Action Machine: Towards Person-Centric Action Recognition in Videos

Jiagang Zhu, Wei Zou, Zheng Zhu, Liang Xu, Guan Huang

***Abstract*—Existing RGB and CNN-based methods in video action recognition mostly do not distinguish human body from the environment, thus easily overfit the scenes and objects of training sets. In this work, we present a conceptually simple, general and high-performance framework for action recognition in videos, aiming at person-centric modeling. The method, called Action Machine, is based on person bounding boxes for instance-level action analysis. It extends the Inflated 3D ConvNet (I3D) by adding a branch for human pose estimation and a 2D CNN for pose-based action recognition. Action Machine can benefit from the multi-task training of action recognition and pose estimation, the fusion of predictions from RGB images and poses. Experiments results are provided on trimmed video action datasets, NTU RGB+D, Northwestern UCLA Multiview Action3D, MSR Daily Activity3D. Action Machine achieves superior performance and generalizes well across datasets.**

***Index Terms*—Video action recognition, Deep learning, Pose estimation.**

I. Introduction

With the release of Kinetics dataset [1], action recognition in videos has shown similar trend as the object recognition due to the ImageNet [2]. A variety of tasks including trimmed video classification [3], [4], temporal action recognition in untrimmed videos [5], [6], spatial-temporal action detection [7], have been quite popular in recent competitions.

To some extent, advances in video action recognition are hampered by the biases in datasets collection, lack of annotations. For example, the videos in UCF-101 [8] and HMDB-51 [9] are rich in scenes and objects, while missing person bounding box annotations. Previous methods [1], [10], [11], which do not directly distinguish human body from videos, tend to predict an action according to the scenes and objects, since convolutional neural networks (CNNs) make it easier to classify the objects and things than human motions. The trained models can easily be distracted by irrelevant cues of videos when recognizing an action. For example, in Fig. 1(b), the video frame with ground-truth class *carry* is predicted as a wrong action *drop trash* by the baseline Inflated 3D ConvNet (I3D) [1]. Presumably, the model has learned that the *trash can* and the action *drop trash* always appear in a video together (Fig. 1(a)). This motivates us to design a model that can explicitly capture human body movements from videos, simultaneously follows the stream of RGB and CNN-based methods in action recognition.

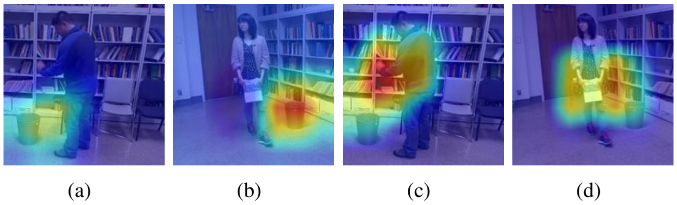

Fig. 1. Visualizing the class-specific activation maps of Inflated 3D ConvNet (I3D) [1] and our method with the Class Activation Mapping [12]. The video frames of two action classes from Northwestern UCLA Multiview Action3D [13] are displayed, i.e., *drop trash*, *carry*, which are acted by a man and a woman respectively. The results of our person-centric modeling method (subfigure (c) and (d)) emphasize the body movements, while the baseline I3D (subfigure (a) and (b)) overfits the *trash can*.

In this work, a person-centric modeling scheme for human action recognition is proposed, called Action Machine, which extends the Inflated 3D ConvNet (I3D) [1] by adding a branch for human pose estimation and a 2D CNN for pose-based action recognition. In details, we use I3D for feature extraction and crop the target persons by bounding boxes. For frame-wise pose estimation, a 2D deconvolution head is added to the last convolutional layer of I3D, in parallel with the existing head for RGB-based action recognition. Following pose estimation, a 2D CNN is applied to the pose sequences for pose-based action recognition. At inference time, the predictions of two classification heads are fused by summation. Some class-specific activation maps of Action Machine are shown in Fig. 1(c) and (d), indicating only the regions that really correspond to the action are activated. The main contributions of this work are summarized as follows:

1) We present a conceptually simple and general framework for action recognition, called Action Machine, aiming at person-centric modeling.
2) The proposed techniques of explicitly modeling human body movements including person cropping, multi-task training of action recognition and pose estimation, the fusion of predictions from RGB images and poses can help to improve the model performance.
3) We showcase the generality of our framework via extensive experiments. Action Machine achieves the state-of-the-art performance on NTU RGB-D [14], Northwestern UCLA Multiview Action3D [13]. In addition, we design a cross-dataset recognition task, which is closer to the

Manuscript received July 31, 2019; accepted September 16. This work is supported in part by the National key research and development program of China under Grant No.2017YFB1300104, the National Natural Science Foundation of China under Grant No.61773374.

J. Zhu, W. Zou, Z. Zhu are with Institute of Automation, Chinese Academy of Sciences, Beijing, China and University of Chinese Academy of Sciences, Beijing, China (email: zhujiagang2015@ia.ac.cn; wei.zou@ia.ac.cn; zhuzheng2014@ia.ac.cn).

L. Xu, G. Huang are Horozon Robotics Co., Ltd, Beijing, China (email: liang.xu@horizon.ai; guan.huang@horizon.ai)

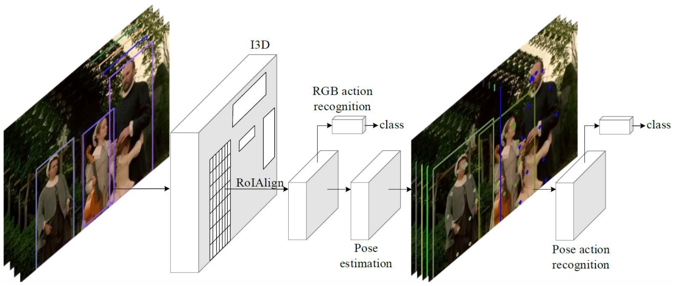

Fig. 2. Action Machine. First, the videos are fed into I3D for RGB-based action recognition. Then a 2D deconvolution head is added to the last convolutional layer of I3D for frame-wise pose estimation. Third, the estimated pose sequences are fed into a 2D CNN for pose-based action recognition. The proposed method is trained in a multi-task manner. Finally, the predictions of two heads for action recognition are fused by summation at inference time.

practical situations where the models have to handle the scenes largely different from the training sets. Action Machine shows significant improvement over the strong baseline I3D, by more than **7-10%** in accuracy, demonstrating the benefits of our person-centric modeling in generalizing across different datasets.

There are also previous works which are related to ours. As a representation typically designed for human pose, PoTion [15] is complementary to standard appearance and motion streams. Chained multi-stream network [16] unifies three sources: RGB images, optical flow and body part mask for action recognition and detection. In [17], Soft-argmax is extended to regress 2D and 3D pose directly, leading to the end-to-end training of pose estimation and action recognition. Different from the above three works, Action Machine is based on I3D, which is easy to train because of transferring pre-trained weights from 2D CNN and does not need the costly optical flow maps compared to two-stream ConvNet [10]. The pose estimation method we use is detection-based, detecting keypoint by regressing heatmap. It can get more accurate pose than the regression-based pose estimation in [17] in our experiments (about 5 AP). Moreover, our method can be applied to multi-person cases because of using RoIAlign [18].

II. Action Machine

Action Machine (Fig. 2) is a person-centric approach for action recognition in videos. It has several key elements: person detection, RGB-based action recognition, pose estimation and pose-based recognition. The details are described next.

RGB-based action recognition. We use the I3D with ResNet-50 [19] as backbone. In order to estimate the pose of each frame, we remove the temporal max pooling after the first stage of I3D. The output feature of the backbone is fed into RoIAlign [18] layer to obtain a tensor with size of 2048×8×7×7, used both by RGB-based action recognition and pose estimation. As shown in Fig. 3, global average pooling is performed after the last convolutional layer of I3D to get a 2048-d feature P_{rgb}.

Consider a dataset of N videos with n categories $\{(X_i, y_i)\}_{i=1}^{N}$, where $y_i \in \{1, \dots, n\}$ is the label. Formally, the prediction can be obtained directly

$$Y_{rgb} = \varphi(W_c P_{rgb} + b_c), \tag{1}$$

where φ is the softmax operation, $Y_{rgb} \in \mathbb{R}^n$. W_c and b_c are the parameters of the fully connected layer. In the training stage, combining with cross-entropy loss, the final loss function is

$$L_r = -\sum_{i=1}^{N} \log(Y_{rgb}(y_i)), \tag{2}$$

where $Y_{rgb}(y_i)$ is the value of the y_i-th dimension of Y_{rgb}.

Pose estimation. Given the output features of I3D, the pose estimation is performed on each time step. Inspired from Mask R-CNN [18], a 2D deconvolution head is added to the last convolutional layer of I3D, as shown in Fig. 3. By default, two deconvolutional layers with batch normalization [20] and ReLU activation [21] are used. Each layer has 256 filters with 4×4 kernel and the stride is 2. Following [22], a 1×1 convolutional layer is added at last to generate predicted heatmaps for all K keypoints (one channel per keypoint) and offsets (two channels per keypoint for the x and y-directions) for a total of $3K$ output channels, where $K = 17$ is the number of keypoints.

Given the image crop, let $f_k(x_i) = 1$ if the k-th keypoint is located at position x_i and 0 otherwise. Here $k \in 1, ..., K$ indexes the keypoint type and $i \in 1, ..., Q$ indexes the pixel locations on the image crop grid. For each position x_i and each keypoint k, we compute the probability $h_k(x_i) = 1$ if $||x_i - l_k|| \leq M$, which means the point x_i is within a disk of radius M from the location l_k of the k-th keypoint. A typical value of M is 25 in a 224×224 image. K such heatmaps are trained by solving a binary classification problem for each position and keypoint independently. For each position x_i and each keypoint k, we also predict the 2D offset vector $F_k(x_i) = l_k - x_i$ from the pixel to the corresponding keypoint. K such vector fields are trained by solving a 2D regression problem for each position and keypoint independently.

The output of the heatmap branch yields the heatmap probabilities $h_k(x_i)$ for each position x_i and each keypoint k. The training target for the heatmap branch $\overline{h}_k(x_i)$ is a map of zeros and ones, with $\overline{h}_k(x_i) = 1$ if $||x_i - l_k|| \leq M$ and 0 otherwise. The corresponding loss function $L_h(\theta)$ is the sum of smooth L_1 loss for each position and keypoint independently

$$L_h(\theta) = \frac{1}{K} \sum_{k=1}^{K} \sum_{i} R(h_k(x_i) - \overline{h}_k(x_i)), \tag{3}$$

where R is the smooth L_1 loss.

For training the offset regression branch, the differences between the predicted and ground truth offsets are penalized by smooth L_1 loss. The offset loss is only computed for positions x_i within a disk of radius M from each keypoint.

$$L_o(\theta) = \frac{1}{K} \sum_{k=1}^{K} \sum_{i:||l_k - x_i|| \leq M} R(F_k(x_i) - (l_k - x_i)), \tag{4}$$

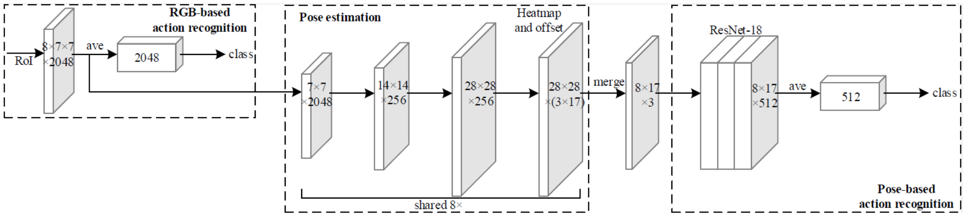

Fig. 3. We extend I3D by adding a branch for human pose estimation and a 2D CNN for pose-based action recognition. Numbers denote spatial resolution and channels. Arrows denote either conv, deconv, or fc layers as can be inferred from context (conv preserves spatial dimension while deconv increases it). The output conv of heatmap and offsetmap is 1×1, deconvs are 4×4 with stride 2. 'res5' denotes the fifth stage of I3D with ResNet-50. '8×' denotes the shared operations of 2D pose estimation on the temporal dimension. In the last of pose estimation head, '(3×17)' denotes the concatenation of 1-channel heatmap and 2-channel offset for 17 keypoints. For the input tensor of pose-based action recognition CNN, '×3' denotes the concatenation of 2-d coordinates and 1-channel confidence.

The final loss function for pose estimation has the form

$$L_p = L_h(\theta) + L_o(\theta), \tag{5}$$

At inference time, for the k-th keypoint, the argmax operation is performed on the k-th heatmap to yield the coarse location

$$x_k = \arg\max_{x_i} \, (h_k(x_i), i \in 1, ..., Q). \tag{6}$$

The accurate coordinate of the k-th keypoint is obtained by adding the corresponding offset $F_k(x_k)$ to x_k.

Pose-based action recognition. The coordinates of 2D pose can be transformed into a tensor of a size $2{\times}T{\times}K$ [17], where T denotes the number of input frames. An extra confidence channel, which is obtained by max pooling over the heatmap and passed to the ReLU activation, is added for each predicted joint to get a $3{\times}T{\times}K$ tensor. Then the tensor is fed into a modified ResNet-18 [19] for pose-based action recognition, as shown in Fig. 3. Due to the low spatial dimension of the input pose sequences, all the pooling operations are removed and all the stride 2 operations in the convolutional layers are replaced with 1. Global average pooling is performed after the last convolutional layer of ResNet-18 to get a 512-d feature. The prediction of pose stream $Y_{paction}$ is optimized with cross-entropy loss

$$L_{paction} = -\sum_{i=1}^{N} \log(Y_{paction}(y_i)). \tag{7}$$

Multi-task training. Action Machine has three tasks: RGB-based action recognition, pose estimation and pose-based action recognition. They are jointly optimized by the following loss function:

$$\begin{aligned} L &= L_r + \lambda_1 L_p + \lambda_2 L_{paction} \\ &= L_r + \lambda_1 (L_h(\theta) + L_o(\theta)) + \lambda_2 L_{paction} \end{aligned} \tag{8}$$

where λ_1 and λ_2 are the loss weights of pose estimation and pose-based action recognition respectively. When jointly training pose estimation with action recognition, λ_1 is set to 0.5 to not influence the main task. Because the gradients of pose-based action recognition don't back-propagate into the pose estimation head, λ_2 is set to 1.0. They are determined by cross-validation. Since the pose features cannot be used in a fully differentiable way, the pose-based action recognition task is trained sequentially.

TABLE I
EXPERIMENTAL SETUP.

Clip length	sampling stride	GPUs	Clips per GPU	optimizer
8	8	2	4	SGD
base lr	schedule	epochs	test crops	evaluation metric
0.01	[42, 68]	85	three spatial crops, 10 times	top-1 accuracy

TABLE II
PERFORMANCE ON NTU RGB+D, ACCURACY(%).

	Pose	RGB	xview	xsub
Lie Group [26]	✓	-	52.8	50.1
H-RNN [27]	✓	-	64.0	59.1
Deep LSTM [14]	✓	-	67.3	60.7
PA-LSTM [14]	✓	-	70.3	62.9
ST-LSTM+TS [28]	✓	-	77.7	69.2
Temporal Conv [29]	✓	-	83.1	74.3
VA-LSTM [30]	✓	-	87.6	79.4
ST-GCN [31]	✓	-	88.3	81.5
SR-TSL [32]	✓	-	92.4	84.8
Chained [16]	✓	-	-	80.8
2D-3D-Softargmax [17]	-	✓	-	85.5
Glimpse Clouds [25]	-	✓	93.2	86.6
PoseMap [24]	✓	✓	95.2	91.7
Action Machine (Ours)	-	✓	**97.2**	**94.3**

Fusion of RGB and pose-based action recognition. In order to combine the strengths of predictions from RGB images and poses, the predicted probabilities of two heads are fused by summation during inference. Other sophisticated fusion methods (e.g., feature concatenation) can also be tried, which is not the focus of this paper.

III. EXPERIMENTS

A. Datasets

The proposed method has been evaluated on video action datasets: NTU RGB+D [14], Northwestern-UCLA Multiview Action 3D (N-UCLA) [13], MSR Daily Activity3D (MSR Daily) [23]. We follow the experimental setup in Table I.

B. Experiments: Action recognition in videos

1) Comparison with state-of-the-art: In this section, Action Machine is compared with other approaches on NTU RGB+D [14], N-UCLA [13]. Results are shown in Table II, III, where ✓ denotes that the corresponding modality is used as the input of model in *testing*. On NTU RGB+D [14], Action Machine outperforms previous state-of-the-art PoseMap [24] by 2 and 2.6 points in top-1 accuracy on cross-view and cross-subject respectively. On N-UCLA [13], compared to Glimpse Clouds [25], Action Machine has a accuracy gain of 4.7 points in average top-1 accuracy on cross-view.

2) Ablation study: There are four basic configurations, detailed next:

TABLE III
PERFORMANCE ON N-UCLA, ACCURACY(%).

	Pose	RGB	xview1	xview2	xview3	Avg
Lie Group [26]	✓	-	-	-	-	74.2
H-RNN [27]	✓	-	-	-	-	78.5
Enhanced viz. [33]	✓	-	-	-	-	86.1
Ensemble TS-LSTM [34]	✓	-	-	-	-	89.2
Glimpse Clouds [25]	-	✓	83.4	89.5	90.1	87.6
Action Machine (Ours)	-	✓	**89.6**	90	94.3	91.3
Action Machine (Ours, NTU pre-training)	-	✓	88.3	**92.2**	**96.5**	**92.3**

TABLE IV
ABLATION STUDIES ON NTU RGB+D, ACCURACY(%). IN THE ROWS WHICH HAVE SLASH /, THE NUMBER ON THE LEFT OF SLASH IS THE ACCURACY OF POSE-BASED ACTION RECOGNITION, THE RIGHT IS THE ACCURACY OF FUSION OF RGB AND POSE RESULTS.

	xview	xsub	xview-s	xsub-s
RGBAction random crop	97.2	92.3	94.3	61.2
RGBAction person crop	**97.7**	93.2	94.5	67.9
KPS RGBAction	97.3	93.8	95.0	71.2
KPS PoseAction RGBAction (ResNet-18)	90.1/97.1	84.9/94.1	87.8/95.9	62.9/72.7
KPS PoseAction RGBAction (ResNet-50)	91.3/97.2	85.5/**94.3**	89.9/**96.1**	66.0/**73.5**

RGBAction random crop. The baseline I3D model takes as inputs the random crops of videos and performs action recognition using RGB feature.

RGBAction person crop. The I3D model uses RoIAlign to obtain person features and performs action recognition using RGB feature.

KPS RGBAction. The I3D model uses RoIAlign to obtain person features, performs action recognition using RGB feature, and adds a head for pose estimation.

KPS PoseAction RGBAction. The I3D model uses RoIAlign to obtain person features, adds a head for pose estimation, and performs action recognition using RGB and pose feature. The model trained from **KPS RGBAction** is used as the pre-trained model. We fix it and only train the ResNet-18 or ResNet-50 for pose-based action recognition. We report the results of pose-based action recognition and the sum fusion of predictions from RGB images and poses.

Results and Analysis. As shown in Table IV, on the cross-subject of NTU RGB+D, person cropping can improve the model accuracy by 0.9 points over random crop. Our full model outperforms the baseline **RGBAction random crop** by 2 points. Due to the high accuracy of baseline, the improvement on the cross-view is not obvious. Similar gain potential can also be observed on the small subsets of NTU RGB+D (xview-s, xsub-s), which are originally used for the fast training and testing in our implementation.

As shown in Table V, on N-UCLA, Action Machine outperforms the baseline I3D by a large margin. Specifically, **RGBAction person crop** with the person cropping technique can improve the accuracy by 1.6 and 4.3 points on *xview1* and *xview3* over the baseline **RGBAction random crop** respectively. Person cropping does not bring accuracy gain on *xview2*, because the test crops of front view images on this dataset are close to that cropped by person boxes. Jointly training pose estimation and RGB-based action recognition, i.e., **KPS RGBAction**, can improve about 3 to 7 points. Overall, using ResNet-18, our final model exceeds the baseline by 7.2 points. By using a stronger backbone, i.e., ResNet-50 for pose-based action recognition and NTU RGB+D pre-training, the accuracies of our models, either solely by poses or the fusion of RGB images and poses, are further improved.

TABLE V
ABLATION STUDIES ON N-UCLA, ACCURACY(%).

	xview1	xview2	xview3	Avg
RGBAction random crop	81.6	82.4	86.3	83.4
RGBAction person crop	83.2	82.4	90.6	85.4
KPS RGBAction	86.3	90	94.9	90.4
KPS PoseAction RGBAction (ResNet-18)	79.7/87.5	81/90.4	87.5/94.1	82.7/90.6
KPS PoseAction RGBAction (ResNet-50)	84.2/**89.6**	81.8/90	88.4/94.3	84.8/91.3
KPS PoseAction RGBAction (ResNet-18, NTU pre-training)	85.5/88.6	88.0/91.6	93.2/**96.5**	88.9/92.2
KPS PoseAction RGBAction (ResNet-50, NTU pre-training)	83.8/88.3	87.6/**92.2**	93.2/**96.5**	88.2/**92.3**

TABLE VI
CROSS-DATASET TESTING ON N-UCLA AND MSR DAILY, ACCURACY(%).

	N-UCLA	MSR Daily
RGBAction random crop	70.0	70.8
RGBAction person crop	70.0	78.4
KPS RGBAction	76.4	79.7
KPS PoseAction RGBAction (ResNet-18)	68.8/76.4	58.2/79.7
KPS PoseAction RGBAction (ResNet-50)	69.2/**77.3**	63.2/**81.0**

Cross-dataset recognition task. This task is designed to imitate the challenge that the models have to handle the unseen scenes when being deployed. The models are trained and tested on different datasets. Specifically, we train our models on NTU RGB+D cross-subject and test them on the test sets of the smaller datasets, i.e., N-UCLA, MSR Daily respectively. We report performance on the shared categories of these datasets. Because of the different sources of videos, the scene contexts and objects in training dataset are largely different from the testing dataset. In this case, a model without capturing human body motion will behave worse than that learns to focus on. Results are shown in Table VI. It is clearly observed that our proposed person-centric modeling techniques including: person cropping, multi-task training of action recognition and pose estimation, the fusion of predictions from RGB images and poses can help to improve the performance of baseline model **RGBAction random crop** on different datasets. Our method shows massive improvement over the baseline I3D, by more than **7-10%** in accuracy. This again indicates that our method really learns to focus on human body movements instead of overfitting the scenes and objects of specific datasets. Though some existing methods based on RGB images may have high performance on some datasets, they can easily be distracted by the non-human stuff when facing new videos with different context. In contrast, Action Machine is more generalizable and extendable.

IV. CONCLUSIONS

In this work, we propose Action Machine for human action recognition in videos. By using person bounding boxes and human poses, Action Machine achieves competitive performance compared with other approaches on video action datasets [13], [14], [23]. In experiments, we show that the models trained based on the full contents of videos tend to overfit the scenes and objects. Instead, Action Machine can generalize well across different datasets by explicitly capturing human body movements. We expect Action Machine be an effective framework in video surveillance and other application scenarios, which demand instance-level action analysis.

References

[1] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4724–4733.

[2] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[3] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[4] D. Purwanto, R. R. A. Pramono, Y. Chen, and W. Fang, "Three-stream network with bidirectional self-attention for action recognition in extreme low resolution videos," *IEEE Signal Process. Lett.*, vol. 26, no. 8, pp. 1187–1191, Aug 2019.

[5] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "Bsn: Boundary sensitive network for temporal action proposal generation," in *Eur. Conf. Comput. Vis.*, 2018, pp. 3–21.

[6] F. Murtaza, M. H. Yousaf, S. A. Velastin, and Y. Qian, "End-to-end temporal action detection using bag of discriminant snippets," *IEEE Signal Process. Lett.*, vol. 26, no. 2, pp. 272–276, Feb 2019.

[7] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6047–6056.

[8] K. Soomro, A. R. Zamir, M. Shah, K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, p. 2012.

[9] H. Kuehne, H. Jhuang, R. Stiefelhagen, and T. Serre, "Hmdb51: A large video database for human motion recognition," in *High Performance Computing in Science and Engineering*, 2013.

[10] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Adv. Neural Inf. Process. Syst.*, ser. NIPS'14, 2014, pp. 568–576.

[11] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.

[12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2016, pp. 2921–2929.

[13] J. Wang, X. Nie, Y. Xia, Y. Wu, and S. Zhu, "Cross-view action modeling, learning, and recognition," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2014, pp. 2649–2656.

[14] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2016, pp. 1010–1019.

[15] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "Potion: Pose motion representation for action recognition," in *IEEE Int. Conf. Comput. Vis.*, 2018.

[16] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2923–2932.

[17] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5137–5146.

[18] K. He, G. Gkioxari, P. Dollr, and R. Girshick, "Mask r-cnn," in *IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Int. Conf. Machine Learning*, 2015, pp. 448–456.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[22] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, July 2017, pp. 3711–3719.

[23] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1290–1297.

[24] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1159–1168.

[25] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, "Glimpse clouds: Human activity recognition from unstructured feature points," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 469–478.

[26] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 588–595.

[27] Yong Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1110–1118.

[28] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *ECCV*, 2016.

[29] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *CVPRW*, 2017.

[30] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2136–2145.

[31] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[32] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Eur. Conf. Comput. Vis.*, 2018, pp. 106–121.

[33] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, 2017.

[34] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in *IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1012–1020.