What Elements are Essential to Recognize Human Actions?

Yachun Li[†], Yong Liu[†], Chi Zhang[§]
[†]Institute of Cyber-Systems and Control, Zhejiang University, [§]Megvii Technology
liyachun@zju.edu.cn, yongliu@iipc.zju.edu.cn, zhangchi@megvii.com

Abstract

RGB image has been widely used for human action recognition. However, it could be redundant to include all information for human action depiction. We thus ask the following question: What elements are essential for human action recognition? To this end, we investigate several different human representations. These representations emphasize dissimilarly on elements (e.g. background context, actor appearance, and human shape). Systematic analysis enables us to find out essential elements as well as unnecessary contents for human action description. More specifically, our experimental results demonstrate the following: Firstly, both context-related elements and actor appearance are not vital for action recognition in most cases. But an accurate and consistent human representation is important. Secondly, essential human representation ensures better performance and cross-dataset transferability. Thirdly, fine-tuning works only when networks acquire essential elements from human representations. Fourthly, 3D reconstruction-related representation is beneficial for human action recognition tasks. Our study shows researchers need to reflect on more essential elements to depict human actions, and it is also instructive for practical human action recognition in real-world scenarios.

1. Introduction

Action recognition is one of the indispensable tasks for video understanding. Undoubtedly, two-stream inspired methods [31, 10, 44, 9] and 3D-convolution-based models [37, 3, 46, 38, 45] have already achieved remarkable performance in large scale video datasets (e.g. Kinetics [3], AVA [11] and UCF101 [32]). However, most research works use full RGB frame as it is. They do not take into account the various elements[1] in original RGB representations[2]. We argue that not all elements in RGB images are essential. Surplus unnecessary information might hinder the learning of essential features for human action recognition. For example, if we are only able to collect *drinking* in the consistent scene such as *classroom*, training the model may leads to the recognition of *the classroom* instead of action *drinking* itself. In this circumstance, model might not be able to recognize the *drinking* action when a person drinks in the restaurant.

Moreover, it is known that current model trained from one dataset easily fails when transferring to another dataset, especially to the dataset in a quite different scene. This issue has been discussed in several works [16, 25, 13, 40, 30]. Pose-related features are systematically studied for the purpose of action understanding [16, 25]. Hara *et al*. [13] attribute this failure to relatively small scale video datasets. [40] and [30] focus on investigation of temporal support. We conjuncture that this failure comes from poor capture of essential elements. This results in the overfitting to excessive elements, e.g. scene objects, person clothes, specific background scene, in source data (which is supported by our cross-dataset experiments).

Therefore, it is important to understand the essential elements to distinguish different human actions. A good human representation with essential elements helps network to capture discriminative features. And this representation is also beneficial to eliminate redundant information in original RGB images across different datasets. In this work, we provide in-depth analyses of what elements are essential for human action recognition. Concretely, we try to answer two questions:

Question 1: *Are all elements from RGB representations essential for human action recognition?*

In order to resolve this confusion, we studied several mask-style human representations. These representations are generated with dataset-provided skeleton annotations or segmentation mask (extracted from state-of-the-art pretrained models). Different representation has its own emphasis on action-related elements. Thus, they could help to seek out the most essential parts for human action recognition.

Our representations explicitly exclude several elements

[1] In this paper, elements refer to the contents included in original RGB images, e.g. background context, actor appearance, and joint movement.

[2] All representations in our work are image-based and have their own emphasis on different elements.

(e.g. background context, actor appearance, and human shape) from original RGB frames. Experiments show sustained improvement over original RGB representation in most cases. This suggests that redundant elements from RGB images indeed hinder the learning of critical features for human action recognition. Better representations, which filter out unnecessary elements and only retain essential elements, leads to superior performance.

Quantitative analysis indicates that scene-related elements are not necessary and may aggravate overfitting to simple background. Further cross-dataset transfer evaluation demonstrates human-focused representation is more generalized than original RGB.

Question 2: *Are there better representations to depict human actions?*

Skeleton-based action recognition is unaffected by complex foreground and background. There are an increasing number of researches in skeleton-based action recognition field [35, 48, 19, 43]. However, skeleton keeps only a small amount of information and its representative ability is limited. As our experiments show superior performance with image-based representations, we argue that image-related representations are more powerful to depict human actions. Therefore, we mainly focus on image-based human representations.

Experimental results have demonstrated that human-focused representations, derived from skeleton and segmentation, obtain better performance than original RGB. We regard these representations as mid-level depictions of human action. Recently, there emerge novel methods in the literature which could be utilized to represent human [1, 7, 12, 39]. Since mid-level features are not robust to large variance of background and human, 3D reconstruction method could provide a more accurate high-level human representation. Those reconstructed outputs have abundant semantic meanings behind recovered 3D human, including fine details of human actions.

Among those 3D reconstruction-related human representation, we investigated DensePose [12] as an ideal representation. DensePose reserves essential fine details of human action and dump unconcerned context elements from raw RGB frames. Evaluation shows that DensePose could be a reliable and practical form for human action depiction, and experiments demonstrate its superior performance over other representations.

In summary, we conclude our observations as:

- Context-related elements and actor appearance are not critical for human action recognition in most cases. It is important to construct an accurate and consistent human representation.
- Human representation with essential elements achieves better performance and cross-dataset transferability in human action recognition task. It is complementary to original RGB representation for general action recognition.
- Fine-tuning technique works only when deep models could learn from essential elements in human representations.
- 3D reconstruction-related representation is beneficial for human action recognition tasks.

Our findings are practical for human action recognition in real-world scenarios, as our essential human representations maximize the efficiency of cross-dataset transfer learning, even when limited data is available. Therefore, it could be instructive for applications in human action recognition field.

2. Related Work

2.1. Action Recognition

General Action Recognition We refer general action recognition as the tasks which do not explicitly require human's present in videos. Research works could be categorized into two mainstream architectures. One category is two-stream convolutional networks where two parallel networks process RGB still appearance and temporal motion respectively [31, 10, 44, 9, 3]. Simonyan and Zisserman [31] first utilize RGB frame and optical flow together for action recognition. Two streams are combined using late fusion. Fusion approach is further studied in [10]. Wang *et al.* [44] extended two-stream networks and presented Temporal Segment Networks (TSN), where full video is divided into segments and then these scores are aggregated through consensus function. Besides, spatiotemporal multiplier network [9] presents motion gating for the interaction of appearance and motion pathways.

Another dominating method is 3D-convolution-based model, where 3D convolution is used to incorporate temporal information. An early work C3D [37] verifies superior performance of 3D convolution over 2D features. Following work explores capacity of 3D convolution [3, 46, 38, 45]. Carreira *et al.* [3] proposed I3D model. 2D ConvNet is inflated to make 3D convolution. Wang *et al.* [45] built non-local block based on 3D model and achieved state-of-the-art performance in Kinetics [3] dataset. Other common dataset for general action recognition includes UCF101 [32] and HMDB51 [18].

Pose-based Action Recognition The pose-related method can be categorized into skeleton-based and image-based. Although skeleton may exclude some useful information, it is still a good form for human representation. These methods are able to obtain competitive results [28, 35, 48, 19, 43]. Tang *et al*. [35] introduce reinforcement learning for frame distillation and use graph-based representation for skeleton learning. Wang *et al*. [43] construct skeleton-related primitives (joints, edges, and surfaces) and use them for human action modeling.

On the other hand, pose features are also employed in image form [16, 5, 6, 47, 51, 8]. Chéron *et al*. [5] designed a new action descriptor based on human pose, where motion and appearance features of different body parts are combined for human action recognition. In addition to RGB appearance stream and optical flow motion stream, Zolfaghari *et al*. [51] incorporated pose stream. Chained connection is used to pass information among these streams. Pose motion (PoTion) forms a new stream and encodes pose features to recognize video action [6]. This pose representation is complementary to appearance and motion path, and improved performance is observed. Clearly, pose forms favorable features to differentiate human actions.

Foreground-related Action Recognition Emphasis on foreground could be beneficial for human action recognition since human action focuses on human [16, 33, 25, 22, 36, 34]. We could not completely separate foreground-related recognition methods from pose-related ones, because pose underlines foreground information to some extent. More weighting on foreground features shows performance promotion in human action recognition task [33]. Actionness map is used to detect action [22]. Its integration with pooling scheme shows accuracy improvement for action recognition.

The most related work to us is from Pishchulin *et al*.'s [25]. Their work attempted to find out ideal activity representations/features for recognition. They thus explored underlying factors that affect method performance and compared holistic methods with pose-based ones in human activity recognition. Experiments show pose-based features outperform holistic features in certain cases. In our work, we focus on human-centric actions and also utilize pose to build image-based representations. However, we aimed at analyzing essential elements for human action recognition. For this purpose, we elaborately experimented with potential elements (e.g. background context and actor appearance) which might influence recognition performance. Consequently, distinct human representations are investigated for better action perception. We show our human-related representations are more essential and achieve superior results over original RGB images.

2.2. Human Representations

Pose Estimation and Segmentation Both pose estimation [2, 24, 4] and segmentation [20, 14, 50] tasks could benefit our work for more generalized human representations. [50] introduced the ADE20K dataset and achieved great segmentation performance. Since some datasets already provide skeleton annotations, we turn to state-of-the-art model [50] for segmentation generation.

3D Human Reconstruction SMPL enables parametric representations of human body shape [21]. Following work employs SMPL model for 3D reconstruction from image [1, 17, 12, 23]. Güler *et al*. introduced a new dense representation of human pose, which is able to model 3D surface of human body [12]. This surface-based modeling is an ideal representation of human since it is weakly related to context elements like background and person appearance. It has been used for pose transfer of a person to different views [23]. The warping module in [23] is capable of recovering full-size high-quality texture from incomplete DensePose prediction. This technique could be used to generate synthetic data for various tasks.

More research studied on various human representations models [7, 39]. These 3D reconstructions could be good inspiration for novel human representation in action recognition task. While some researches utilize 3D models for data generation [41, 26], we could, in turn, employ these models for human representations.

3. Approaches

This section describes how we utilized human-related mid-level and high-level features. Different combinations of these features and RGB image have their own emphasis on different elements.

3.1. Skeleton Heatmap

Instead of designing hand-engineered features from skeleton joints data, we utilize the strong representative ability of image-based representations. Thus, we generate heatmap from skeleton annotations with Gaussian kernel. Given K joints of frame t, their corresponding Gaussian heatmaps of each joint are $\{h_{t1}, h_{t2}, \cdots, h_{tK}\}$. Our *skeleton heatmap* is expressed as

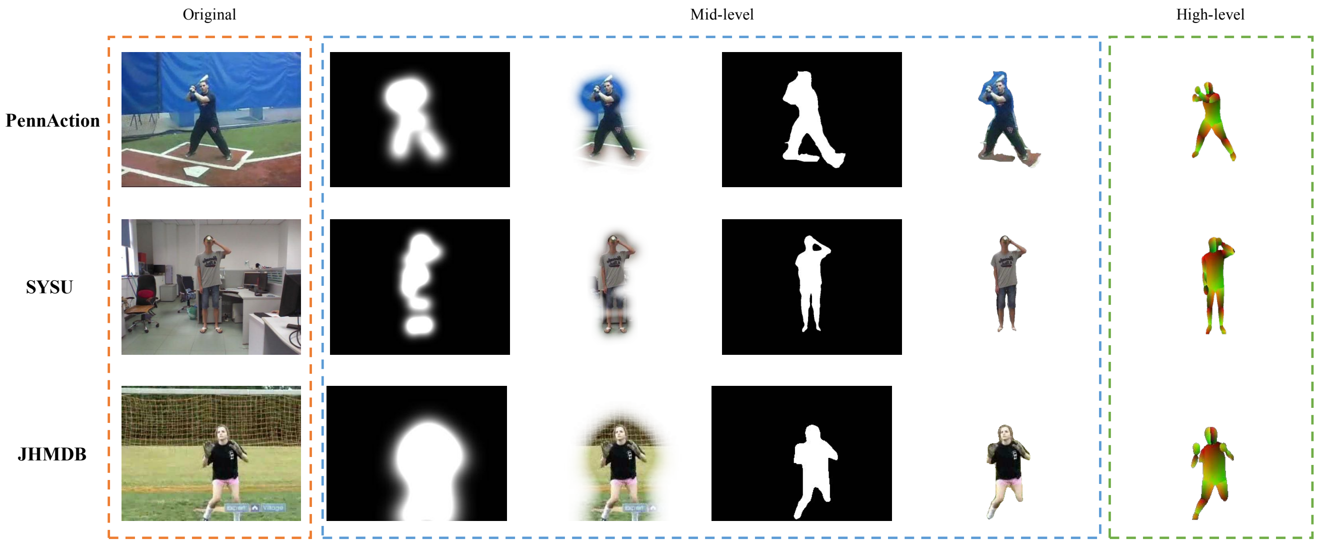
$$H_t = \max(\sum_{k=1}^{K} h_{tk}, 1) \tag{1}$$

The *skeleton heatmap* itself could be a suitable representation of human actions, which eliminates background context, actor appearance as well as human shape from image. It reserves more information about joint movements and provides a consistent representation with smooth boundaries.

Besides, we would also combine RGB color information C_t with mask-style heatmap H_t to build another skeleton-related image representation, which is

$$HC_t = H_t \odot C_t \tag{2}$$

where $\odot$ denotes element-wise multiplication. This multiplication reintroduces actor appearance.



(a) Original RGB (b) Skeleton Heatmap (c) RGB⊙Skeleton Heatmap (d) Segmentation Mask (e) RGB⊙Segmentation Mask (f) DensePose

Figure 1. Overview of human representations for human action recognition.

3.2. Segmentation Mask

Semantic segmentation outputs pixel-level predictions for each input image. [50] proposed to parse images into 150 object and stuff classes with their Cascade Segmentation Module. We utilized their pre-trained models to extract person segmentation. Our *segmentation mask* M_t is a binary mask where 1 indicates human and 0 represents other classes. *Segmentation mask* M_t has good description of human shape, and excludes background context and actor appearance elements as *skeleton heatmap* H_t does.

Similarly, the integration of RGB color map C_t and *segmentation mask* M_t is also an alternative option to depict human action:

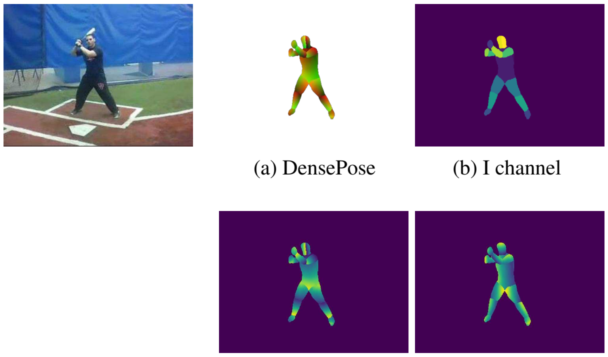
$$MC_t = M_t \odot C_t \tag{3}$$

3.3. DensePose

In addition to aforementioned mid-level human representations, we investigated a novel human representation *DensePose* [12] for the purpose of human action recognition. The 3D surface-based representation depicts dense human poses, and it can be directly derived from RGB image.

DensePose generates UV map for 24 body parts from a single image, and it finally constructs a 3-channel image result. These 3 channels consist of one patch number channel and two UV coordinates channels (see Fig. 2). UV maps present rich information about 3D human pose/shape and hence are beneficial to human-centric action recognition.

With *DensePose* representation, more smooth and consistent body shape is reserved. It also eliminates background and appearance elements as a human action representation.



(a) DensePose (b) I channel (c) U channel (d) V channel

Figure 2. Example of DensePose representation (*baseball swing*) on PennAction dataset. (a) Combined color image, (b)(c)(d) Corresponding IUV for 3-channel color image, where I represents body parts and U/V for UV coordinates.

4. Experiments

4.1. Datasets and Settings

Penn Action Dataset (PennAction) [49]. PennAction was collected from various online videos, consisting of 15 action classes. It contains 2326 videos and provides 13 keypoints annotation along with RGB clips. The dataset is roughly divided 50/50 for training/testing.

SYSU 3D HOI Dataset (SYSU) [15]. The Kinect captured RGB-D dataset focuses on human-object interactions. There are 480 video clips and each sample is provided with RGB frames, depth sequence and skeleton data (20 joints). 40 subjects performed 12 different action activities in the limited lab environment.

Joint-HMDB (JHMDB) [16]. This dataset is a subset of HMDB51 [18] and provides human joints annotation. It has overall 928 clips with 31838 frames, providing pose, segmentation and dense optical flow ground truth for human actions.

NTU RGB+D (NTU) [28]. NTU is one of the large scale datasets containing RGB videos as well as corresponding skeleton (25 keypoints). It consists of 56880 RGB+D video sequences with more than 4 million frames. 60 action classes were performed by 40 distinct subjects and 3 cameras were used to record the action at the same time. Two splitting configurations, cross-subject (CS) and cross-view (CV), are proposed for evaluation. All the samples are captured with Kinect cameras in indoor controlled scene.

UCF101 [32]. UCF101 is the most widely used dataset for action recognition, consisting of 101 action classes, with over 13k clips and 27 hours of video data. It is based on videos in the wild.

Implementation. We choose non-local neural networks (non-local) [45] for our experiments and train models for up to 200 epochs. Input size is set to 256 × 256 and the kernel size of global average pooling layer is adjusted accordingly. Images are sampled every 8 frames to form an 8-image video clip. These clips are latter fed to non-local model. We use the Stochastic Gradient Descent (SGD) optimizer with different learning rate and batch size for different datasets. For testing, single clip evaluation is reported and we do not conduct data augmentation for better validation performance. All the experiments are based on PyTorch framework.

4.2. Analysis

To better understand what elements in a single image are more important for human-focused action recognition, we experiment with human representations (see Section 3) on different datasets.

Our human representations emphasize dissimilarly on foreground or background related components. We would like to know: Does the background context, actor appearance, human shape and texture, joints positions or other elements matter more in our human action recognition task?

What elements are essential for human action recognition and what is better representation?

We first test our baseline model non-local neural networks on large scale action dataset NTU, in order to verify the representation ability of image-based deep models on human-related action recognition tasks. We tested on harder splitting setting, cross-subject (CS), and reported evaluation performance.

As shown in Table 1, our non-local baseline achieves state-of-the-art performance on NTU dataset, reaching 90.4% accuracy. Non-local gives high-accurate results compared to either skeleton-based or RGB image-based methods. We, therefore, consider our baseline as a powerful model for human-centric action recognition.

Method	**Input**	**CS**
VA-LSTM [48]	skeleton	79.4
DPRL [35]	skeleton	83.5
SR-TSL [29]	skeleton	84.8
DA-Net [42]	RGB	88.1
Non-local (our baseline)	RGB	90.4

Table 1. Comparisons of accuracy(%) on NTU RGB+D dataset with cross-subject setting.

Above fundamental experiment shows top performance of our baseline non-local with original RGB inputs. As we aim at finding out the essential elements for human action depiction, we then carried out experiments with original RGB representation and 5 proposed human presentations on three datasets (PennAction, SYSU, and JHMDB). These datasets are either from controlled lab environments or from challenging complex scenes.

To clarify, we termed overall 6 image-based representations as *Original RGB* (*RGB*), *Skeleton Heatmap* (*SkHeatmap*), *RGB⊙Skeleton Heatmap* (*RGB⊙SkHeatmap*), *Segmentation Mask* (*SegMask*), *RGB⊙Segmentation Mask* (*RGB⊙SegMask*) and *DensePose* (*DPose*). For *skeleton heatmap* representation, we use keypoint annotations from datasets. For *segmentation mask* as well as *DensePose* representation, we generate them in advance using pre-trained model. Fig. 1 gives an overview of our human representations in 3 different datasets. We could see from those samples that *SkHeatmap* gives a consistent human action representation with smooth boundaries. As we are using ground truth skeleton annotation for heatmap generation, it is more accurate and not influenced by imperfect predictions of assistant models (i.e. semantic segmentation model and DensePose model). Even though both segmentation mask and DensePose are precise human descriptions in most cases, segmentation quality falls when action is performed in complex environment, and DensePose representation is relatively accurate and more human-related.

PennAction: As described in Section 4.1, PennAction is obtained from online videos. It includes pose-related action (e.g. baseball pitch, bench press, jump rope and sit up) both in indoor and outdoor complex scene, which is challenging with large variation in appearance and motion.

We show our evaluation results in Table 2a. All of our proposed human representations achieve more accurate performance than original RGB frame representations on PennAction dataset. Our mask-form human representations (i.e. *SkHeatmap*, *RGB⊙SkHeatmap*, *SegMask* and *RGB⊙SegMask*) explicitly highlight human-related fore-

Representation	Acc.
RGB	88.48
SkHeatmap	94.10
RGB⊙SkHeatmap	90.64
SegMask	92.51
RGB⊙SegMask	90.82
DPose	**94.29**

(a) PennAction performance

Representation	Acc.
RGB	80.00
SkHeatmap	70.42
RGB⊙SkHeatmap	72.75
SegMask	70.00
RGB⊙SegMask	75.42
DPose	**82.08**

(b) SYSU performance (setting-2)

Representation	Acc.
RGB	45.52
SkHeatmap	56.34
RGB⊙SkHeatmap	47.02
SegMask	46.64
RGB⊙SegMask	42.54
DPose	**57.84**
PMask	**64.18**
RGB⊙PMask	54.85

(c) JHMDB performance (split1)

Table 2. Human representations evaluation results on PennAction, SYSU and JHMDB. Best accuracy is highlighted in bold.

ground elements, which implies emphasis on human is favorable in this scenario.

The generated skeleton heatmap gives a clear and distinct boundary of actors' behavior without full inclusion of actor, while segmentation mask gives a complete depiction of action range and human shape. Both *RGB⊙SkHeatmap* and *RGB⊙SegMask* introduce richer information about actor's appearance, but they also bring some redundant elements such as clothes color. We found that *SkHeatmap* obtains the best performance 94.10% among these 4 mask-like mid-level human representations. This indicates the articulated action description is essential and scene-related elements are unnecessary for human action recognition. Besides, a relatively consistent human representation with smooth boundaries is helpful to our task.

Our model recognizes human actions well without abundant RGB stimulus. It might be due to the large variance among different action classes. In this case, changing of joint positions or human shape already provides enough information to distinguish actions. Thus, extra RGB elements are redundant. PennAction videos have various and complex context scene. With the help of mid-level auxiliary features, we are able to explicitly filter out unnecessary background pixels and RGB appearance elements. It ensures more effective learning from essential elements to differentiate action classes.

We got best action recognition performance 94.29% with *DPose*, which includes human-related fine details yet eliminates redundant scene-related elements. Like *SegMask* representation, *DPose* includes more subtle changes in human shape and body parts. Moreover, *DPose* is more robust to complex scenes, and it gives relatively consistent predictions across different datasets as *SkHeatmap* does. Top performance of *DPose* suggests that context-related elements are not necessary for human action recognition.

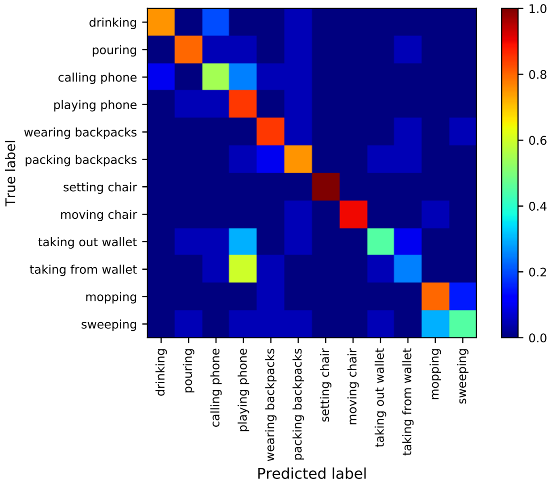


Figure 3. The confusion matrix of *SegMask* representation evaluation results on SYSU dataset.

SYSU: SYSU is captured in indoor controlled environment, which means it contains limited and specific background. Due to the relatively small size of this dataset, our 3D model might more easily overfit to unconcerned features. Besides, we observed SYSU actions have a relatively small range of motion. Our mask-form representations might be ill-posed since they are not adaptive to subtle changes within mask.

As reported in Table 2b, *SkHeatmap*, *RGB⊙SkHeatmap*, *SegMask* and *RGB⊙SegMask* give inferior recognition accuracy compared to original RGB inputs, which is in line with our concern. The confusion matrix of *SegMask* in Fig 3 shows that it could not distinguish *taking from wallet vs. playing phone* and *mopping vs. sweeping* well.

We could find generated segmentation mask gives more accurate human shape representation in SYSU controlled scene. In this case, similar performance of *SkHeatmap* and *SegMask* implies accurate human boundary is important factor for action recognition (inaccurate *SegMask* results in inferior performance to *SkHeatmap* in PennAction), which is consistent to the observation from [27] that action recognition cares about boundaries. Additional RGB appearance could help to recognize actions with similar human shape and joint movement, e.g. mopping and sweeping (see Fig 4).

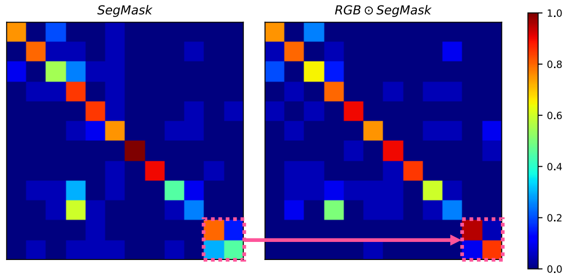


Figure 4. Change of confusion matrix after adding additional RGB appearance to *SegMask* representation on SYSU dataset. Clear improvement in certain actions indicates that RGB could be helpful only when *SegMask* is not enough to distinguish these certain human actions, e.g. mopping and sweeping as shown in box.

It hence leads to better performance of *RGB⊙SkHeatmap* (from 70.42% to 72.75%) and *RGB⊙SegMak* (from 70.00% to 75.42%).

DPose still obtains best recognition results of 82.08%. This shows our high-level *DPose* representation includes more essential elements for human action recognition, and it is still ideal to describe human actions in controlled scene.

JHMDB: In addition to skeleton annotation, JHMDB provides ground truth segmentation annotation derived from puppet model. Thus, we also experimented with its annotated segmentation. Corresponding human representations are named *Puppet Mask (PMask)* and *RGB⊙Puppet Mask (RGB⊙PMask)*.

Evaluation results are shown in Table 2c. Two RGB-combined mask representations (*RGB⊙SkHeatmap* and *RGB⊙SegMak*) get similar results as original RGB inputs do. As their performance is worse than corresponding human representations without RGB appearance, this again confirms the dispensability of RGB appearance elements.

Varying pixel-level predictions on the boundaries hinder the performance improvement of *SegMask*. On the contrary, accurate human representation from *SkHeatmap* shows better performance 56.34%. Not surprisingly, *DPose* still gives best results (57.84%) among these 5 human representations and further demonstrates its ability for human action representation.

We observe improved performance with ground truth segmentation mask. *PMask* achieves 64.18% accuracy. Combination with RGB presents decreased result of 54.85%. In the one hand, it indicates a consistent mask representation with accurate boundaries helps human action recognition task. In the other hand, we could see RGB appearance is not informative when we have "perfect" descriptions of human shape.

Dedicated experiments on three datasets reveal that background context elements are not critical for human action recognition. Neither is actor appearance. A relatively accurate and consistent human representation could be helpful for action recognition. Therefore, we could seek a more effective way to represent human instead of using raw RGB inputs, especially when limited data is available. This could help neural networks to concentrate more on essential elements and discard unnecessary elements. A proper representation ensures capture of key features for human action recognition task.

Representations	**scratch**	**softmax**		**finetune**	
		Acc.	Diff.	Acc.	Diff.
RGB	80.00	42.92	↓ 37.08	78.75	↓ 1.25
SegMask	70.00	55.42	↓ 14.58	73.75	↑ 3.75
RGB⊙SegMask	75.42	50.83	↓ 24.59	78.75	↑ 3.33
DPose	82.08	68.33	↓ 13.75	85.83	↑ 3.75

Table 3. Cross-dataset transfer learning results from PennAction to SYSU dataset.

Is the representation general to datasets from different sources?

To further analyze representative capacity of different human representation, we conducted transfer learning experiments. This also benefits in-depth understanding of essential elements for human action recognition.

We chose PennAction and SYSU for this part. These 2 datasets are different in captured scene and action categories. We argue that good human representation should be able to capture the most essential elements. Thus, it could well represent human actions across the datasets. For this purpose, we transfer the models pre-trained on PennAction to SYSU dataset. Experiments are based on two transfer fashion: *softmax* and *finetune*. *Softmax* transfer learning means we only train last softmax layer to fit different action classes of two datasets. *Finetune* indicates all layers are set to trainable during transfer experiments.

Since different datasets provide skeleton annotations of different joint number, *SkHeatmap*-related representations are not appropriate in our transfer learning experiments. So we perform transfer analysis with *RGB*, *SegMask*, *RGB⊙SegMask* and *DPose* human representations.

Experimental results are in Table 3. There are various degree of performance dropping on *softmax* transfer setting. Compared with *RGB* of 37.08% and *RGB⊙SegMask* of 24.59% decreasing, *SegMask* and *DPose* show less reduction of 14.58% and 13.75%. *RGB* gives worst performance (42.92%). *DPose* obtains best performance 68.33% with least accuracy dropping 13.75%. It is an evidence that *RGB* easily overfits to unessential elements in specific dataset. On the contrary, including more essential elements ensures learning of action-related features for human action recognition.

In *finetune* transfer fashion, we observe improvement on *SegMask*, *RGB⊙SegMask* and *DPose*. It shows distillation of essential information can help network to learn hu-

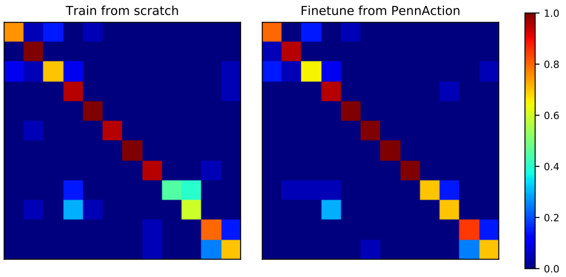


Figure 5. Confusion matrices of *DPose* representation with *train from scratch* and with *finetune from PennAction* on SYSU dataset.

Representations	Acc.
RGB	83.56
SegMask	58.18
RGB$\odot$*SegMask*	65.87
DPose	65.69
RGB + *SegMask*	84.46
RGB + *RGB*$\odot$*SegMask*	83.40
RGB + *DPose*	**86.10**

Table 4. Results of our human representations on UCF101 dataset (split1). Note that these representations are not designed for general action recognition but still obtain moderate performance on UCF101 dataset.

man actions more effectively. And these representations enable learning of more critical action-related features, which are general even across different datasets. As illustrated in Fig 5, regular and consistent improvement of *DPose* demonstrates its potential as an ideal human representation with well-defined surface and fine details of human actions.

We observe inferior fine-tuning result of *RGB* (78.75%) compared with SYSU *RGB* trained from scratch (80.00%). It is a bit opposite to our intuition. Generally, fine-tuning from pre-trained model achieves better performance than the one which is trained from scratch. For the fine-tuning of *RGB* from PennAction to SYSU, it again suggests deep model with raw RGB inputs focuses on unconcerned elements (such as background context and actor appearance) instead of human action itself. Therefore, we conclude that fine-tuning technique only works when deep models could learn from essential features in human representations.

Does the representation help original RGB model performance boosting?

Through previous analysis, we show that our human representations are able to catch more essential elements, and they achieve better performance in human-centric action recognition. We would like to investigate if our model learns complementary features against original RGB features in general action recognition scenario. Thus, we conducted experiments on UCF101.Similarly, we generate segmentation mask and DensePose of UCF101 videos using pre-trained model. The 4 distinct representations, *RGB*, *SegMask*, *RGB*$\odot$*SegMask* and *DPose*, alone are trained on UCF101. We reported results in Table 4.

Experiments show that our human representations achieve moderate performance on general action dataset UCF101. It indicates human pose-related features/representations could also be an important clue for general action recognition. In order to explore whether our human representations are complementary to original RGB representations, we combine them together via score average fusion (Table 4). Both *RGB* + *SegMask* and *RGB* + *DPose* obtain more accurate recognition of 84.46% and 86.10% respectively. This suggests that two representations *SegMask* and *DPose* help to catch complementary features to original RGB input.

On the contrary, we found average fusion of *RGB* and *RGB*$\odot$*SegMask* scores gives worse performance. Although *RGB*$\odot$*SegMask* alone achieves relatively high accuracy of 65.87%, fusion result implies that non-local does not learn extra action features except RGB appearance for those action classes. This is consistent with previous observation that RGB appearance is not essential and might even be cumbersome for human action recognition.

5. Conclusions

We experimentally show that current RGB image is not an effective representation for human action recognition. And it easily overfits to background context and actor appearance of specific dataset. On the contrary, our essential human representations are able to obtain better performance. It suggests these representations more efficiently catch essential elements and dump interfering information from RGB images.

Cross-dataset transfer evaluation implies that pretraining on a distinct dataset with pure RGB input is not a wise choice for action recognition tasks. Only representations with essential elements ensure the boosting performance of fine-tuned networks. And we show more reasonable cross-dataset transferability with our representations. Thus, this paper could be instructive for human action recognition in real-world scenarios. Furthermore, we present a new perspective to represent human as reconstruction results for action recognition. 3D reconstruction-related human representation DensePose achieves surprisingly well performance. This encourages researchers to rethink the form of essential human representation. We believe our work could help action recognition researchers to better understand intrinsic elements behind human actions.

Acknowledgments. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1302003 and by the Fundamental Research Funds for the Central Universities (2018QNA5012).

References

[1] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 2, 3

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CVPR*, 2017. 3

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1, 2

[4] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018. 3

[5] G. Chéron, I. Laptev, and C. Schmid. P-cnn: Pose-based cnn features for action recognition. In *ICCV*, 2015. 3

[6] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid. Potion: Pose motion representation for action recognition. In *CVPR*, 2018. 3

[7] E. Dibra, H. Jain, A. C. Öztireli, R. Ziegler, and M. H. Gross. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In *CVPR*, 2017. 2, 3

[8] W. Du, Y. Wang, and Y. Qiao. Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. In *ICCV*, 2017. 3

[9] C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spatiotemporal multiplier networks for video action recognition. In *CVPR*, 2017. 1, 2

[10] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 1, 2

[11] C. Gu, C. Sun, S. Vijayanarasimhan, C. Pantofaru, D. A. Ross, G. Toderici, Y. Li, S. Ricco, R. Sukthankar, C. Schmid, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 1

[12] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2, 3, 4

[13] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet. In *CVPR*, 2018. 1

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 3

[15] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *CVPR*, 2015. 4

[16] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, 2013. 1, 3, 5

[17] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3

[18] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 2, 5

[19] C. Li, Z. Cui, W. Zheng, C. Xu, and J. Yang. Spatio-temporal graph convolution for skeleton based action recognition. 2018. 2, 3

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 3

[21] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015. 3

[22] Y. Luo, L.-F. Cheong, and A. Tran. Actionness-assisted recognition of actions. In *ICCV*, 2015. 3

[23] N. Neverova, R. A. Güler, and I. Kokkinos. Dense pose transfer. In *ECCV*, 2018. 3

[24] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 3

[25] L. Pishchulin, M. Andriluka, and B. Schiele. Fine-grained activity recognition with holistic and pose based features. In *GCPR*, 2014. 1, 3

[26] A. Ranjan, J. Romero, and M. J. Black. Learning human optical flow. In *BMVC*, 2018. 3

[27] L. Sevilla-Lara, Y. Liao, F. Guney, V. Jampani, A. Geiger, and M. J. Black. On the integration of optical flow and action recognition. *GCPR*, 2018. 6

[28] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016. 3, 5

[29] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In *ECCV*, 2018. 5

[30] G. A. Sigurdsson, O. Russakovsky, and A. Gupta. What actions are needed for understanding human actions in videos? In *ICCV*, 2017. 1

[31] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1, 2

[32] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01*, 2012. 1, 2, 5

[33] W. Sultani and I. Saleemi. Human action recognition across datasets by foreground-weighted histogram decomposition. In *CVPR*, 2014. 3

[34] C. Sun, A. Shrivastava, C. Vondrick, K. Murphy, R. Sukthankar, and C. Schmid. Actor-centric relation network. In *ECCV*, 2018. 3

[35] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou. Deep progressive reinforcement learning for skeleton-based action recognition. In *CVPR*, 2018. 2, 3, 5

[36] A. Tran and L.-F. Cheong. Two-stream flow-guided convolutional attention networks for action recognition. *ICCV Workshop*, 2017. 3

[37] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 1, 2

[38] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 1, 2

[39] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *ECCV*, 2018. 2, 3

[40] G. Varol, I. Laptev, and C. Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517, 2018. 1

[41] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017. 3

[42] D. Wang, W. Ouyang, W. Li, and D. Xu. Dividing and aggregating network for multi-view action recognition. In *ECCV*, 2018. 5

[43] H. Wang and L. Wang. Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection. *IEEE Transactions on Image Processing*, 27(9):4382–4394, 2018. 2, 3

[44] 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 *ECCV*, 2016. 1, 2

[45] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, 2018. 1, 2, 5

[46] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 1, 2

[47] D. Zhang, G. Guo, D. Huang, and J. Han. Poseflow: A deep motion representation for understanding human behaviors in videos. In *CVPR*, 2018. 3

[48] 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 *ICCV*, 2017. 2, 3, 5

[49] W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013. 4

[50] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 3, 4

[51] 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 *ICCV*, 2017. 3