Deep Moving Poselets for Video Based Action Recognition

Effrosyni Mavroudi Lingling Tao René Vidal
Center for Imaging Science, Johns Hopkins University, Baltimore MD 21218, USA
{emavrou1@, ltao4@, rvidal@cis.}jhu.edu

Abstract

We propose a new approach to action classification in video, which uses deep appearance and motion features extracted from spatio-temporal volumes defined along body part trajectories to learn mid-level classifiers called deep moving poselets. A deep moving poselet is a classifier that captures a characteristic body part configuration, with a specific appearance and undergoing a specific movement. By having this mid-level representation of a body part be shared across action classes and by learning it jointly with action classifiers, we obtain a representation that is interpretable, shared and discriminative. In addition, by using sparsity-inducing norms to regularize action classifiers, we can reduce the number of deep moving poselets used by each class without hurting performance. Experiments show that the proposed method achieves state-of-the-art performance on the popular and challenging sub-JHMDB and MSR Daily Activity datasets.

1. Introduction

Human action recognition is an active research field with a plethora of applications, such as surveillance and human-robot interaction. Despite continuous efforts from the research community, it still remains a challenging task due to variations in the execution of actions, variations in scale and viewpoint as well as occlusions. Even deep learning methods [8, 20, 3] have not yielded significant advances over hand-crafted descriptors such as improved dense trajectories (iDT) [25]. This could be attributed to limited training data, since video collection and annotation are expensive and time-consuming tasks. Another possible explanation is that many of the popular deep learning frameworks for action recognition operate on whole frames or cropped patches, which might not be sufficiently discriminative of where and when the action occurs. One solution could be to focus on salient spatio-temporal regions [30, 36]. Video sub-volumes along human body parts would be ideal for it, since intuitively an action can be thought as a composition of action primitives, i.e., characteristic configurations and movements of body parts. This was confirmed by Cheron et al. [1], who proposed to use P-CNN features extracted along body part tracks. Still the straightforward approach of extracting deep features around body parts at each frame, temporally pooling these features using max or average pooling, and then classifying the video using a linear SVM, is not always able to outperform results obtained by iDT [1].

The success of iDT relies on not only an effective description of appearance and motion, but also mid-level representations such as the Fisher Vector (FV) [17], which capture higher order statistics about features. Indeed, the work of [31] recently showed that combining deep features with high dimensional encodings such as FV can further boost performance in action detection. However, one shortcoming of such mid-level features is their lack of interpretability as they do not shed light into the spatial configurations, movements and hand-object interactions that are most discriminative for each action. To address this issue, another line of research focuses on training mid-level part classifiers/templates [5, 34, 23, 11] which, besides their interpretability, perform very well, and are computationally efficient as their responses lead to a much lower-dimensional representation of a video in comparison to high dimensional encodings such as FV. However, these methods have not yet explored the approach of learning mid-level classifiers based on deep features, which was recently shown to outperform mid-level classifiers learned with hand-crafted features for the task of human pose retrieval [6].

In this paper we propose a new approach to action classification in video, which combines the strengths of deep features, part-based features, and discriminative mid-level representations. Specifically, we propose to extract deep features from short spatio-temporal volumes extracted around the 2D trajectories of a hierarchy of body parts ranging from small rigid parts, like arms, to the whole body. These features are then fed to mid-level spatio-temporal part classifiers that are designed to be sensitive to specific body part pose configurations (e.g., hands above head) and movement patterns (e.g., hand moving forward), hence they are naturally defined to be *interpretable*. By having a common dictionary of mid-level classifiers for each body part, our

978-1-5090-4822-9/17 $31.00 © 2017 IEEE
DOI 10.1109/WACV.2017.20

mid-level representation is naturally *shared* among different action classes, e.g., a mid-level classifier for a hand-to-mouth movement and pose configuration can be shared between actions *drink* and *eat*. Moreover, by jointly learning mid-level and action classifiers, and properly regularizing the action classifier weights, our mid-level classifiers are designed to be *discriminative* in the sense that not all of them are needed to classify an action. For instance, the appearance of arms when playing the guitar or the motion of a leg when kicking a ball are examples of mid-level parts that are discriminative.

In summary, the contributions of this work are three-fold. First, we learn interpretable, shared and discriminative mid-level classifiers that capture spatio-temporal configurations of body parts during different phases of an action. We call our mid-level classifiers *deep moving poselets*, as they are a natural extension of the moving poselet representation of skeletal data proposed in [22]. One difference is that 3D joint positions are typically measured in the case of skeletal data, while in the case of video data, 2D joint positions are in general unknown and need to be annotated or estimated. Furthermore, local features around joints are not sufficient for capturing the rich information conveyed by the appearance and the motion of whole body parts. Our second contribution is a method for constructing short tubelets around a hierarchy of body parts using 2D joint positions and a sliding window. Deep features extracted from such tubelets provide a powerful descriptor for the appearance and motion of body parts. Finally, we evaluate our method on two challenging datasets and show that it achieves state-of-the-art performance on both of them. We also show that sparsity-inducing norms can be integrated in our framework to reduce the number of deep moving poselets used by each class without hurting performance.

2. Related Work

The most popular action recognition methods are based on holistic features extracted from whole videos, which are typically fed directly to a classifier, such as a linear SVM, or are first encoded by a a high dimensional feature encoder, such as FV. Typical features are either hand-crafted, such as iDT [26], or deep, like the two-stream convolutional network [20] and other networks [8, 30]. While such features yield state-of-the-art results, most of them lack interpretability in terms of body parts involved and their characteristic movements. Instead, we focus on body parts by building short spatio-temporal volumes along their trajectories and extracting features from such volumes.

Many works are based on pooling features extracted from local spatio-temporal volumes, such as densely selected video sub-volumes [16, 19], volumes from a spatio-temporal pyramid [35], randomly selected volumes [5, 33], volumes enclosing the action of interest [23] or the actor [21] and volumes around objects by linking object proposals [34]. In this work, we focus on the appearance and motion of both individual body parts as well as the whole human body. There are many ways of describing a body part in video data: for instance, one can use the position and velocity of joints, as in pose-based methods, e.g., [24, 9], or can exploit appearance information by describing a spatio-temporal neighborhood around each joint belonging to the body part [11] or inferring a coarse tubelet that encloses the body part using 2D joint position information [1]. We adopt the latter strategy and place our volumes of interest around body part trajectories, with the difference that our tubelets span short temporal windows rather than the whole video as in [1]. This allows us to learn action primitives that are compact in space (capturing the appearance of a certain body part) and in time (capturing short movement patterns of a certain body part). We also build a two-layer model on top of these features, which aims at jointly training discriminative part classifiers (deep moving poselets) and action classifiers, while in [1] they used a simple linear SVM.

Learning mid-level feature representations for action recognition is also an active research field. Some approaches learn linear classifiers based on image patches [14, 18, 4], while others, including ours, learn them based on spatio-temporal volumes [35, 5, 34, 23, 11]. A common step among many of these approaches is to learn a separate set of discriminative linear classifiers/poselets for each action, employing methods such as seeding, expansion and selection [29, 34], multiple instance learning [35, 19] or clustering and hard mining [33]. In contrast, we jointly train deep moving poselets and action classifiers, without having to resort to clustering and hard-mining, which yields a set of discriminative part classifiers per action class. Moreover, as we will show, our method encourages deep moving poselets to be shared among action classes and is therefore more scalable. Joint learning of mid-level linear classifiers and action classifiers for video action recognition has been also exploited by [18, 11]. Our approach differs from these two works not only in the volumes selected and features extracted from them, but also in the model used. Instead of using a (latent) Structural SVM, we prefer the two-layer model proposed in [22] for finding moving poselets from skeleton data, because it can be trained in an efficient way using stochastic gradient descent.

3. Deep Moving Poselets: A Mid-Level Feature Representation for Action Recognition

In this section we describe our approach to extracting deep features from spatio-temporal volumes defined along human body part trajectories as well as the proposed model for learning mid-level spatio-temporal part detectors (deep moving poselets). Our approach is illustrated in Fig. 1.

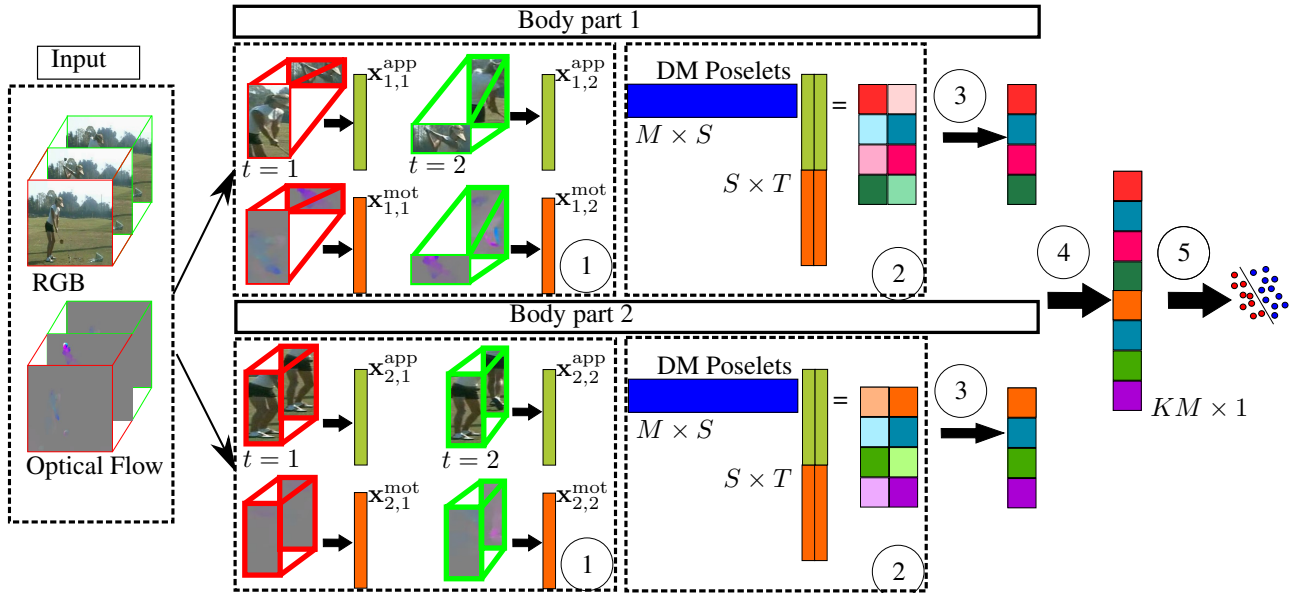


Figure 1. Our framework consists of 5 main steps: 1) extract spatio-temporal volumes (tubelets) along the trajectories of human body parts in a sliding window fashion and compute CNN features for the description of the appearance ($\mathbf{x}^{\text{app}}_{k,t}$) and motion ($\mathbf{x}^{\text{mot}}_{k,t}$) of each one of these tubelets, 2) compute responses of all deep moving poselets for each tubelet, i.e., each body part and temporal window, 3) max-pool the responses for each deep moving poselet and body part temporally, 4) concatenate these responses and 5) apply a linear SVM action classifier. Deep moving poselets and action classifiers are trained jointly. For illustration purposes, we assume $K = 2$ body parts, $T = 2$ temporal windows and $M = 4$ poselets per body part. (Best viewed in colour).

3.1. Low-level Features from Body Part Tubelets

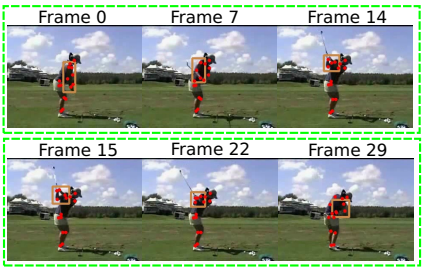
Let the input be a video of a person performing an action. Assume that the 2D image positions of all joints, for all frames are either annotated or estimated using pose estimation methods, such as [32]. Assume also that the human body is divided in K body parts, such as arms, legs, upper body, lower body, full body, etc., where each part is defined as a collection of joints (see Fig. 2a).

Tubelets. For each body part, we define a 2D bounding box containing all the joints for that body part. By tracking the motion of this bounding box throughout the video, we obtain a spatio-temporal tube, which we subdivide into T (possibly overlapping) *tubelets* of L consecutive frames, as illustrated in step 1 of Fig. 1. More specifically, to form a tubelet surrounding body part k, e.g., the right arm, we find the tightest bounding box that encloses all the joints belonging to body part k. Since joint positions do not specify the spatial extent of each body part, which depends on its scale, we enlarge the bounding box of body part k by $m_x^{(k)}$ and $m_y^{(k)}$ pixels in the horizontal and vertical direction, respectively, to include both the body part as well as the context around it. Including context is very important since contextual cues, such as the place where the action takes place or the objects used, are highly discriminative of the action. For large body parts, such as the full body, upper body and lower body, bounding boxes can be coarse so that they contain more background, while for small body parts, such as left and right leg, bounding boxes should ideally be tight enough so that they focus on the corresponding body part. Also note that, unlike in the case of skeleton data where body parts such as right and left arm have a unique set of features defined by the positions and velocities of the joints belonging to each one of them, in the case of video data, regardless of the way that a bounding box surrounding a body part is constructed, it can also partially or fully surround other body parts, as shown in Fig. 2b. This is caused by the high complexity of human poses and the variety of camera viewpoints, and constitutes an additional challenge for learning mid-level part classifiers from video.

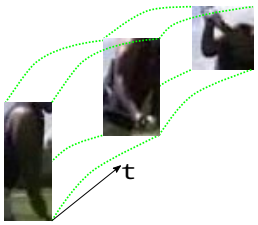
Deep features. From each tubelet, we extract appearance and motion features and encode them into a representation of fixed size S. Although our framework can make use of any features extracted from a spatio-temporal volume, we choose to use CNN features. Specifically, we extract deep features from each body part tubelet using the deep network architecture in [1], which consists of a spatial CNN operating on RGB patches and a motion CNN operating on optical flow patches. Since the size of a tubelet patch changes with time, the inputs to both CNNs are resized to 224×224. Using these networks we compute for each frame a 4096-dimensional appearance vector and a 4096-dimensional motion vector (fc7 layer outputs). We then compute a representation for each tubelet by ap-

Body part	Joint Set
Torso	neck, r_hip, l_hip
Right arm	r_shoulder, r_elbow, r_wrist
Left arm	l_shoulder, l_elbow, l_wrist
Right leg	r_hip, r_knee, r_ankle
Left leg	l_hip, l_knee, l_ankle
Torso & head	head, neck. belly, head, l_shoulder, r_shoulder
Upper body	head, {Right arm}, {Left arm}
Lower body	hip, {Right leg}, {Left leg}
Full upper body	head, {Right arm}, {Left arm}, r_hip, l_hip
Full body	all joints

(a)



(b)



(c)

Figure 2. Body part tubelets. (a) Body parts defined in terms of joints forming their boundaries. (b) Example of two tubelets along body part *right arm* for two temporal windows of a video belonging to action *golf*. Each tubelet is represented by a sampled sequence of frames (first, middle and last) and a bounding box. (c) Zooming into a tubelet along body part *right arm*.

plying temporal max-pooling to both appearance and motion descriptors, i.e., by computing the maximum value of each entry of these descriptors over all frames inside the tubelet. The tubelet descriptors for appearance and motion are then separately normalized by dividing them by the average L_2-norm of the CNN features computed from the training set and are concatenated into the final feature vector $\mathbf{x}_{k,t} = \left[\mathbf{x}^{\text{app}}_{k,t}, \mathbf{x}^{\text{mot}}_{k,t}\right] \in \mathbb{R}^S$ for the tubelet corresponding to the kth body part and the tth temporal window.

3.2. Deep Moving Poselets: An Interpretable and Discriminative Mid-Level Representation for Action Classification

Given spatio-temporal features extracted from multiple tubelets in a video, which correspond to a collection of K body parts, our goal is to learn an interpretable, shared, and discriminative mid-level representation for action classification. Following the moving poselet representation of skeletal data proposed in [22], our mid-level representation for video data consists of a set of dictionaries $\mathbf{D}_k \in \mathbb{R}^{S\times M}$, one per body part $k \in \{1,\dots,K\}$, where each column of $\mathbf{D}_k$ represents a linear classifier. In the case of skeletal data, these classifiers were tuned to be sensitive to specific 3D pose configurations and motion patterns of the corresponding body part. In the case of video data, these classifiers are tuned to be sensitive to specific 2D pose configurations and motion patterns of a body part as well as specific appearance patterns obtained from deep features, hence the name *deep moving poselets*. Formally, let $\mathbf{X}_k = [\mathbf{x}_{k,1},\dots,\mathbf{x}_{k,t},\dots,\mathbf{x}_{k,T}] \in \mathbb{R}^{S\times T}$ be the time-series of features from the kth body part. Assuming that $\mathbf{D}_k$ is a dictionary of M linear classifiers with biases $\mathbf{d}_k \in \mathbb{R}^M$, the response map obtained after applying $(\mathbf{D}_k, \mathbf{d}_k)$ to the tth data point of the time series for the kth body part is given by $\mathbf{h}_{k,t} = \mathbf{D}_k^\top \mathbf{x}_{k,t} + \mathbf{d}_k$. Intuitively, the response map captures the similarity between each one of the patterns represented by the mid-level classifiers and the pattern described by feature $\mathbf{x}_{k,t}$. Specifically, a high response indicates high similarity, while a low response indicates low similarity.

To obtain a single descriptor for the time series $\mathbf{X}_k$ from the kth body part, we apply temporal max-pooling to the response map $\mathbf{h}_{k,t}$ for the M mid-level classifiers in $\mathbf{D}_k$. Formally, we compute an M-dimensional vector $\mathbf{f}$, which we call the *moving poselet activation vector*, whose jth entry is given by $\mathbf{f}(\mathbf{X}_k;\mathbf{D}_k,\mathbf{d}_k)^{(j)} = \max_{t=1,\dots,T} \mathbf{h}^{(j)}_{t,k}$. The final representation $\mathbf{F}$ for all K time series $\mathbf{X} = \{\mathbf{X}_k\}_{k=1}^K$ extracted from the video is obtained as the concatenation of the representations of the K time series, i.e.,

$$\mathbf{F}(\mathbf{X};\mathbf{D},\mathbf{d}) = [\mathbf{f}(\mathbf{X}_1;\mathbf{D}_1,\mathbf{d}_1),\dots,\mathbf{f}(\mathbf{X}_K;\mathbf{D}_K,\mathbf{d}_K)]\,, \quad (1)$$

where $\mathbf{D} = \{\mathbf{D}_k\}_{k=1}^K$ and $\mathbf{d} = \{\mathbf{d}_k\}_{k=1}^K$.

To classify a video, we first apply a rectified linear unit $\phi(x) = \max(x, 0)$ to each entry of its representation $\mathbf{F} \in \mathbb{R}^{KM}$, and then give it as the input to linear action classifiers $\{(\mathbf{W}_c, b_c)\}_{c=1}^C$, where C is the number of classes and $\mathbf{W}, \mathbf{b}$ are the weight and bias matrices. This gives the *class activation scores* $g_c = \mathbf{W}_c^\top \phi(\mathbf{F}(\mathbf{X},\mathbf{D},\mathbf{d})) + b_c$, from which the class of the video is obtained as $c^* = \text{argmax}_c\, g_c$.

3.3. Joint Learning of Deep Moving Poselet Classifiers and Action Classifiers

Given N training examples, $\{\mathbf{X}^{(n)}, Y^{(n)}\}_{n=1}^N$, where $\mathbf{X}^{(n)}$ is a set of K time series of length T_n extracted from the nth video sequence and $Y^{(n)} \in \{1,\dots,C\}$ is the action class associated with that video, the moving poselet classifiers $\mathbf{D}$ and $\mathbf{d}$ are learned jointly with the action classifiers $\mathbf{W}$ and $\mathbf{b}$ by solving the following optimization problem:

$$\min_{\substack{\mathbf{D},\mathbf{d},\\ \mathbf{W},\mathbf{b}}} \sum_{c=1}^{C}\sum_{n=1}^{N} \max(0, 1 - Y_{cn}(\mathbf{W}_c^\top \phi(\mathbf{F}(\mathbf{X}^{(n)},\mathbf{D},\mathbf{d})) + b_c)) + \lambda(R_W(\mathbf{W}) + R_D(\mathbf{D})). \quad (2)$$

Here Y_{cn} is equal to 1 if sample n has label c and -1 otherwise, while R_W and R_D are the regularizers for action classifiers and deep moving poselets, respectively, and λ is

a regularization parameter. We can use the ℓ_2-norm regularizer for both deep moving poselet classifiers and action classifiers, i.e., $R_D(\mathbf{D}) = \|\mathbf{D}\|_F^2$ and $R_W(\mathbf{W}) = \|\mathbf{W}\|_F^2$, respectively, as suggested in [22]. Since we use a single dictionary of poselets per body part, which is shared for all actions, and the ℓ_2 regularizer penalizes all entries in $\mathbf{W}$ equally so that all action classifiers can select as many poselets as needed, the ℓ_2 regularizer encourages sharing of deep moving poselets among action classifiers. This captures the intuition that different actions can be obtained as a combination of the same motion primitives. At the same time, this makes our action recognition framework more scalable with respect to the number of classes than other approaches which learn a separate set of discriminative mid-level classifiers for each action. On the other hand, we can encourage each action classifier to select a small subset of deep moving poselets depending on the complexity of the action (e.g., a complex action may need more poselets than a simple action) by using a sparsity-inducing norm to regularize $\mathbf{W}$, such as the ℓ_1-norm $\|\mathbf{W}\|_1$. As suggested in [13] for applications in object recognition, we can balance sharing and sparsity by using a convex combination of the ℓ_2 and ℓ_1 norms, such as the elastic net regularizer [37], $R_W(\mathbf{W}) = \alpha\|\mathbf{W}\|_F^2 + (1-\alpha)\|\mathbf{W}\|_1$, where the first term promotes sharing, while the second one promotes sparsity. In addition, it is also possible to use sparsity-inducing norms with respect to the number of body parts involved in an action, which facilitates interpreting which body parts and body movements are most discriminative for an action, though we do not explore such regularizers in this paper.

In summary, by using proper regularization, our framework can learn a compact, shared, discriminative and interpretable mid-level representation of actions from video.

4. Experimental Evaluation

We evaluate our method on two challenging datasets for action recognition: sub-JHDMB [7] and MSR Daily Activity 3D [27]. These datasets provide both RGB videos in which the whole human body is visible and 2D joint positions, which we need for constructing bounding boxes enclosing body parts. First, we report our results on each dataset and compare them with the state of the art. Next, we examine the effect of different model components and show qualitative results suggesting that the learned deep moving poselets are discriminative, shared and interpretable.

4.1. Implementation Details

We implemented our model in Keras [2] and trained it using backpropagation with a batch size of 10 samples, a learning rate of 0.01, a momentum of 0.9, 200 epochs and a regularization parameter $\lambda = 10^{-4}$. The learning rate was reduced in half every 20 and 50 epochs for sub-JHMDB and MSR Daily Activity 3D, respectively.

4.2. Experiments on sub-JHMDB

The sub-JHMDB dataset consists of 316 realistic videos belonging to 12 full body action classes, such as catch, golf, swing baseball, run and walk. This dataset is challenging due to variations in position, scale, viewpoint of human actors, as well as varying video quality and camera motion.

For our experiments on sub-JHMDB, we form tubelets around body parts: *torso*, *right arm*, *left arm*, *right leg*, *left leg*, *torso and head*, *upper body*, *lower body*, *full body*. We enlarge tight bounding boxes by $m_x = \epsilon w$ and $m_y = \epsilon h$, where h, w are the height and width of the bounding box and ϵ is chosen as 0.5 for body parts *upper body*, *lower body*, *full body* and 0.25 for the remaining body parts. We set the number of deep moving poselets per body part to $M = 50$ after 5-fold cross-validation. Besides, we use the ℓ_2-norm regularizer for both mid-level classifiers and action classifiers. We use a sliding window of length $L = 15$ frames and a step of 5 frames.

Table 1 shows the average classification accuracy over the three standard train/test splits suggested by the authors [7] on the sub-JHMDB dataset using either the human annotated 2D joint positions provided with this dataset or the pose estimated joint annotations provided by [1]. Our method obtains state-of-the-art results for both configurations. Note that we do not use the joint positions as features, but only to coarsely localize bounding boxes around body parts, so there is no fair comparison between our results and methods that explicitly use pose features. Specifically, when using pose-estimated positions, we observe from the last column of Table 1 that our method outperforms all state-of-the-art methods, even those using explicit pose features, such as joint positions, velocities and angles. Note that our method is more robust to inaccurate pose estimation results than pose-based methods, such as [7]. In particular, although our action classification performance drops by 9% when moving from joint annotations to pose estimated joints, this drop is smaller than the one reported by Jhuang et al. [7], who use explicit pose features. In their case, performance drops from 75.5% to 52.9%. Our method also outperforms P-CNN [1], which is another work based on body parts, by using tubelets which are compact in space and time and learning discriminative mid-level classifiers.

4.3. Experiments on MSR Daily Activity 3D

The MSR Daily Activity 3D dataset consists of 320 videos of 10 subjects performing 16 different daily activities in two different settings: "sitting on sofa" and "standing". Although these videos are captured in a controlled setting resembling a living room and there are no significant viewpoint and scale changes as in the sub-JHMDB dataset, this dataset is also very challenging because joint positions are captured by a Kinect device and are hence very noisy, especially when subjects sit or walk behind the sofa.

Method	Features	Accuracy (%)	
		GT Joints	PE Joints
DT [7]	RGB	46.0	
NTraj [7]	2D Pose	75.1	54.1
DT + NTraj [7]	RGB + 2D Pose	75.5	52.9
MST-AOG [28] [15]	RGB + 2D Pose	-	45.3
AOG [15]	RGB + 2D Pose	-	61.2
[11]	RGB + 3D Pose	77.5	-
P-CNN [1]	RGB + Bps	72.5	66.8
Ours	RGB + Bps	**79.2**	**70.2**

Table 1. Recognition accuracy in the sub-JHMDB dataset. (RGB features: any features extracted from RGB videos, including optical flow features, GT joints: annotated joints, PE joints: pose estimated joints, Bps: body parts.)

Method	Accuracy (%)
RGB	
STIP [10] [28]	54.5
DT [25] [34]	71.7
MST-AOG [28][1]	73.1
IPM [34]	**83.3**
RGB + Pose	
IPM+Joints [34]	**89.3**
RGB + Body parts	
Ours	**84.4**

Table 2. Recognition accuracy in the MSR Daily Activity 3D dataset using 2D joint positions captured by a Kinect device.

Method	Accuracy (%)
app, full body, no sliding window	60.3
mot, full body, no sliding window	66.1
app+mot, full body, no sliding window	74.3
app+mot, all bps, no sliding window	77.7
app+mot, all bps, with sliding window	79.2

Table 3. Contribution of each component to recognition accuracy on sub-JHMDB using annotated joints. (app: appearance features, mot: motion features, bps: body parts)

For our experiments on MSR Daily Activity 3D, we form tubelets only around body parts: *upper body*, *lower body* and *full body*, because joint positions from Kinect are very noisy, and it is hard to center tubelets around fine-grained body parts such as arms. Since human scale in this dataset is almost the same for all subjects and actions, we enlarge tight bounding boxes by a fixed number of pixels, $m_x = m_y = 40$. We set the number of deep moving poselets per body part to $M = 333$ after 5-fold cross-validation. We use ℓ_2-norm regularizer for both deep moving poselet and action classifiers. We also use a sliding window of length $L = 15$ frames and a step of 5 frames.

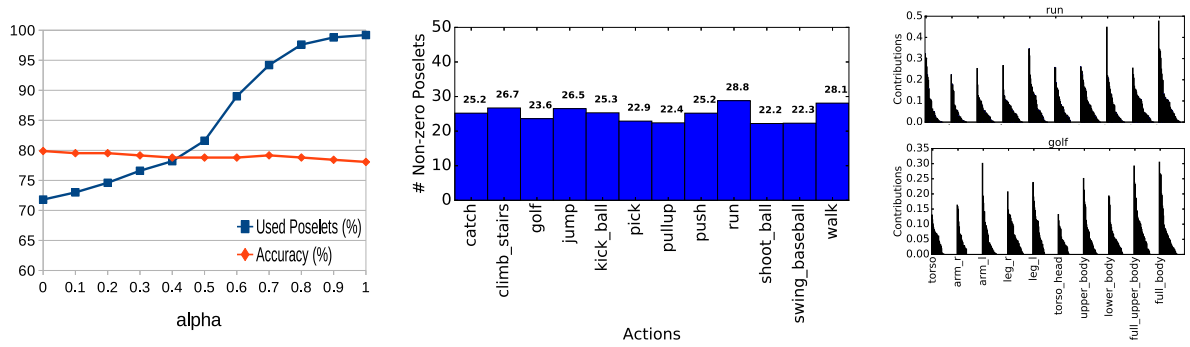
Table 2 compares our method to state-of-the-art methods based on RGB and/or 2D joint information on the MSR Daily Activity 3D dataset. We do not compare against methods that rely on 3D skeleton and/or depth cues, since our method does not use these information streams. We use the cross-subject evaluation method proposed in [27], namely we use subjects 1, 3, 5, 7 and 9 for training, and the rest for testing. Our method obtains results comparable to the state of the art. The best performing method on this dataset, IPM [34], also finds discriminative mid-level part classifiers, which are however centered around object proposals. These mid-level classifiers are beneficial in this case, since 10 out of the 16 action classes of this dataset involve some kind of human-object interaction. Nevertheless, our method is more general, since it does not restrict itself in actions that involve hand-object interaction. Also note that our method outperforms IPM when no extra information about the 2D joint locations is used.

4.4. Ablation Analysis

In Table 3 we analyze the contribution of the key components of our proposed method, namely the contribution of a) combining appearance and motion information streams, b) forming short tubelets in a sliding window fashion and c) using body parts. To accomplish this analysis, we start by using just the full body tubelet that spans the whole video and using either appearance or motion features to learn 500 deep moving poselets. As expected, motion information alone is more important than appearance information. The difference in performance can also be attributed to the CNNs used, since the spatial CNN was trained on ImageNet for image classification, while the optical flow CNN was trained on UCF101 for the task of action classification [1]. Combining both information streams we get an increase of 8%. Considering all 10 body parts and still using tubelets spanning the whole length of the video, we get an average action classification accuracy of 77.7% for 50 deep moving poselets per body part. Finally, by introducing a sliding window of 15 frames and forming spatio-temporal tubes that are compact in space and time, we get an additional increase of 2%.

In the next experiment, we showcase the superiority of our two-layer model in comparison to single layer models. Specifically, we show how our two-layer model can improve the result obtained by Cheron et al. [1] on sub-JHMDB with annotated joints using SVM classifiers. Using their code and parameters we extract P-CNN features from the 5 body parts they use (full image, full body, upper body and hands) and aggregate them in time using temporal max-pooling. P-CNN with SVM obtains an accuracy of 72.5%. Using our model with 50 poselets per body part (250 in total) and without using a sliding window, to allow

[1]MST-AOG [28] is trained using both RGB videos and 3D joint positions for training, but is applied on RGB videos for testing



(a) (b) (c)

Figure 3. (a) Percentage of used deep moving poselets and classification accuracy in sub-JHMDB for varying values of the elastic net regularizer parameter α (results averaged over 3 splits). (b) Average number of deep moving poselets used per body part for the 12 classes in sub-JHMDB. (c) Distribution of absolute action classifier weights per body part for actions *golf* and *run* ($\alpha = 0.5$, split $= 2$).

Method	Accuracy (%)
P-CNN + SVM [1]	72.5
P-CNN + DMPs (no sliding window)	74.3
P-CNN + DMPs (with sliding window)	76.9

Table 4. Effect of mid-level representation on action recognition accuracy on sub-JHMDB with manually annotated joints (DMPs: deep moving poselets).

for a fair comparison with [1], we get 74.3%, as shown in Table 4. Adding the sliding window, we reach 76.9% accuracy. This justifies the use of deep moving poselets as a mid-level representation.

This final experiment examines the effect of the regularizer used on the action classifier weights $R_W(\mathbf{W})$. In the previous experiments we used an ℓ_2-norm regularizer, $R_W(\mathbf{W}) = \|\mathbf{W}\|_F^2$, which encourages the utilization of all deep moving poselets belonging to all body parts by all actions. Here, we use the elastic net regularizer, $R_W(\mathbf{W}) = \alpha\|\mathbf{W}\|_F^2 + (1-\alpha)\|\mathbf{W}\|_1$, which promotes selecting a smaller number of poselets for each action class.

Fig. 3 shows the number of deep moving poselets used by at least one action class together with the action classification accuracy for various values of α ($\alpha = 0$ corresponds to ℓ_1-norm regularization, while $\alpha = 1$ corresponds to ℓ_2-norm regularization). It can be observed that increasing sparsity, and therefore reducing the number of used poselets, does not hurt performance. We also compare the number of discarded poselets between an easy and a hard to classify action , where difficulty in classifying an action is measured in terms of the average recognition accuracy. For instance, *golf* is an easy to classify action with recognition accuracy 90%, while *run* is a hard to classify action with recognition accuracy 30%. Indeed, our method chooses to use more poselets to discriminate *run* from other similar classes, e.g., *walk* than to classify *golf*.

4.5. Qualitative results

To begin with, we visualize some of the learned deep moving poselets for the sub-JHMDB dataset with annotated joints. First, for each action class, we find the 5 deep moving poselets with the highest action classification weights for that action. Then, for each chosen deep moving poselet, we find the body part tubelet in the training data that gives the highest activation score. In Fig. 4 we show examples of deep moving poselets that are highly discriminative for actions *catch*, *swing baseball* and *pullup*. For instance, for the action *catch*, our framework utilizes poselets that capture discriminative spatio-temporal configurations, such as the configuration of the whole body at two phases of the action, namely waiting for the ball and catching the ball, the configuration of the upper body when catching the ball and finally the characteristic pose and movement of the lower body. Note that the tubelet that gives the highest activation score for each one of these poselets chosen by each class does not necessarily belong to the same class, but still it captures a salient movement of the class.

Our next experiment delves deeper into this desirable property of some of our deep moving poselets, which capture poses/movements that are shared among action classes. For each deep moving poselet we find from training data the tubelet from each action that has the highest response to the poselet and sort them in descending order based on their response. Fig 5 shows 5 deep moving poselets that capture spatio-temporal configurations that are shared by at least 3 actions. For example, the first row shows a poselet that captures the "left hand crossing torso" movement in actions *golf*, *kick ball* and *swing baseball*. The second row shows a poselet that is sensitive to the upwards motion of the upper body, shared by *pull up*, *catch* and a specific instance of *kick ball*. Another deep moving poselet captures the "torso parallel to ground" configuration, present in actions such as *pick*, *push* and some instances of *jump*.

Figure 4. From left to right: each column shows the 5 most significant deep moving poselets for action classes *catch*, *swing baseball*, *pullup*. Each deep moving poselet is represented by the body part tubelet with the highest activation from training data of sub-JHMDB with annotated joints. Each tubelet is illustrated by a sampled sequence of frames (first, middle and last) and a bounding box.

Figure 5. Examples of deep moving poselets shared among action classes in the sub-JHMDB dataset (split 2). Each row shows 3 tubelets from different classes with high activations for a specific poselet. Each tubelet is represented by a sampled sequence of frames (first, middle and last) and a bounding box.

5. Conclusion

We have presented a new method for video-based action classification, which constructs tubelets around body parts, describes them using CNN features and efficiently learns a set of mid-level classifiers called *deep moving poselets*, which capture characteristic spatio-temporal configurations of body parts in different phases of actions. Joint learning of mid-level and action classifiers leads to state-of-the-art performance in two challenging datasets, as well as shared and discriminative deep moving poselets. Although our method constructs tubelets based on 2D joint positions, our experiments showed that it can be successfully combined with a pose estimation algorithm and it is robust to pose estimation errors. Promising future directions include the use of spatio-temporal regions obtained by action proposal networks as well as extending our framework to action detection.

6. Acknowledgements

The authors thank Colin Lea for his insightful comments. This work was supported by NIH grant R01HD87133-01.

References

[1] G. Chéron, I. Laptev, and C. Schmid. P-CNN: Pose-Based CNN Features for Action Recognition. In *IEEE International Conference on Computer Vision*, pages 3218–3226, Dec. 2015. 1, 2, 3, 5, 6, 7

[2] F. Chollet. Keras. https://github.com/fchollet/keras, 2015. 5

[3] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional Two-Stream Network Fusion for Video Action Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016. 1

[4] G. Gkioxari and J. Malik. Finding action tubes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 759–768, June 2015. 2

[5] A. Jain, A. Gupta, M. Rodriguez, and L. S. Davis. Representing Videos Using Mid-level Discriminative Patches. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2571–2578, June 2013. 1, 2

[6] N. Jammalamadaka, A. Zisserman, and C. V. Jawahar. Human pose search using deep poselets. In *IEEE International Conference on Automatic Face and Gesture Recognition*, volume 1, pages 1–8, May 2015. 1

[7] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards Understanding Action Recognition. In *IEEE International Conference on Computer Vision*, pages 3192–3199, Dec. 2013. 5, 6

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1, 2

[9] I. K. Khai Tran and S. Shah. Modeling motion of body parts for action recognition. In *British Machine Vision Conference*, pages 64.1–64.12. BMVA Press, 2011. 2

[10] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005. 6

[11] I. Lillo, J. Carlos Niebles, and A. Soto. A Hierarchical Pose-Based Approach to Complex Action Understanding Using Dictionaries of Actionlets and Motion Poselets. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1981–1990, 2016. 1, 2, 6

[12] I. Lillo, A. Soto, and J. C. Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[13] H. Lobel, R. Vidal, and A. Soto. Learning shared, discriminative, and compact representations for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2218–2231, 2015. 5

[14] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3177–3184, 2011. 2

[15] B. X. Nie, C. Xiong, and S. C. Zhu. Joint action recognition and pose estimation from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1293–1301, June 2015. 6

[16] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action Recognition with Stacked Fisher Vectors. In *European Conference on Computer Vision*, Sept. 2014. 2

[17] 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, 2010. 1

[18] M. Raptis and L. Sigal. Poselet key-framing: A model for human activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 2

[19] M. Sapienza, F. Cuzzolin, and P. H. S. Torr. Learning Discriminative Space–Time Action Parts from Weakly Labelled Videos. *International Journal of Computer Vision*, 110(1):30–47, Oct. 2013. 2

[20] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Neural Information Processing Systems*, pages 568–576, 2014. 1, 2

[21] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao. A Multi-Stream Bi-Directional Recurrent Neural Network for Fine-Grained Action Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1961–1970, 2016. 2

[22] L. Tao and R. Vidal. Moving poselets: A discriminative and interpretable skeletal motion representation for action recognition. In *ChaLearn Looking at People Workshop 2015*, 2015. 2, 4, 5

[23] D. Tran and L. Torresani. EXMOVES: Mid-level Features for Efficient Action Recognition and Video Analysis. *International Journal of Computer Vision*, 119(3):239–253, Apr. 2016. 1, 2

[24] C. Wang, Y. Wang, and A. L. Yuille. An Approach to Pose-Based Action Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 915–922, June 2013. 2

[25] H. Wang, A. Klaser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 1, 6

[26] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *IEEE International Conference on Computer Vision*, pages 3551–3558, Dec. 2013. 2

[27] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 5, 6

[28] J. Wang, X. Nie, Y. Xia, Y. Wu, and S. C. Zhu. Cross-View Action Modeling, Learning, and Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2649–2656, June 2014. 6

[29] L. Wang, Y. Qiao, and X. Tang. Motionlets: Mid-level 3d parts for human motion recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 2

[30] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4305–4314, June 2015. 1, 2

[31] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative CNN video representation for event detection. In *IEEE Con-*

ference on Computer Vision and Pattern Recognition, pages 1798–1807, June 2015. 1

[32] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 3

[33] W. Zhang, M. Zhu, and K. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *IEEE International Conference on Computer Vision*, 2013. 2

[34] Y. Zhou, B. Ni, R. Hong, M. Wang, and Q. Tian. Interaction part mining: A mid-level approach for fine-grained action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3323–3331, June 2015. 1, 2, 6

[35] J. Zhu, B. Wang, X. Yang, W. Zhang, and Z. Tu. Action recognition with actons. In *IEEE International Conference on Computer Vision*, 2013. 2

[36] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao. A Key Volume Mining Deep Framework for Action Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1991–1999, 2016. 1

[37] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005. 5