# Computer Vision Lab SS16
# P-CNN: Pose-based CNN Features for Action Recognition

Ahmed Abdelbaki
University of Bonn
Bonn, Germany
abdelbak@iai.uni-bonn.de

## Abstract

*This work targets human action recognition in video based on the state of art deep convolutional neural network (CNN). Recent methods represents actions by statistics of local video features,here we argue for the importance of a representation derived from human pose. We are using an approach called Pose-based Convolutional Neural Network descriptor (P-CNN) for action recognition. The descriptor aggregates motion and appearance information through patches of human body parts. We investigate different CNN model fro the P-CNN approach and experiment P-CNN features obtained both for automatically estimated and manually annotated human poses. We evaluate our method on the recent and challenging JHMDB dataset. Our evaluation shows consistent improvement over the original approach.*

## 1. Introduction

Recognition of human actions is an important step toward fully automatic understanding of dynamic scenes. Despite significant progress in recent years, action recognition remains a difficult challenge. Common problems stem from the strong variations of people and scenes in motion and appearance. Other factors include subtle differences of fine-grained actions, for example when manipulating small objects or assessing the quality of sports actions.

The majority of recent methods recognize actions based on statistical representations of local motion descriptors [14, 10, 17]. These approaches are very successful in recognizing coarse action (standing up, hand-shaking, dancing) in challenging scenes with camera motions, occlusions, multiple people, etc. Global approaches, however, are lacking structure and may not be optimal to recognize subtle variations, e.g. to distinguish correct and incorrect golf swings or to recognize fine-grained cooking actions

In this work, we are using the P-CNN approach which is a new action descriptor based on human poses. Provided with tracks of body joints over time,it combines motion and appearance features for body parts. In [4], they explore CNN features obtained separately for each body part in each frame due to the recent success of CNN [8, 11]. We use appearance and motion-based CNN features computed for each track of body parts and investigate different patch extraction mechanisms which are based on human joints pose. Moreover, we examine different CNN architecture for the feature extraction by comparing between the "VGG-F" Net [2] and Microsoft Residual Networks(ResNet) [6]. The CNN features extraction pipeline of (P-CNN) is illustrated in figure 1.

The rest of the paper/technical report is organized as follows. Related work is discussed in Section2. Section3 introduces the P-CNN approach and CNN features extraction. Our contribution of the improved P-CNN is explained in Section4. We present the Dataset used in our experiments in Section6 and summarize the experimental results in Section7.

## 2. Related work

Action Recognition in the last decade has been dominated by local and hand crafted features. In particular, Dense Trajectories (DT) features[17] combined with Fisher Vector (FV)[13] aggregation have shown magnificent results for many challenging benchmarks.

The latest breakthroughs in Convolutional Neural Networks (CNN) have impacted a significant progress in image classification and other vision tasks.Although the applications of CNNs to action recognition in video[15, 18] has shown only limited improvements so far. P-CNN approach extend previous global CNN methods and address action recognition using CNN descriptors at the local level of human body parts.

Most of the recent methods for action recognition deploy global aggregation of local video descriptors. Such representations provide invariance to numerous variations

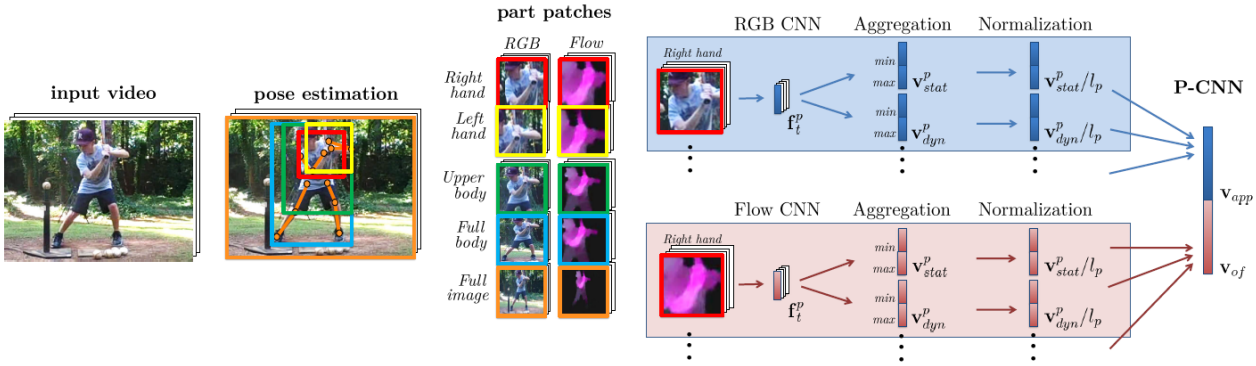


Figure 1. P-CNN features. From left to right: Input video. Human pose and corresponding human body parts for one frame of the video. Patches of appearance (RGB) and optical flow for human body parts. One RGB and one flow CNN descriptor $f_t^p$ is extracted per frame *t* and per part *p* (an example is shown for the human body part *right hand*). Static frame descriptors $f_t^p$ are aggregated over time using *min* and *max* to obtain the video descriptor $v_{stat}^p$. Similarly, temporal differences of $f_t^p$ are aggregated to $v_{dyn}^p$. Video descriptors are normalized and concatenated over parts *p* and aggregation schemes into appearance features $v_{app}$ and flow features $v_{of}$. The final P-CNN feature is the concatenation of $v_{app}$ and $v_{of}$.

in the video but may fail to capture important spatio-temporal structure. For fine-grained action recognition, previous methods represented person-object interactions by joint tracking of hands and objects[12] or, by linking object proposals[19]. Alternative methods represent action using positions and temporal difference of body joints. Hence, it seems that Pose Estimation is a very important task for action recognition;however, reliable human pose estimation is still a challenging task. A recent study [7] reports significant gains provided by dynamic human pose features in case of the availability of reliable pose estimation. P-CNN extend the work [7] and designed a new CNN-based representation for human actions combing positions, appearance and motion of human joints.

## 3. P-CNN: Pose-based CNN features

In [4], they claim that human pose is essential for action recognition. They use positions of body joints to define informative image regions. Moreover, such regions are represented with motion-based and appearance-based CNN descriptors. Such descriptors are extracted at each frame and then aggregated over time to form a video descriptor, see figure 1 for an overview.
The details are explained below.

To generate P-CNN features, optical flow is computed for each consecutive pair of frames using [1] method. They borrow inspiration from [5], the values of the motion field $v_x, v_y$ are transformed to the interval [0,255] by $\tilde{v}_{x|y} = av_{x|y} + b$ where $a = 16$ and $b = 128$. The values below 0 and above 255 are truncated. The transformed flow maps are saved as images with three channels corresponding to motion $\tilde{v}_{x|y}$ and the flow magnitude. Given a video frame and the corresponding positions of body joints, the RGB patches and flow patches are cropped for the following body parts: *right hand, left hand, upper body, full body and full image* as illustrated in Figure1. Each patch is resized to 224 x 224 pixels to match the CNN input layer. To represent appearance and motion patches, Two distinct CNNs are used with architecture similar to AlexNet[8]. Both networks contain 5 convolutional and 3 fully-connected layers. The output of the second fully-connected layer with $k = 4096$ values is used as a frame descriptor ($f_t^p$). For RGB patches,they use the publicly available VGG-f network from [2] that has been pre-trained on the ImageNet ILSVRC-2012 challenge dataset. Regarding the flow patches, the motion network provided by [5] are being used which was pre-trained for action recognition task on the you tube video UCF101 dataset.
Given descriptors $f_t^p$ for each part *p* and each frame *i* of video, the proceeding step is to aggregate $f_t^p$ over all frames to obtain a fixed-length video descriptor. They consider the *min* and *max* aggregation by computing the minimum and maximum values for each descriptor *i* over *T* video frames

$$\begin{aligned} m_i &= \min_{1 \le t \le T} f_t^p(i), \\ M_i &= \max_{1 \le t \le T} f_t^p(i) \end{aligned} \tag{1}$$

The static video descriptor for part p is defined by the concatenation of time-aggregated frame descriptors as

$$v_{stat}^p = [m_1, ..., m_k, M_1, ..., M_k]^T. \tag{2}$$

To capture temporal evolution of per-frame descriptors, the temporal differences will be in form $\Delta f_t^p = f_{t+\Delta t}^p - f_t^p$ for $\Delta t = 4$ frames. Similar to (1), the aggregation of

temporal differences values is computed for minimum $\Delta m_i$ and maximum $\Delta m_i$ of $\Delta f_t^p$ and concatenate them into the dynamic video descriptor

$$v_{stat}^p = [\Delta m_1, ..., \Delta m_k, \Delta M_1, ..., \Delta M_k]^T. \quad (3)$$

Finally, video descriptors for motion and appearance for all parts and different aggregation schemes are normalized and concatenated into the P-CNN feature vector. The normalization is performed by dividing video descriptors by the average $L_{2-norm}$ of the $f_t^p$ from the training set.The concatenation of static and dynamic descriptors will be denoted by Static+Dyn.
The final dimension of our P-CNN is $(5x4x4K)x2 = 160K$, i.e., 5 body parts, 4 different aggregation schemes, 4K-dimensional CNN descriptor for appearance and motion. P-CNN training is performed using a linear SVM.

## 4. IPCNN: Improved P-CNN

In the previous section, we discussed the details of the P-CNN approach which is in brief extracting the video descriptors $V_{Static+Dyn}^p$ for each body part p aggregated across all the frames. The frame descriptor $f_t^p$ is computed from the appearance-based and motion-based CNNs pre-trained models. In this section, we will discuss the improved version of P-CNN which is replacing the "VGG-F" apperance CNN with the state of art image classification very Deep CNN called Deep Network(ResNet)[6].

According to ImageNet 2015, ResNet out performs all the other approaches like GoogleNet[16] and Vgg-f as illustrated in figure 2.An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. This result won the 1st place on the ILSVRC 2015 classification task. On the ImageNet dataset they evaluated residual nets with a depth of up to 152 layers-8x deeper than VGG nets but still having lower complexity. Hence, it is proven that deeper net is better performance in case there is no degradation problem and ResNet is a stack of residual blocks with very deep architecture that solves the vanishing gradients problem, see figure 3.

In the context of action recognition and P-CNN method, we apply the same approach on JHMDB dataset6 by replacing "VGG-F" appearance-based model with ResNet-50 (50 layers Residual Blocks). In contrast with 3, the output of the last average pooling layer with k=2048 values is used as appearance frame descriptor $f_t^p$. This improved version of P-CNN (IP-CNN) out performs the original approach by 3 4%; for more details read section 7.

## 5. P-CNN without Pose Estimation

In this section, we want to prove that the claim in [4] regarding the importance of the human pose in the action recognition task is correct. Removing the pose information from the patch extraction process and use another naive approach will prove by contradiction how crucial the human joints positions to the task.

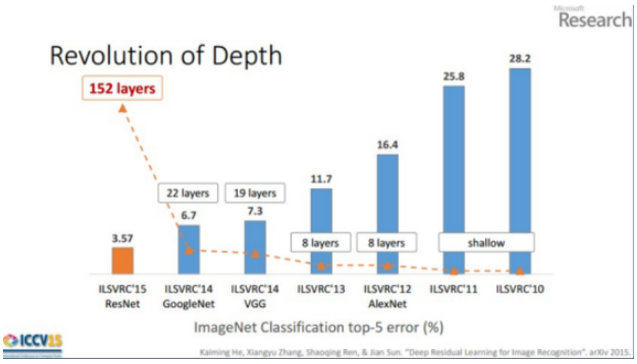


Figure 2. Performance of different approaches in ImageNet 2015 competition.

| | JHMDB | | |
|---|---|---|---|
| | GT | Pose[3] | diff |
| P-CNN | 74.6 | 66.8 | 13.5 |
| HLPF | 77.8 | 25.3 | 52.5 |

Table 1. Impact of automatic pose estimation versus ground-truth pose (GT) for P-CNN features and HLPF[7].Results are presented for JHMDB (% accuracy )

In [4], they compared P-CNN using ground truth pose to the automatic pose estimation; the summary of the results is shown in Table 1. On the other hand, we tried different approaches for parts extraction without using the pose information.The first approach named "NoPose" is to crop an image into quarters and a center one regardless the human bounding box and the second approach "NoPose-BB" is similar to the previous one but extract the crops within the human bounding box;see figure 4. The results have proven the claim of the importance of human pose in our task and this will be discussed in Section 7.

## 6. Datasets

In our experiments, we use a well-known dataset called JHMDB (Joint-annotated Human Motion Data Base)[7]. We present it in the following.

**JHMDB** is a subset of **HMDB**[9], see figure 5. It contains 21 human actions, such as *brush hair, climb, golf, run or sit*. Video clips are restricted to the duration of the action. There are between 36 and 55 clips per action for a total of 928 clips. Each clip contains between 15 and 40 frames of size 320 X 240. Human pose is annotated in each of the 31838 frames. There are 3 train/test splits for the JHMDB dataset and evaluation averages the results over these three

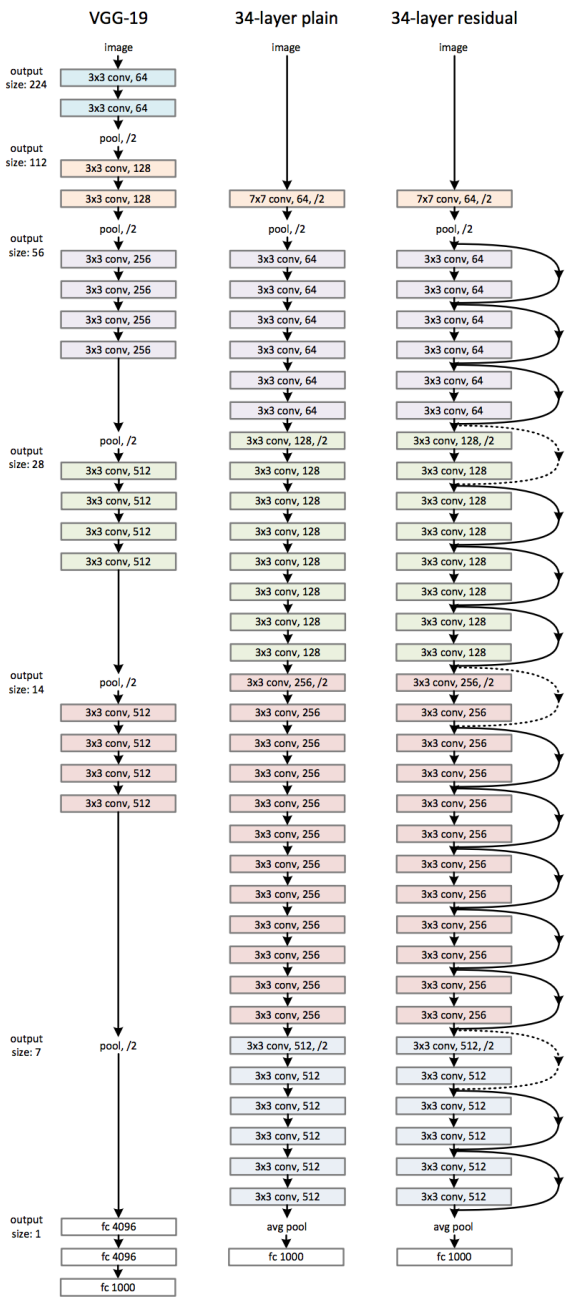


Figure 3. Example network architectures for ImageNet. Left: the VGG-19 model (19.6 billion FLOPs) as a reference. Middle: a plain network with 34 parameter layers (3.6 billion FLOPs). Right: a residual network with 34 parameter layers (3.6 billion FLOPs). The dotted shortcuts increase dimensions.

splits. The metric used is accuracy: each clip is assigned an action label corresponding to the maximum value among the scores returned by the action classifiers.

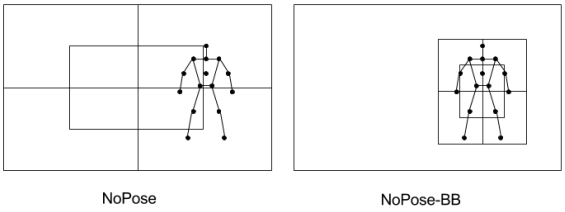


Figure 4. Illustration of the cropping approaches without Pose information. ***NoPose*** : crops of an image subdivided in quarters and a center crop regardless the human bounding box. ***NoPose-BB***: similar to *NoPose* but within the human bounding box.

Figure 5. Examples from JHMDB dataset

| Parts | JHMDB-GT | | |
|---|---|---|---|
| | App | OF | App + OF |
| Hands | 46.3 | 54.9 | 57.9 |
| Upper body | 52.8 | 60.9 | 67.1 |
| Full body | 52.2 | 61.6 | 66.1 |
| Full body | 43.3 | 55.7 | 61.0 |
| All | **60.4** | **69.1** | **73.4** |

Table 2. Performance of appearance-based (App) and flow-based (OF) P-CNN features. Results are obtained with maxaggregation for JHMDB-GT (% accuracy)

## 7. Experimental results

This section describes our experimental results and highlight important experiments done on the original approach in [4]. First, we present the evaluation of the complementarity of different human parts in Section7.1. We then compare the performance of the Original P-CNN and our improved P-CNN with ResNet in Section7.2.

### 7.1. Performance of human part features

Table 2 compares the performance of human part CNN features for both appearance and flow on JHMDB-GT (the JHMDB dataset with ground-truth pose). We can observe that all human parts (hands, upper body, full body) as well as the full image have similar performance and that their combination improves the performance significantly. We

can also observe that flow descriptors consistently outperform appearance descriptors by a significant margin for all parts as well as for the overall combination *All*. Moreover, we can observe that the combination of appearance and flow further improves the performance for all parts including their combination *All*. This is the pose representation used in the rest of the evaluation.

### 7.2. Performance of different pre-trained CNN models (Vgg-f vs ResNet)

## References


[1] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *European conference on computer vision*, pages 25–36. Springer, 2004.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[3] A. Cherian, J. Mairal, K. Alahari, and C. Schmid. Mixing body-part sequences for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2353–2360, 2014.

[4] G. Chéron, I. Laptev, and C. Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3218–3226, 2015.

[5] G. Gkioxari and J. Malik. Finding action tubes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 759–768, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[7] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3192–3199, 2013.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[9] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[12] B. Ni, V. R. Paramathayalan, and P. Moulin. Multiple granularity analysis for fine-grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 756–763, 2014.

[13] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer, 2010.

[14] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[15] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[17] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[18] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4694–4702, 2015.

[19] Y. Zhou, B. Ni, R. Hong, M. Wang, and Q. Tian. Interaction part mining: A mid-level approach for fine-grained action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3323–3331, 2015.