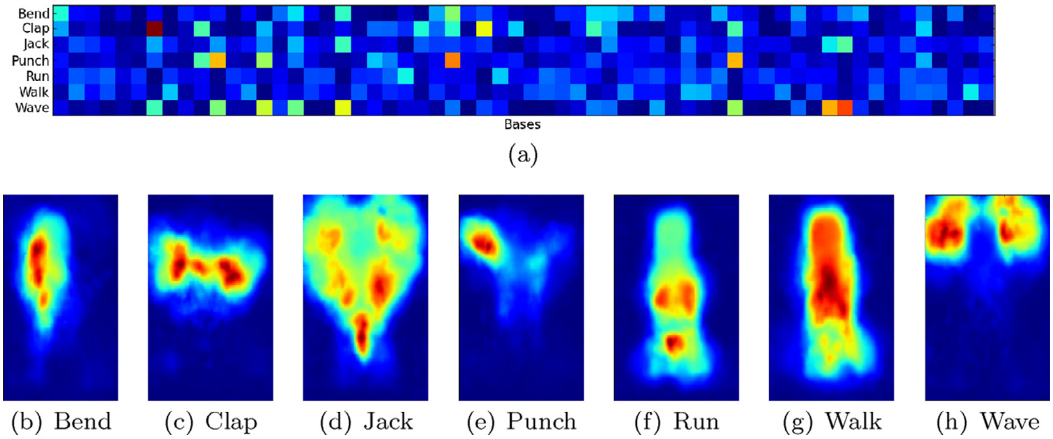

Fig. 4. (a) Relative importance of bases with respect to different actions as characterized by $P(\mathbf{w}_k|a_i)$. Note that actions flows can be approximated by only a small number of basis vectors. (b–h) Examples of action signatures resulting from Eq. (2).

While the quantity $P(\mathbf{v}_j|l_k)$ indicates the overall probability of the occurrence of a visual word at a specific location, the action-specific likelihood $P(\mathbf{v}_j|a_i, l_k)$ further specifies the relevant importance of visual words at different locations for different actions. This is achieved by computing

$$P(\mathbf{v}_j|a_i, l_k) = \frac{\sum_{\mathbf{f}\in\mathbf{F}^{(a_j)}} \chi(\mathbf{v}_j, l_k)}{|\mathbf{F}^{(a_j)}|\,|L|} \tag{8}$$

where $\mathbf{F}^{(a_j)} \subset \mathbf{F}$ are those training images that contain examples of action a_i.

5. Experimental results

Our experimental evaluation of the proposed approach mainly addresses two tasks: (i) the matching of video-based action-specific regions of interests to important body parts in still images (Section 5.1) and (ii) the classification of action images using regions of interest or signatures (Section 5.2).

In order to learn action specific regions of interest, we used videos of different actions available in the Weizmann and KTH data sets. As these videos show little change in background and view-point, they allow us to focus on estimating the importance of different body parts for different actions. In particular, we consider the following actions *Bending*, *Clapping*, *Jacking*, *Punching*, *Running*, *Walking*, and *Waving*. We used the bounding boxes provided by Yao et al. [42] and resized them to common size of 96 × 64 pixels. To determine optical flows, we considered the methods due to Lucas–Kanade [26] and Farnebäck [13]. In both cases, we used the corresponding OpenCV implementations, however, similar to [36], we finally adopted the Farnebäck algorithm as we observed a higher efficiency and a more robust performance in the extraction of our actions signatures. Finally, all of the results reported below were obtained using 200 basis flows $\mathbf{w}_i$.

In order to evaluate the proposed approach on the target domain, i.e. still images, we used 270 images from the H3D [4] and the VOC2011 [46] datasets which we also resized to a resolution of 96 × 64 pixels. Each of these images shows a person performing an action. Fig. 5 shows several examples for each action. Note that most of the images involve background clutter and occlusion.

5.1. Interest regions and salient body parts

We first evaluate as to how far regions of interest extracted by our approach described in Section 3 correspond to locations of human body parts in real images. To this end, we make use of the manually annotated positions of limbs or joints that are available in the H3D and VOC2011 data sets. In particular, we determine the joint probability distribution of actions, interest regions, and body parts. Given the locations of a body part b_j in an image of action a_i, we have

$$\begin{aligned} P(b_j, \mathbf{w}_k, a_i) &= P(b_j|\mathbf{w}_k, a_i)P(\mathbf{w}_k|a_i)P(a_i) \\ &= P(b_j|\mathbf{w}_k)P(b_j|a_i)P(\mathbf{w}_k|a_i)P(a_i) \end{aligned} \tag{9}$$

where $P(b_j|\mathbf{w}_k)$ is chosen to be inversely proportional to the Euclidean distance between the location of b_j and the center of a region in $\mathbf{w}_k$. The prior $P(a_i)$ is assumed to be uniform. The conditional distribution $P(b_j|a_i)$ is obtained by marginalizing over the K bases and all training images corresponding to action a_i. Consequently, $P(b_j|a_i)$ can be understood to encode the relative importance of different body parts for different action a_i.

We used all 270 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 compare our interest regions to key points extracted by two popular detectors, the Harris detector [17] and the SIFT key point detector [10]. 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 mimicking the conditional distribution $P(b_j|a_i)$ discussed in Section 3.

Fig. 6 compares results from our method for extracting interesting regions from video data to the ones obtained from using Harris and SIFT key points. The visualization emphasizes the relative importance of the body parts for a particular action given different sampling schemes. The size of the plotted body part corresponds to the sampling rate or the importance of locations around that part. We observe that, in case of Harris and SIFT key points, head and feet dominate other limbs regardless of the action (Fig. 6 rows 1 and 2). Moreover, in these cases, the probabilities for other body parts are almost uniformly distributed and do not convincingly relate to different actions. For example, body parts naturally related to the activity of *Clapping*, i.e. elbows and hands, achieve rather low scores compared to other limbs or parts.

On the other hand, our approach exhibits logically coherent relationships between body parts and actions (Fig. 6 third row). Compare, for instance, the varying importance of different body parts for *clapping* and *running*. Clearly, the lower body parts are

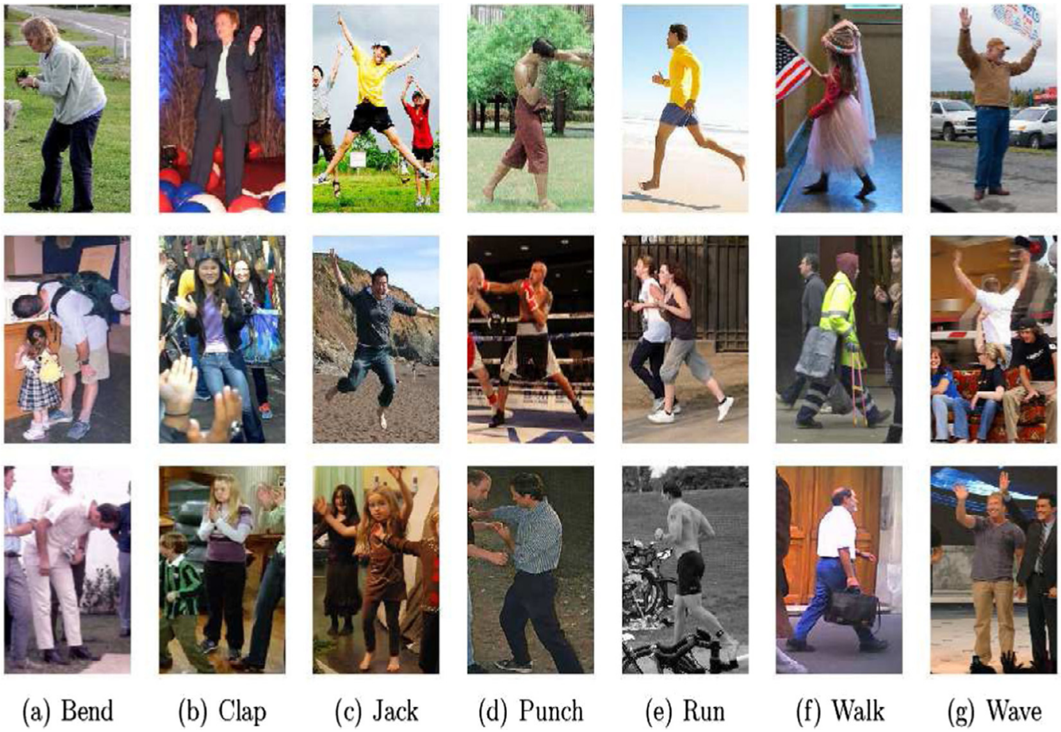

Fig. 5. Examples of images showing different actions.

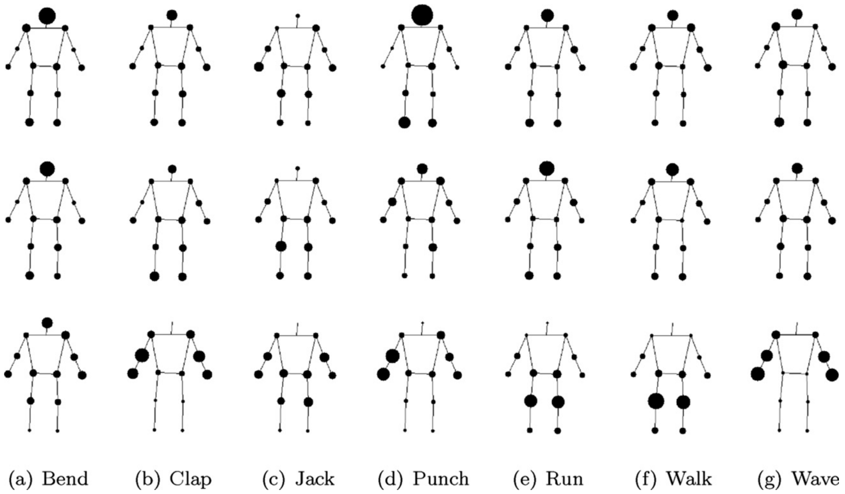

Fig. 6. Stick figures depicting the relevance of different body parts for different actions. Important key points computed using the Harris detector (first row) and SIFT detector (second row) hardly correlate to action-specific body parts; interest regions from our approach correlate better (third row).

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 *Clapping* or *running* when compared to *bending*. We expect that this favorable property of the approach proposed in this article can ultimately be used in order to establish rigorous and discriminative action models through an informed sampling phase that concentrates on distinctive patterns of an action rather than on random samples.

5.2. Action classification

Having established the effectiveness of our approach to identify relevant regions (body parts) for different actions, we next evaluate its utility for action recognition in still images. To this end, we consider all images in our dataset without any annotation of body parts or joints. For training, we again divide each image into L rectangular regions (see Section 4), where we use a 16 × 16 grid of overlapping cells. Within each cell we extract an L_2 normalized 6-bin local histogram of oriented gradients. To compute a vocabulary $\mathbf{V}$ of visual words, we considered different numbers of words and observed good performance for 64–120 words. Next, we determined the probability distributions $P(\text{word}|\text{location})$ and $P(\text{word}|\text{action,location})$ (see Eq. (6)). Note again that $P(\text{location}|\text{action})$ is determined by action signatures learned from videos.

Given a query image, we identify the best matching visual word $\mathbf{v}(l_k)$ at each location and assign the query image the action with the highest likelihood

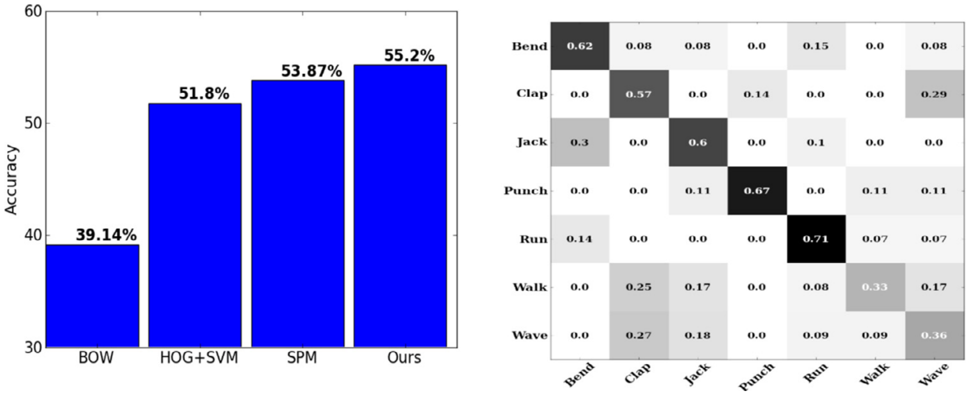

Fig. 7. (left) Classification accuracies using various approaches. (right) Confusion matrix for action classification by the proposed approach.

$$\text{argmax}_{a_i} \sum_{k=1}^{L} P(a_i|\mathbf{v}(l_k), l_k) \tag{10}$$

$$= \text{argmax}_{a_i} \sum_{k=1}^{L} \frac{P(\mathbf{v}(l_k)|a_i, l_k)P(l_k|a_i)}{P(\mathbf{v}(l_k)|l_k)} \tag{11}$$

We used 5-fold cross validation and achieved an accuracy of about 55.2% accuracy for action recognition. We compared our approach to the standard BoW approach based on Spatial Pyramid (SPM) binning and to other global template matching technique using the histograms of oriented gradients (HOG). For SPM, we densely sample local features every 6 pixels at multiple scales and compute SIFT features [10]. Then we construct a codebook using K-means and use the codebook to encode the extracted local features from the image. The codes are pooled afterwards over three levels of the spatial pyramid of the image plane [44]. Fig. 7 compares the performance to the baseline methods and show the confusion matrix obtained by approach. Note that most ambiguity is due to the actions of *waving*, *jacking*, and *clapping*, as all of them share similar body parts configuration. On the other hand, actions such as *punching* and *bending* are accurately classified by our approach.

Finally, we evaluated the impact of the number of training videos on our saliency map $P(\text{location}|\text{action})$. In the extreme case, when training videos are not present, our model assigns a uniform distribution of location importance to $P(\text{location}|\text{action})$. As discussed earlier, this is similar to orderless BoW model and as anticipated the performance is close to it (42.71%). By using only two training videos per action, our model achieves an accuracy as high as HOG+SVM (51.02%). The best performance of 55.2% was obtained by considering only 4 videos per action. Adding more training videos did not significantly improve the overall classification. To conclude, it appears that only a few videos are sufficient for constructing discriminative action signatures that covers all different execution styles of an action.

6. Conclusion and future work

We have presented a novel approach to the human action recognition in still images based on the notion of regions of interest. Since human activities are inherently dynamic, we proposed to analyze optical flow fields extracted from video sequences showing human activities in order to learn about salient regions for action recognition. 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. We applied this information in a generative Bayesian model for action classification which integrates information as to regions of interests and local spatial image features. In an experimental validation, we found a clear relationship between regions of interest determined by our approach and action-specific body parts. We also found that the proposed approach yields higher action recognition accuracies than three recent baseline methods. This is noteworthy since our approach fundamentally differs from existing approaches for action recognition in still images. Firstly, although we consider rather low-level signal 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 subsequently facilitates an informed sampling of key points in still image. Secondly, 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 better 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 scenes or images of partly occluded human bodies. This article therefore established a baseline for video-based feature selection and classification towards action recognition in still images.

References

[1] A. Agarwal, B. Triggs, A local basis representation for estimating human pose from cluttered images, in: ACCV, 2006.

[2] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, SURF: speeded up robust features, CVIU 110 (2008) 346–359.

[3] S. Cheema, A. Eweiwi, C. Bauckhage, Human activity recognition by separating style and content, in: Pattern Recognition Letters, vol. 34, 2013.

[4] L. Bourdev, J. Malik,Poselets: body part detectors trained using 3d human pose annotations, in: ICCV, 2009.

[5] E. Vig, M. Dorr, D. Cox, Space-variant descriptor sampling for action recognition based on saliency and eye movements, in: ECCV, 2012.

[6] S. Mathe, C. Sminchisescu, Dynamic eye movement datasets and learnt saliency models for visual action recognition, in: ECCV, 2012.

[7] L. Itti, C. Koch, A saliency-based search mechanism for overt and covert shifts of visual attention, Vis. Res. 40 (2000) 1489–1506.

[8] A. Coates, A. Ng, The importance of encoding versus training with sparse coding and vector quantization, in: ICML, 2011.

[9] V. Deltaire, I. Laptev, J. Sivic, Recognizing human actions in still images: a study of bag-of-features and part-based representations, in: BMVC, 2010.

[10] D.G. Lowe, Distinctive image features from scale-invariant keypoints, IJCV 60 (2004) 91–110.

[11] D. Donoho, V. Stodden, When does non-negative matrix factorization give a correct decomposition into parts?, in: NIPS, 2004.

[12] E. Nowak, F. Jurie, B. Triggs, Sampling strategies for bag-of-features image classification, in: ECCV, 2006.

[13] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: SCIA, 2003.

[14] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, TPAMI 32 (2010) 1627–1645.

[15] J. Gall, A. Yao, N. Razavi, L. Van Gool, V.S. Lempitsky, Hough forests for object detection, tracking, and action recognition, TPAMI 33 (2011) 2188–2202.

[16] A. Gilbert, J. Illingworth, R. Bowden, Action recognition using mined hierarchical compound features, TPAMI 33 (2011) 883–897.

[17] C. Harris, M. Stephens, A combined corner and edge detection, in: In Alvey Vision Conference, 1988.
[18] H. Jhuang, T. Serre, L. Wolf, T. Poggio, A biologically inspired system for action recognition, in: ICCV, 2007.
[19] S. Johnson, M. Everingham, Learning effective human pose estimation from inaccurate annotation, in: CVPR, 2011.
[20] B. Klingenberg, J. Curry, A. Dougherty, Non-negative matrix factorization: ill-posedness and a geometric algorithm, PR 42 (2008) 918–928.
[21] A. Kovashka, K. Grauman, Learning a hierarchy of discriminative space–time neighborhood features for human action recognition, in: CVPR, 2010.
[22] I. Laptev, On space–time interest points, IJCV 64 (2005) 107–123.
[23] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: CVPR, 2008.
[24] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–799.
[25] J. Liu, Y. Yang, I. Saleemi, M. Shah, Learning semantic features for action recognition via diffusion maps, CVIU 116 (2012) 361–377.
[26] L. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Imaging Understanding Workshop, 1981.
[27] S. Maji, L. Bourdev, J. Malik, Action recognition from a distributed representation of pose and appearance, in: CVPR, 2011b.
[28] M. Malinowski, M. Fritz, Learning smooth pooling regions for visual recognition, in: BMVC, 2013.
[29] P. Matikainen, M. Hebert, R. Sukthankar, Trajectons: action recognition through the motion analysis of tracked features, in: ICCV Workshop on Video-Oriented Object and Event Classification, 2009.
[30] C. Schmid, R. Mohr, C. Bauckhage, Evaluation of interest point detectors, IJCV 37 (2000) 151–172.
[31] G. Sharma, F. Jurie, C. Schmid, Discriminative spatial saliency for image classification, in: CVPR, 2012.
[32] Y. Song, L. Goncalves, P. Perona, Unsupervised learning of human motion, TPAMI 25 (2003) 814–827.
[33] M. Sun, S. Savarese, Articulated part-based model for joint object detection and pose estimation, in: ICCV, 2011.
[34] C. Thurau, V. Hlavac, Pose primitive based human action recognition in videos or still images, in: CVPR, 2008.
[35] C. Thurau, K. Kersting, M. Wahabzada, C. Bauckhage, Convex non-negative matrix factorization for massive datasets, KAIS 29 (2011) 457–478.
[36] H. Wang, A. Klaeser, C. Schmid, L. Cheng-Lin, Action recognition by dense trajectories, in: CVPR, 2011.
[37] H. Wang, M.M. Ullah, A. Klaeser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: BMVC, 2009.
[38] D. Weinland, R. Ronfard, E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition, CVIU 115 (2011) 224–241.
[39] A. Eweiwi, M. Cheema, C. Bauckhage, Discriminative joint non-negative matrix factorization for human action classification, in: GCPR, 2013.
[40] W.Yang, Y. Wang, G. Mori, Recognizing human actions from still images with latent poses, in: CVPR, 2010.
[41] A. Yao, J. Gall, G. Fanelli, L. Van Gool, Does human action recognition benefit from pose estimation?, in: BMVC, 2011a.
[42] A. Yao, J. Gall, L. Van Gool, A hough transform-based voting framework for action recognition, in: CVPR, 2010.
[43] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L. Guibas, L. Fei-Fei, Human action recognition by learning bases of action attributes and parts, in: ICCV, 2011b.
[44] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: CVPR, 2006.
[45] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the wild, in: CVPR, 2009.
[46] M. Everingham, L. Van Gool, C. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results, 2011. <http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html> .