applied sciences

MDPI

Article

Body-Part-Aware and Multitask-Aware Single-Image-Based Action Recognition

Bhishan Bhandari, Geonu Lee and Jungchan Cho *

College of Information Technology, Gachon University, Seongnam 13120, Korea; bhandari@gc.gachon.ac.kr (B.B.); lkw3139@gc.gachon.ac.kr (G.L.)

* Correspondence: thinkai@gachon.ac.kr; Tel.: +82-31-750-5328

Received: 13 January 2020; Accepted: 20 February 2020; Published: 24 February 2020

Abstract: Action recognition is an application that, ideally, requires real-time results. We focus on single-image-based action recognition instead of video-based because of improved speed and lower cost of computation. However, a single image contains limited information, which makes single-image-based action recognition a difficult problem. To get an accurate representation of action classes, we propose three feature-stream-based shallow sub-networks (image-based, attention-image-based, and part-image-based feature networks) on the deep pose estimation network in a multitasking manner. Moreover, we design the multitask-aware loss function, so that the proposed method can be adaptively trained with heterogeneous datasets where only human pose annotations or action labels are included (instead of both pose and action information), which makes it easier to apply the proposed approach to new data on behavioral analysis on intelligent systems. In our extensive experiments, we showed that these streams represent complementary information and, hence, the fused representation is robust in distinguishing diverse fine-grained action classes. Unlike other methods, the human pose information was trained using heterogeneous datasets in a multitasking manner; nevertheless, it achieved 91.91% mean average precision on the Stanford 40 Actions Dataset. Moreover, we demonstrated the proposed method can be flexibly applied to multi-labels action recognition problem on the V-COCO Dataset.

Keywords: convolutional neural network; deep learning; action recognition; pose estimation; multitask learning

1. Introduction

The action recognition problem [1,2] can be solved using a video or a single image. However, video-based action recognition has a delay (required for receiving all video frames) and a large computational complexity, which makes it impractical for embedded devices with limited resources [3]. Moreover, image-based action recognition is the basis for applications such as video-based action recognition and visual question answering. Hence, we focus on action recognition from a single image.

To infer the human action, the human pose is an essential component [4–7] because a variety of human behaviors can be categorized based on the pose (such as standing and sitting). However, human pose information is not enough for the deep understanding of human actions because many actions have the same pose structure. For example, "Raising arms" can be further categorized into fine-grained classes such as "Brushing teeth" and "Waving hands". Therefore, action recognition can be considered as a fine-grained classification problem.

Many studies on fine-grained classification [8–10] have been focusing on local information, because the distinctive features are in the local area. In terms of action recognition, the foreground around the human joints can be the key local information for classifying fine-grained actions. Moreover, the most important local information is the objects that interact with the person. Hence, the primary focus should be on these areas. Based on the above analysis, we propose a body-part-aware and

attention-based action recognition method using the pose information, as depicted in Figure 1. It consists of three streams: (1) image-based action recognition; (2) attention-based action recognition; and (3) body-part-based action recognition. Moreover, the information that describes the human action should be considered together, which leads us to multitask learning for human pose estimation and action recognition.

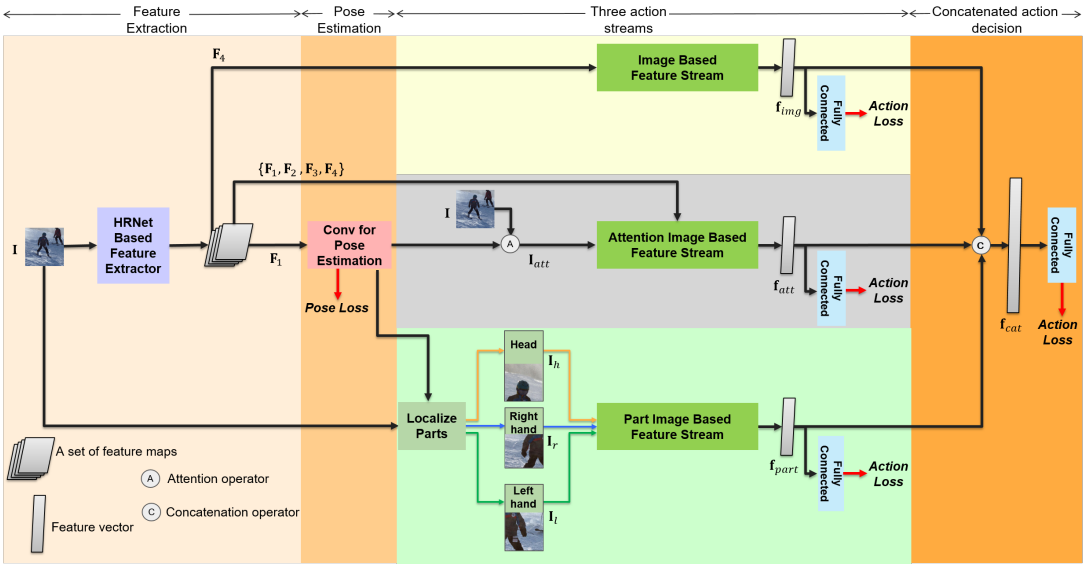

Figure 1. Proposed method. The proposed method is one end-to-end trainable convolutional neural network. The HRNet-based feature extractor is shared between pose estimation and action recognition task in a multitasking manner.

By definition, multitask learning is learning of M number of tasks, where all or some subset of the tasks are related, thus improving the generalized learning of models of these tasks by using knowledge of all or some of the M tasks [5,11,12]. Regardless of the advantages of multitask learning, the performance of individual tasks suffers when the tasks are learned simultaneously. It is similar to the problem of earlier and later classifiers discussed in [13]: an earlier classifier degrades the performance of the later classifier because early classifiers interfere with later classifiers. This performance drop can be mitigated with the use of dense connectivity, which connects each layer with all subsequent layers and avoids features optimized only by the early classifier. Recently, Sun et al. [14] proposed a pose estimation method with a structure based on dense connectivity. In our method, we use the High-Resolution Net (HRNet) [14] as a feature extractor, as depicted in Figure 1.

There is another problem in applying multitask learning to fine-grained action classification. Because deep learning has become the de-facto standard for studies on human behavior analysis, many data required to train models in a supervised manner. However, creating a dataset is difficult and time-consuming. Specifically, not enough datasets contain both human pose annotations and action labels to solve this fine-grained action recognition problem. Moreover, for various applications of intelligent systems, a low computational cost is necessary, and the system must be able to reuse an existing dataset for human behavior analysis. To solve this problem, we propose action recognition through an adaptive multitask (multitask-aware) loss function, which is designed to handle even datasets where only human pose annotations or only action labels are available (instead of both pose and action information).

Our contributions can be summarized as follows:

- Human action is an organic combination of the human pose, area of interaction, and interacting objects. In this context, we propose a deep network to efficiently fuse all the human-centric information based on multitask learning.

- The loss function in the proposed deep network is designed by a multitask-aware loss function. Thereby, the proposed method can efficiently handle various datasets, even if they do not include both human pose annotations and action labels.
- According to the experimental results, the proposed method achieved 91.91% mean average precision (mAP) on the Stanford 40 Actions Dataset [15], while having very little effect on the pose estimation task in our multitask learning framework. Moreover, we demonstrated that the proposed method can be flexibly applied in multi-labels action recognition problem on the V-COCO Dataset [16].

The rest of this paper is organized as follows. In Section 2, we introduce related work. The proposed method is presented in Section 3. Section 4 shows our extensive experiments. We report experimental results on the MPII Human Pose Dataset [17], Stanford 40 Actions Dataset [15], and V-COCO Dataset [16]. This is followed by conclusion in Section 5.

2. Related Works

2.1. Human Pose Estimation

Several studies have been published on the pose estimation [14,18–22]. Xiao et al. [21] stacked deconvolution layers over the last convolution stage in ResNet [23]; this simple architecture achieves better results than previous complex methods [19,20]. Sun et al. [14] started with a high-resolution sub-network and gradually added high-to-low sub-networks forming new stages; each sub-network receives information from other parallel sub-networks, which leads to rich high-resolution representation and improved spatial precision. They also achieved superior performance for pose detection on two benchmark datasets: the COCO Keypoint Detection Dataset [24] and the MPII Human Pose Dataset [17]. The proposed method exploits the essence of multitask learning by simultaneously learning the human joints and human actions because both tasks are human-centric, and intermediate results of the pose estimation task benefit the action recognition task.

2.2. Single-Image-Based Action Recognition

The problem we focus on in this paper is single-image-based action recognition, which has been actively studied due to its importance [25]. However, traditional person detectors [4,26] that rely on hand-crafted features were not adequate for proposing candidates bounding boxes for action classification. Khan et al. [26] proposed an action-specific person detection based on transfer learning to improve the quality of bounding boxes. Recently, deep learning has produced successful results in single-image-based action recognition [26–30]. Zhang et al. [28] proposed action recognition with only action labels, which learns the action masks and extracts the features from the objects in an action. Focusing on the part information, Zhao et al. [27] localized seven body parts by using an independent human pose estimator and trained a body-part-based action classification model based on manually annotated semantic part labels. Although this work [27] is similar to our proposed method, we estimate three body parts from the proposed multitask-based framework, instead of an independent human pose estimator, and do not use manual annotations for part action classification.

There is another research direction on recognizing human interaction with objects [16,31,32], which deviates from our main focus, as we use heterogeneous datasets for the human pose task and the action task in a multitasking manner. Moreover, we do not use the object information directly, i.e., we do not use object detectors.

2.3. Fine-Grained Classification Problems

Many fine-grained classification problems can be solved with part-based methods and visual attention methods [8–10]. Liu et al. [8] applied the attention model on part patches and achieved competitive accuracy on benchmark datasets. Zhang et al. [9] proposed an approach for fine-grained recognition by picking deep filters and training discriminative detectors, avoiding the need for

object/part annotations. Xiao et al. [10] presented two attention pipelines: object-level (predictions driven by several relevant patches of an image) and part-level (driven by detected parts); the final classification merges the results of both pipelines. Although intra-class variation problems are solved by proper localization techniques [8–10], fine-grained classification tasks suffer from inter-class similarity. Dubey et al. [33] addressed the problem of inter-class similarity affecting the feature learning in fine-grained classification tasks by introducing pair-wise confusion in output activations, hence improving the generalizability of networks.

Local information about human joints helps to recognize fine-grained classes of human actions. We use the local information around the human joints by using heat maps for body joints, which is similar to what was proposed by Fukui et al. [34].

2.4. Multitask Learning

Multitask learning has been applied to several areas, including human behavior understanding [5,11,12]. Luvizon et al. [5] reinforced that human pose is extremely relevant to action recognition and presented a single framework that performs end-to-end trainable multitasking to learn the 2D pose, 3D pose, and action recognition. Kim et al. [11] applied multitask learning to a baseline hypothesis that captioning data may be helpful for action classification because both are human-centric, and the captioning data are usually related to human actions. Similarly, Ranjan et al. [12] applied multitask learning to face detection, landmark localization, face pose estimation, and gender recognition tasks. They used the fact that the lower layers respond to edges and corners and, therefore, contain better localization properties, which can be used for landmark localization and pose estimation tasks. For face detection and gender recognition tasks, they used the deeper layers, as they are suitable for these complex tasks.

By multitask learning of the human pose and action, we cut the complexity of action learning but achieve comparable results with HRNet [14] for the pose estimation task, although there is an additional action recognition task. As discussed above, only few datasets contain both pose and action annotations. To overcome this limitation, the proposed method of action recognition through an adaptive multitask loss function is designed to be applicable if either the pose or action annotation is available.

3. Proposed Method

Human pose information is a key for action recognition problems, especially with single images. Therefore, the proposed method uses HRNet-based [14] human pose estimation network as a backbone network. Figure 1 depicts the overview of the proposed method, in which HRNet-based [14] feature extractor is used for human pose estimation and action classification in a multitasking manner. However, the shared feature extractor is not enough to classify fine-grained action classes based on the same human pose. To get an accurate representation of action classes, we add three feature-stream-based shallow sub-networks on the deep pose estimation network: (1) image-based feature stream; (2) attention-image-based feature stream; and (3) part-image-based feature stream. The features represented by these three streams of action recognition are complementary, as shown in our experiments in Section 4. Notably, Figure 1 does not depict independent multiple networks but it is one end-to-end trainable convolutional neural network.

3.1. HRNet: Multitasking Feature Representation

The proposed method uses HRNet [14] to represent multitasking features of input image $\mathbf{I}$, as depicted in Figure 1. A set of feature maps $\{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4\}$ is extracted from the image with decreasing order of resolution and increasing order of channels. The highest resolution of the feature maps $\mathbf{F}_1$ is fed through a 3×3 convolution layer: "Conv for Pose Estimation" (in Figure 1), which adjusts the number of channels to the number of human joints and returns K heat maps, where the kth heat map gives the location confidence of the kth joint.

Moreover, the same set of multitasking feature representation from the HRNet-based feature extractor is reused for action recognition. The image-based feature stream uses the lowest resolution of feature maps $\mathbf{F}_4$ (as depicted in Figure 1). The heat maps for body joints from "Conv for Pose Estimation" as a byproduct of the pose estimation are used to generate an attention-guided image as an input of the attention-image-based feature stream. Additionally, the rich information from the set of feature maps $\{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4\}$ are also passed through the attention-image-based feature stream; thereby, the sub-network for the stream can be shallow. For the part-image-based feature stream, we use the human joints information to localize the body parts.

3.2. Image-Based Feature Stream

The image-based feature stream uses the lowest resolution of the features maps $\mathbf{F}_4$ from the HRNet-based feature extractor for the action recognition task. As presented in Figure 2, average pooling is performed on the feature maps $\mathbf{F_4}$ to produce feature vector $\mathbf{f}_{img}$, followed by passing through a fully connected layer. The action can be trained from an action loss. The detailed configuration of the image-based feature stream is provided in Section 3.5.

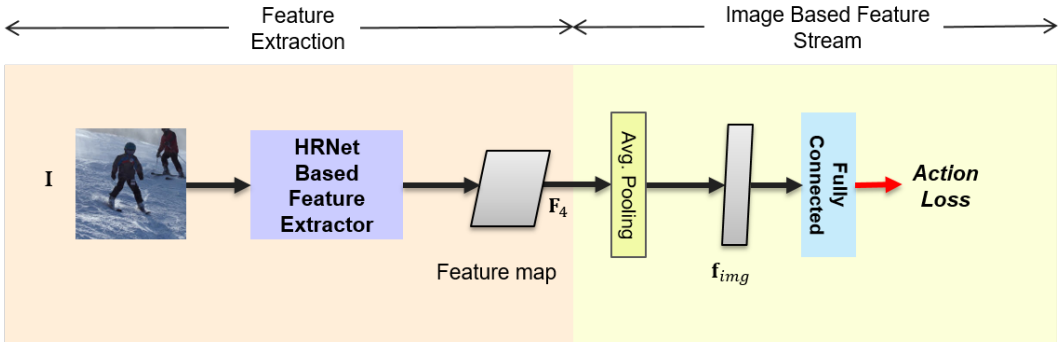

Figure 2. Image-based feature stream. The figure depicts the shared HRNet-based feature extractor and image-based feature stream depicted in Figure 1. The background colors are the same as Figure 1.

3.3. Attention-Image-Based Feature Stream

In this section, we discuss the attention-image-based feature stream for foreground analysis and the significance of this stream for feature extraction. Figure 3 presents the flow of this stream. Inspired by Fukui et al. [34], we use an attention-guided image for the input of the attention-image-based feature stream:

$$\mathbf{I}_{att} = \mathbf{I} \circ (\mathbf{1} + \mathbf{A}) = \mathbf{I} + \mathbf{I} \circ \mathbf{A}, \tag{1}$$

where $\circ$ denotes element-wise product, $\mathbf{I}$ is the input image, and $\mathbf{1}$ is a tensor whose size is the same as the input image and all elements are one. $\mathbf{A}$ is the attention map. To generate attention map $\mathbf{A}$, heat maps for body joints from the "Conv for Pose Estimation" are passed through "conv-bn-relu" to adjust the number of channels to the input image. This is followed by the upsampling layer, so that the generated attention map $\mathbf{A}$ has the same dimensions as the input image.

The "attention-ed" image $\mathbf{I}_{att}$ is passed through the convolution, batch normalization [35], and rectified non-linear unit (ReLU [36]), followed by max pooling. As depicted in Figure 3, the attention-image-based feature stream is based on a ResNet18 [23], which is shallow. We aid the shallow stream with skip connections from the feature maps $\{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4\}$ extracted by the HRNet, in which the feature maps $\{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4\}$ are concatenated with each output of the layers in ResNet18 [23] by using transition blocks, as depicted in Figure 3. Finally, average pooling is performed to create the feature vector $\mathbf{f}_{att}$ for the attention-image-based feature stream. The detailed configuration of the attention-image-based feature stream with information such as filter size and stride is provided in Section 3.5.

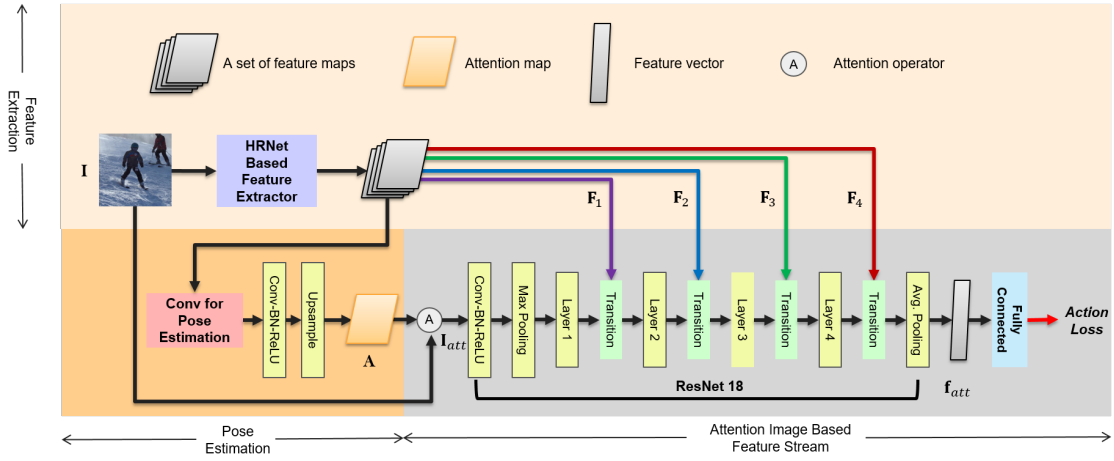

Figure 3. Attention-image-based feature stream. The figure depicts the shared feature extractor, pose estimation, and attention-image-based feature stream depicted in Figure 1. The background colors are the same as Figure 1.

We discuss the effect of the attention-image-based feature stream in Section 4.5: it enhances the learning of the feature extraction network in our model, which, in turn, improves the localization of the image-based feature stream.

3.4. Part-Image-Based Feature Stream

Although the pose is highly informative to recognize a variety of actions, some actions share a similar pose structure. Therefore, human action recognition can be considered as a fine-grained classification problem, which we address by our part-image-based feature stream. Instead of using an independent pose estimator to localize joint locations in the previous methods [27], we use the intermediate pose estimation result of the proposed multitask-based framework.

Figure 4 depicts the details of the part-image-based feature stream. The "Conv for Pose Estimation" in Figure 4 produces the heat maps for body joints, which are used along with the input image $\mathbf{I}$ to localize the body parts. The first column images in Figure 5 depict the intermediate pose estimation results, which is used to create body part images.

For the input of part-image-based feature stream, it is important to crop properly sized patches centering around the estimated joint locations. That is, a cropped part image should have a consistent sized body part, regardless of the size of the human in the input image $\mathbf{I}$. For this, we calculate a cropping base length l as in Equation (2) and the width and the height of a part patch is set to two times the base length l.

$$l = \frac{1}{4}\sqrt{(max_x - min_x)^2 + (max_y - min_y)^2}, \tag{2}$$

where max_x and max_y are the maximum values along the x-coordinate and y-coordinate, respectively, among all the joint coordinates. Similarly, among all the joint coordinates, min_x and min_y are the minimum values along x-coordinate and y-coordinate, respectively. Based on Equation (2), we crop three part images (head, right hand, and left hand), which are further resized to the same size as the input image $\mathbf{I}$. We denote the cropped and resized images for a head, right hand, and left hand as $\mathbf{I}_h$, $\mathbf{I}_r$, and $\mathbf{I}_l$, respectively. The second to fourth column images in Figure 5 depict the examples of $\mathbf{I}_h$, $\mathbf{I}_r$, and $\mathbf{I}_l$.

Individually, $\mathbf{I}_h$, $\mathbf{I}_r$, and $\mathbf{I}_l$ go through a ResNet18 network [23] with shared weights to generate three body part-based feature vectors ($\mathbf{f}_h$, $\mathbf{f}_r$, and $\mathbf{f}_l$). Then, the three part's feature vectors ($\mathbf{f}_h$, $\mathbf{f}_r$, and $\mathbf{f}_l$) are concatenated to form a single-part-based feature vector $\mathbf{f}_{part}$. This feature vector $\mathbf{f}_{part}$ is passed through the fully connected layer to produce confidence scores for the actions. The detailed

configuration of the part-image-based feature stream, such as filter size and stride, is provided in Section 3.5.

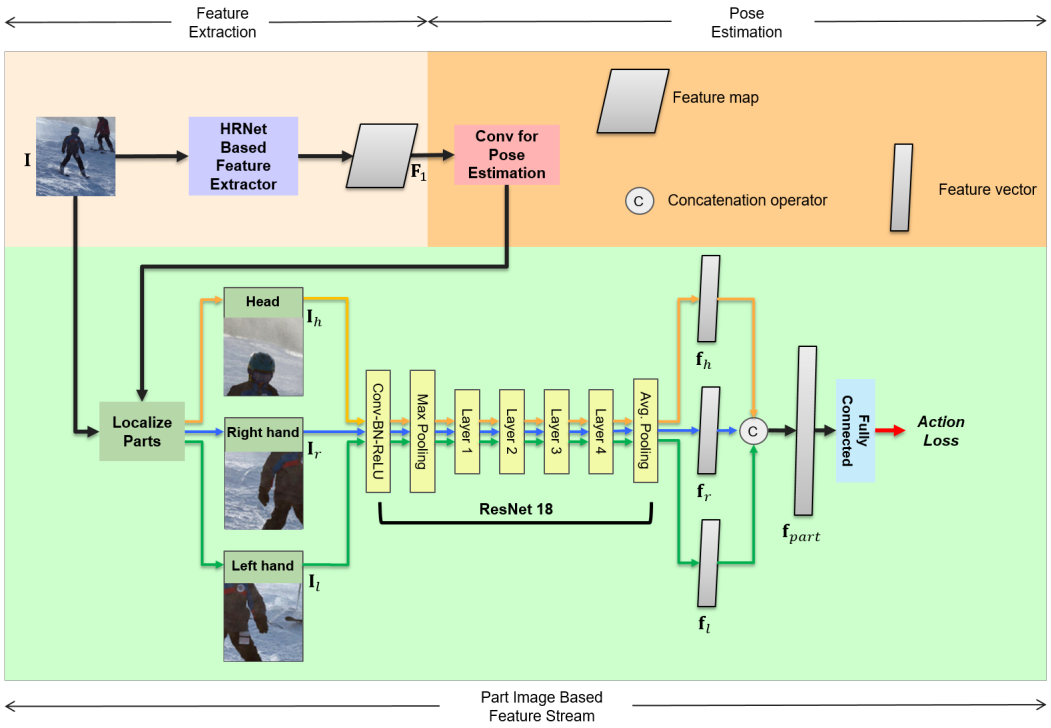

Figure 4. Part-image-based feature stream. The figure depicts the shared feature extractor, pose estimation, and part-image-based feature stream shown in Figure 1. The background colors are the same as Figure 1.

Figure 5. An example of parts localization for the part-image-based feature stream based on the V-COCO dataset [16]. The localized parts are input to the part-image-based feature stream.

3.5. Detailed Structure of the Three Feature Streams for Action Recognition

Table 1 only depicts the configuration of the three feature streams in the proposed method because the configuration of the HRNet-based feature extractor is the same as the original HRNet [14]. The image-based feature stream uses the HRNet-based multitasking feature $\mathbf{F}_4$ for the action task; therefore, additional layers are just average pooling and fully connected layers. The attention-image-based feature stream and part-image-based feature stream are based on ResNet18 [23]. ResNet18 has a sequence of four main blocks and we denote them as Layers 1–4, respectively. Each Layer block consists of two basic blocks.

We follow notations used in ResNet18 [23]: The bracket $[\cdot]$ in Table 1 is used to represent the convolution filter size and the number of filters. For example, Layer 1 in the attention-image-based feature stream consists of two basic blocks. $[3 \times 3, 64 \setminus\setminus 3 \times 3, 64]$ represents a basic building block where two convolutional layers are used and the filter size is 3×3 and the number of filters is 64. Furthermore, by the first basic block in each Layer, down-sampling is performed with a stride of 2 and

then it is followed by batch normalization [35] and ReLU [36] non-linearity. The second basic block in each layer is followed by batch normalization [35] (see [23] for more details).

Table 1. Configuration of the three feature streams (FS) in the proposed method for action recognition. Attention image-based FS and part Iimage-based FS are based on ResNet18 [23]. The layer names are the same as in Figures 2–4 (see [23] for more details). Notably, there are additional transition layers unlike ResNet18 [23].

Layer Name	Image Based FS	Attention Image Based FS	Part Image Based FS
Conv	-	$[7 \times 7, 64]$ stride = 2	$[7 \times 7, 64]$ stride = 2
Max Pooling	-	$[3 \times 3]$ stride = 2	$[3 \times 3]$ stride = 2
Layer 1	-	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$
Transition	-	$[3 \times 3, 64]$	-
Layer 2	-	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$
Transition	-	$[3 \times 3, 128]$	-
Layer 3	-	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$
Transition	-	$[3 \times 3, 256]$	-
Layer 4	-	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$
Transition	-	$[3 \times 3, 512]$	-
Avg. Pooling	Average Pooling	Average Pooling	Average Pooling
Fully Connected	#classes-d	#classes-d	#classes-d
	softmax	softmax	softmax

In the attention-image-based feature stream, there are additional transition layers unlike a ResNet18 network because the feature maps $\{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4\}$ extracted by the HRNet-based feature extractor are concatenated with the output of each layer (Layer 1, Layer 2, Layer 3, or Layer 4), as depicted in Figure 3. Furthermore, the transition layer is followed by batch normalization [35] and ReLU [36] non-linearity.

3.6. Multitask-Aware Loss Function

The proposed method is based on multitask learning of human pose and action. The multitask loss function consists of two losses in learning as

$$\mathcal{L}(\mathbf{I}_i, \mathbf{p}_i, y_i) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_p(\mathbf{p}_i) \mathcal{L}_p(\mathbf{I}_i, \mathbf{p}_i) + \lambda \mathbb{1}_a(y_i) \mathcal{L}_a(\mathbf{I}_i, y_i), \tag{3}$$

where λ is a hyperparameter to balance the losses between the pose and action tasks. N is the number of samples, $\mathcal{L}_p(\cdot)$ is the loss in pose estimation, and $\mathcal{L}_a(\cdot)$ is the loss in learning the human action. $\mathbf{I}_i$ is the ith input image and its ground truth labels for a human pose and action are $\mathbf{p}_i$ and y_i, respectively. All elements of $\mathbf{p}_i$ are set to -1 if there is no available ground truth for the ith input image in the

human pose dataset. Similarly, y_i is set to -1 if there is no available ground truth for the ith input image in the action dataset. $\mathbb{1}_p(\mathbf{p}_i)$ is an indicator function that evaluates to 1 if the pose annotation ($\mathbf{p}_i$) is available and 0 otherwise. Similarly, $\mathbb{1}_a(y_i)$ evaluates to 1 if the action annotation (y_i) is available and 0 otherwise. Therefore, for each training sample, we calculate the pose loss and/or the action loss depending on the availability of the respective annotations. For the human pose loss, $\mathcal{L}_p(\cdot)$, we use a weighted mean squared error as in [14]. The action loss $\mathcal{L}_a(\cdot)$ is a combination of loss in learning the action through multiple feature streams, as in Equation (3).

$$\mathcal{L}_a(\mathbf{I}_i, y_i) = \mathcal{L}_{img}(\mathbf{I}_i, y_i) + \mathcal{L}_{att}(\mathbf{I}_i, y_i) + \mathcal{L}_{part}(\mathbf{I}_i, y_i) + \mathcal{L}_{concat}(\mathbf{I}_i, y_i) \tag{4}$$

Here, $\mathcal{L}_{img}(\cdot)$, $\mathcal{L}_{att}(\cdot)$, and $\mathcal{L}_{part}(\cdot)$ are the loss in learning human action using the image-based feature stream (Section 3.2), attention-image-based feature stream (Section 3.3), and part-image-based feature stream (Section 3.4), respectively. $\mathcal{L}_{concat}(\cdot)$ is the loss in learning the human action from the concatenated features from the three feature streams. The loss functions for the action recognition task are a focal loss [37] and a Binary Cross Entropy (BCE) loss [38] (Sections 4.2.2 and 4.2.3, respectively), depending on the different action recognition tasks.

4. Experiments

4.1. Experimental Setup

The proposed algorithm used a top-down human pose estimator [14] as a backbone to estimate a human pose. To operate this kind of pose estimators, we assumed that the location of a person can be determined by using a human detector on the image. Following Sun et al. [14], we used the human box information to adjust the scale and position of the person in the image. This is because black areas exist in resulting images (as depicted in our experiment figures). Notably, it is a common preprocessing in human pose estimation research. For the proposed method, a RGB image was used, which was resized to sizes 256×256 and 512×512. The resized image to 256×256 was used for the input $\mathbf{I}$ of the proposed method and the image of size 512×512 was used for a crop base image for the three part images. The cropped images from the crop base image 512×512 were followed by resizing to 256×256 for the input of part-image-based stream. All input images were normalized with a mean and standard deviation of ImageNet [39].

As a feature extractor, HRNet-W32 was used. The HRNet-W32 and ResNet18 modules in the proposed method were initialized by ImageNet pretrained models. The batch size was 16. Training was performed using the Adam optimizer [40] with the initial learning rate of 1×10^{-5}. Training was performed until 210 epochs with the decaying rate factor of 0.1 on 170 and 200 epochs. For data augmentation, we used affine transformation with the rotation factor of 90 degrees and the scale factor of 0.25 for the MPII human pose dataset (we followed the affine data augmentation in [14], but changed the rotation angle from 30 to 90 degrees). The rotation factor of 10 degrees and the scale factor of 0.25 were used for the Stanford 40 Actions Dataset. Flip image augmentation was used. Hyperparameter λ in Equation (3) was set to 0.00075.

4.2. Datasets

4.2.1. MPII Human Pose Dataset

The MPII Human Pose Dataset [17] has 24,920 images, including 40,522 person instances labeled with the joints. The dataset consists of instances from YouTube videos with diverse human activities and events. Each person instance has 16 labeled joints. Among the 40,522 total person instances, we used 28,821 and 11,701 for training and testing, respectively. To evaluate the pose estimation, we used the head-normalized probability of the correct keypoint (PCKh score). The PCKh score is the standard metric in the MPII Human Pose Estimation [17]. Specifically, we used PCKh@0.5, which regards 50% of the length of the head segment as a threshold for joint localization accuracy, i.e.,

if the distance between the estimated joint position and the ground-truth position is less than 0.5 times the length of the head segment, it is considered accurate.

4.2.2. Stanford 40 Actions Dataset

We performed extensive experiments on the Stanford 40 Actions Dataset [15]. The dataset includes 40 diverse action classes (such as applauding, blowing bubbles, brushing teeth, fishing, fixing a bike, taking photos, phoning, and walking a dog). The dataset contains images collected from Google, Flickr, and Bing. We use the standard training and testing sets (4000 train images and 5532 test images).

4.2.3. V-COCO Dataset

The V-COCO (Verbs in COCO) Dataset [16] contains a total of 10,346 images, including 16,119 people instances. Each of these instances has binary labels for 26 different action classes. The dataset is split into "train", "val" and "test" sets. The contents of the dataset come from the COCO Dataset [24]. The "train" and "val" split are from the "COCO train" set while the "test" split is from the "COCO val" set. The "train" split contains 2533 images and 3932 people instances. The "val" split contains 2867 images and 4499 people instances. The "test" split contains 4946 images and 7768 people instances. We used the "train" and "val" splits for training and testing the proposed method, respectively.

4.3. Evaluation of the Pose Estimation Task on the MPII Human Pose Dataset

Table 2 compares the experimental results of the pose estimation task in our multitask-based framework with the performance of only a single pose estimation framework [14]. Here, "pose_hrnet_w32" is the pose estimator trained by the author of [14] and "pose_hrnet_w32_aug90" is a retrained version with the same rotation augmentation as explained in Section 4.1. As is the nature of multitask learning, the pose estimation task in our method suffers from the additional action recognition task. However, the performance drop is very low, considering the additional action recognition task due to the multitask-aware loss function in Equation (3) and the dense connectivity in the HRNet [14], as we expected. It demonstrates that performing pose estimation and action recognition using our approach (adaptive multitask learning) saves the time and effort required for annotations. It is very useful, because there are very few data with pose-level and action-level annotations together. Other part-based methods [29] that do not use the pose-level annotations train multiple part detectors to first obtain coarse parts and further learn on these coarse parts to obtain the most discriminative parts, which are computationally expensive. Unlike other methods [27], the proposed action recognition method benefits from abundant information, such as the pose and multiresolution features, while being computationally inexpensive and achieving improvement in overall generalized performance.

Table 2. Evaluation of the pose estimation task in our multitask-based method versus pose estimation as a single task on the MPII Human Pose Dataset [17]. The performance metric used is PCKh@0.5 [17]. Bold font indicates the best result.

Arch	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
pose_hrnet_w32 [14]	97.1	95.9	90.3	**86.4**	89.1	**87.1**	**83.3**	90.3
pose_hrnet_w32_aug90 [14]	**97.3**	**96.1**	**90.9**	86.2	**89.2**	87.0	82.8	**90.4**
proposed_method	96.7	95.4	88.7	83.3	87.5	84.4	79.4	88.5

4.4. Evaluation of the Action Recognition Task on the Stanford 40 Actions Dataset

In this section, we present experimental results for action recognition by highlighting the importance of the feature streams and compare the performance with the existing methods on the Stanford 40 Actions Dataset. The action recognition task in the proposed method has three complementary streams. Each of these streams was trained with a loss function. Thus, we report the mAP on each of these streams and the concatenation-based feature in Table 3. The mAP obtained on individual feature streams are: 85.57%

(image-based feature stream in Section 3.2), 85.38% (attention-image-based feature stream in Section 3.3), and 84.27% (part-image-based feature stream in Section 3.4). On concatenation-based action decision in Figure 1 (Image + Attention + Part), we obtained mAP of 91.76%. As depicted in Figure 1, each of three streams is a component of the entire end-to-end trainable network. Thus, the concatenation feature can learn the combination of features extracted by the three streams, achieving better performance than those by three streams. Additionally, we evaluated the ensemble of the estimated action recognition result from multiple streams of the proposed method because the ensemble technique [41] is a method commonly used to improve generalized performance. We denote it as "Ensemble of (Image + Part + Concatenation)" in Table 3. The obtained mAP is 91.91%. We report the ensemble of (Image + Part + Concatenation) as the proposed method for comparison.

Table 3. Comparison of mAP obtained through feature representation from multiple streams for action recognition on the Stanford 40 Actions Dataset [15]. Bold font indicates the best result.

Feature Representation	mAP
Image-based feature stream	85.57
Attention-image-based feature stream	85.38
Part-image-based feature stream	84.27
Concatenation-Based Action Decision (Image + Attention + Part)	91.76
Proposed: Ensemble of (Image + Part + Concatenation)	**91.91**

Figure 6 depicts the average precision (AP) of detailed action classes obtained from three feature streams and the result of the ensemble technique in Table 3. The features learned from the three feature streams are complementary; the performance of one is better than the others in some action classes, and vice-versa. Some notable classes are "Blowing bubbles" (95.00%), "Brushing teeth" (84.99%), "Phoning" (73.92%), and "Smoking" (73.11%) that perform better for the part-image-based feature stream in comparison to image-based feature stream and attention-image-based feature stream. One thing common to these action classes is that the major distinction is based on interacting objects and the parts of the human body. The part-image-based feature stream is designed to focus on body parts (head, right hand, and left hand) and objects around them. Such action classes have higher APs for part-image-based feature stream. However, on other actions such as "Jumping", "Running", and "Walking a dog", the part-image-based feature stream performs worse than other feature streams. The image-based feature stream obtained an AP of 97.62% on "Jumping" and 93.61% on "Running". Similarly, attention-image-based feature stream obtained an AP of 97.41% and 93.47%, respectively, on the same set of actions. For actions such as "Jumping", "Running", "Walking the dog", and "Gardening", the human pose and background information is much more relevant than the detailed body parts information. Overall, the "Proposed: Ensemble of (Image + Part + Concatenation)" is the best. It means that each of the feature streams learns complementary features for action recognition, as we intended.

In Table 4, we evaluate our methods by comparing them with the existing ones on the Stanford 40 Actions Dataset. Among the other methods (Action-Specific Detectors [26], VGG-16&19 [42], TDP [29], ResNet-50 [23], and Action-Mask [28]), the performance of the method in [27] is the best. Part-Based Network in [27] and Part Action Network in [27] use the body joints information to localize seven body parts and feed the network with the person bounding box image along with the seven body part images. Moreover, the Part Action Network [27] uses manually labeled part action annotations on the Stanford 40 Actions Dataset. The Part-Based Network obtained 89.3% mAP and the part action network obtained 91.2% mAP, whereas our proposed method (91.91%) uses only three body parts and does not use additional annotations for body parts. The body joints in the proposed method are localized in a multitasking manner, unlike the method in [27], which uses an independent human pose estimator.

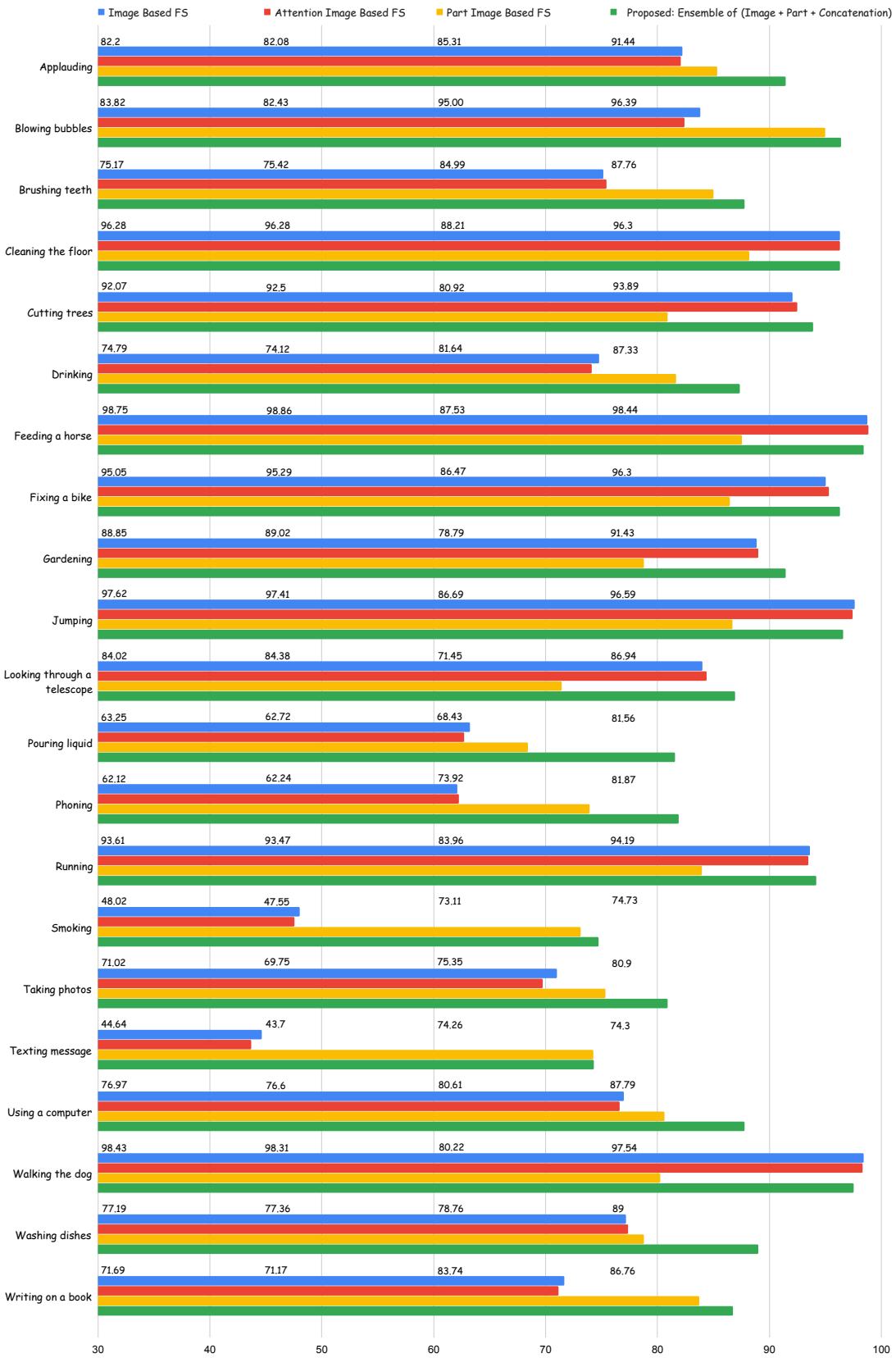

Figure 6. Detailed performance evaluation (average precision) of our proposed method. FS, feature stream. Results reported are on the Stanford 40 Actions Dataset [15]. Only some action classes are represented in the figure.

Table 4. Comparison of mAP with other methods on the Stanford 40 Actions Dataset [15]. Bold font indicates the best result.

Method	mAP
Action-Specific Detectors [26]	75.4
VGG-16&19 [42]	77.8
TDP [29]	80.6
ResNet-50 [23]	81.2
Action Mask [28]	82.6
Part-Based Network in [27]	89.3
Part Action Network in [27]	91.2
Proposed Method	**91.91**

4.5. Evaluation of the Action Recognition Task on the V-COCO Dataset

To further validate the usefulness of our multitask-aware loss function, we performed experiments on the V-COCO Dataset [16]. Because the V-COCO Dataset is designed for human object interaction benchmark, a single image contains multiple human actions and multiple object labels. However, our focus was to perform action recognition for a single person. Thus, we created training and testing images by transforming each person at the center using a bounding box. We solved multi-labels action recognition task for the centered person on the V-COCO Dataset. We used the BCE [38] loss function for the multi-labels problem. This is different from the action loss function used on the Stanford 40 Actions Dataset. This shows the flexibility of multitask-aware loss function in Equation (3). Table 5 depicts comparison of the mAP obtained on the V-COCO Dataset through different feature streams and that of the of ensemble of multi-streams. The mAP obtained are 64.59% (image-based feature stream), 66.30% (attention-image-based feature stream), 62.50% (part-image-based feature stream) and 72.36% (Proposed: Ensemble of (Image + Part + Concatenation)).

Table 5. Comparison of mAP obtained through feature representation from multiple streams for action recognition on the V-COCO Dataset [16]. Bold font indicates the best result.

Feature Representation	mAP
Image-based feature stream	64.59
Attention-image-based feature stream	66.30
Part-image-based feature stream	62.50
Proposed: Ensemble of (Image + Part + Concatenation)	**72.36**

To analyze the effectiveness of different feature streams in our proposed method, we performed experiments by training the models independently with these feature streams excluded from the architecture and including them. In Table 6, "Baseline 1 (trained with Image)" represents the basic multitask learning framework with only one stream (image-based feature stream) for action recognition. "Baseline 2 (trained with Image + Part)" represents performance when including the part-image-based feature stream on training the deep network. "Proposed (trained with Image + Attention + Part)" is trained including three feature streams (image-based, attention-image-based, and part-image-based). As can be seen in Table 6, Baseline 1 achieved a mAP of 68.08%, which was improved by a large margin when trained with the part-image-based feature stream included, reaching a mAP of 69.15% (Baseline 1). Overall, the proposed training method that includes the three modules for action recognition (i.e., image-based, part-image-based, and attention-image-based feature streams) obtained the mAP of 72.36%.

Table 6. Comparison of mAP with the baseline methods on the V-COCO Dataset [16]. Bold font indicates the best result.

Methods	mAP
Baseline 1 (trained with Image)	68.08
Baseline 2 (trained with Image + Part)	69.15
Proposed (trained with Image + Attention + Part)	**72.36**

In Figure 7, we compare the class activation mappings [43] of the image-based feature stream on Baseline 1 (trained with Image), Baseline 2 (trained with Image + Part), and Proposed (trained with Image + Attention + Part). According to the experimental results, the image-based feature stream in Baseline 1 can focus on both global and local information. Because Baseline 1 consists of only one feature stream for action recognition (the image-based feature stream), it is optimized to focus on both global and local information. On Baseline 2, a part-image-based feature stream is included in training, which is designed to focus on the local information. Thus, the image-based feature stream can focus on complementary global information. This, however, degrades the accuracy of the image-based feature stream, as seen in Figure 7. We found the image-based feature stream is improved by the attention-image-based feature stream on training i.e., "Proposed method", as shown by the confidence values in Figure 7. Since attention-image-based feature stream focuses on global information and the feature extractor is a shared one, the image-based feature stream tends to focus on the local information regarding the human body. Therefore, the role of attention-image-based feature stream is to enhance the feature extraction capability of the image-based feature stream.

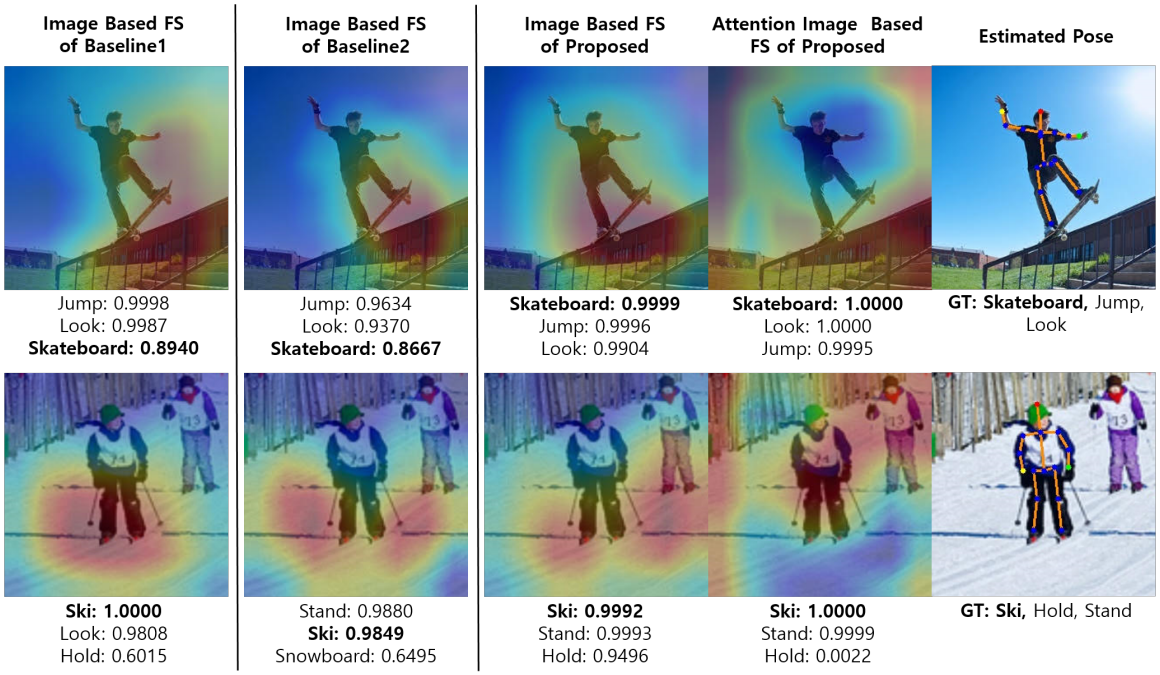

Figure 7. FS denotes Feature Stream. The first column depicts the results of the image-based feature stream on Baseline 1. The second column depicts the results of the image-based feature stream on Baseline 2. The third column depicts the results of the image-based feature stream with the proposed method. The fourth column depicts the results of the attention-image-based feature stream with the proposed method. The fifth column depicts the joints location from the pose estimation task. The results reported are on the V-COCO Dataset [16]. Class activation maps [43] comparison for the bold GTs ("Skateboard" in first row and "Ski" in second row).

4.6. Qualitative Results

We propose a single end-to-end trainable multitask learning method for pose estimation and action recognition. The MPII Human Pose Dataset [17] contains only pose annotations for training. However, as can be seen in Figure 8, we present both pose estimation and action recognition results on this dataset. Similarly, in Figure 9, we present both pose estimation and action recognition results on the Stanford 40 Actions Dataset [15], even though the dataset is only meant for action recognition and does not contain pose annotations for training. Further, we present experimental results on the V-COCO Dataset [16] in Figure 10. The V-COCO Dataset contains multi-labels for actions. Thus, the proposed method can even handle heterogeneous datasets while simultaneously performing both pose estimation and action recognition task.

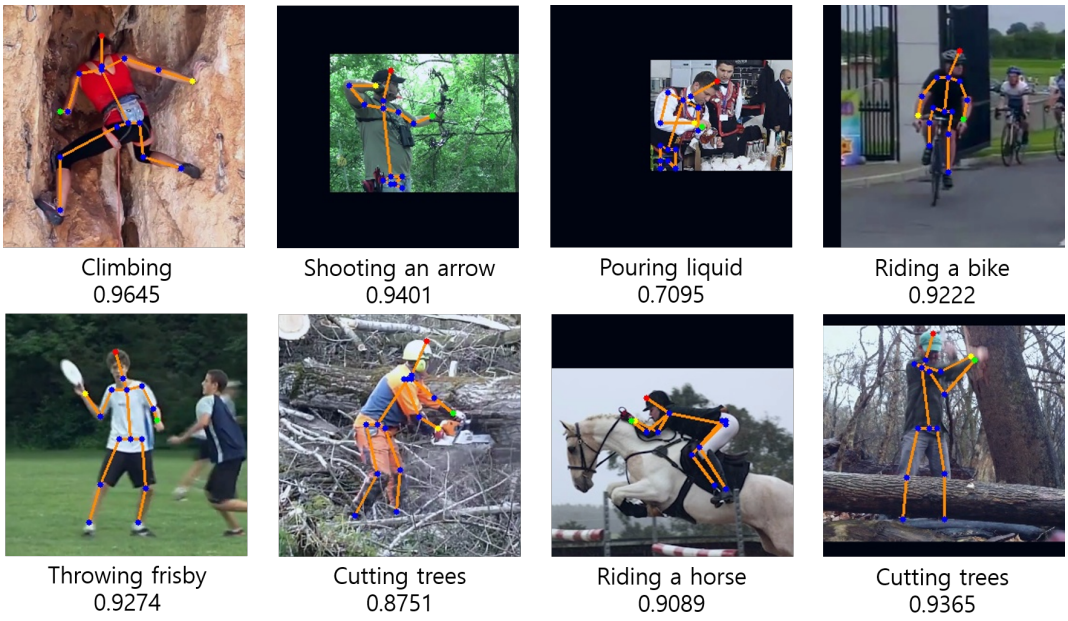

Figure 8. Pose estimation and action recognition results on the MPII Human Pose Dataset [17]. Predicted action classes along with confidence scores are shown below the individual results.

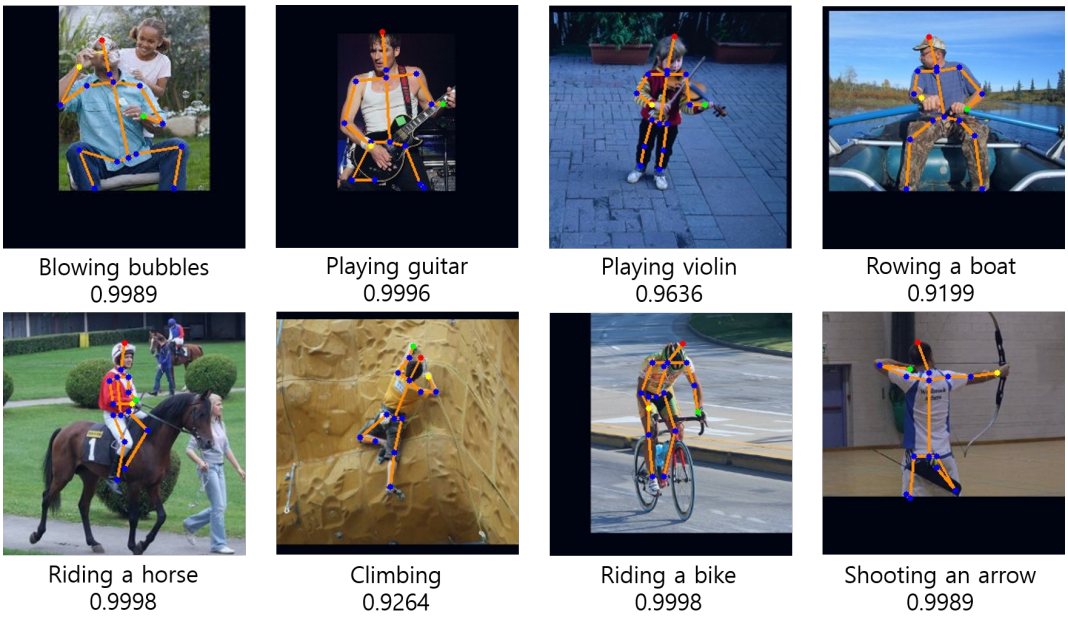

Figure 9. Pose estimation and action recognition results on the Stanford 40 Actions Dataset [15]. Predicted action classes along with confidence scores are shown below the individual results.

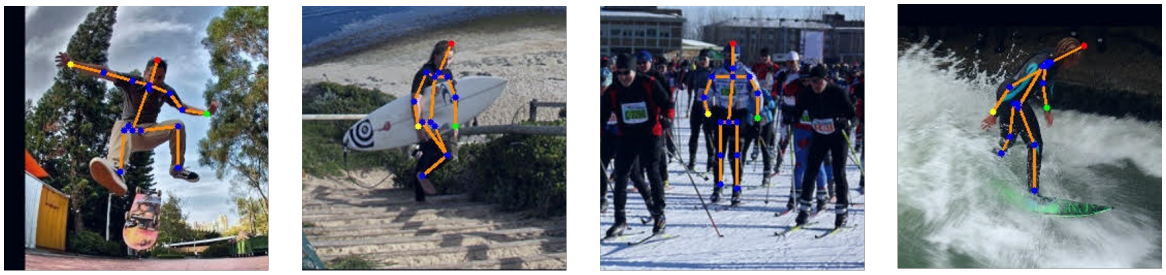

Figure 10. Pose estimation and action recognition results on the V-COCO Dataset [16], which has multi-labels for a person. The ground truth labels and scores for the ground truths are shown below the individual results.

5. Conclusions

In this paper, we propose a multitask-aware single-image-based action recognition method for intelligent systems. Specifically, according to our findings, human action is best represented using multiple streams: (1) image-based; (2) attention-based; and (3) part-based. We showed in our experiments that the features represented by these streams are complementary and aid in the final classification of the human action. Moreover, the proposed deep networks for action recognition are less expensive in computations than the previous method while attaining better results. We believe that our method contributes to the progress of real-time intelligent sensor systems.

Author Contributions: J.C. conceived the idea and designed and performed the experiments; B.B. and G.L. performed the experiments. B.B. and J.C. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019R1F1A1058666), in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. B0101-15-0266, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis), and in part by the Gachon University research fund of 2019 (No. GCU-2019-0775).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Phyo, C.N.; Zin, T.T.; Tin, P. Complex Human–Object Interactions Analyzer Using a DCNN and SVM Hybrid Approach. *Appl. Sci.* **2019**, *9*, 1869. [CrossRef]
2. Yang, H.; Zhang, J.; Li, S.; Lei, J.; Chen, S. Attend it again: Recurrent attention convolutional neural network for action recognition. *Appl. Sci.* **2018**, *8*, 383. [CrossRef]
3. Cho, J.; Lee, M. Building a Compact Convolutional Neural Network for Embedded Intelligent Sensor Systems Using Group Sparsity and Knowledge Distillation. *Sensors* **2019**, *19*, 4307. [CrossRef] [PubMed]
4. Yang, H.D.; Lee, S.W.; Lee, S.W. Multiple human detection and tracking based on weighted temporal texture features. *Int. J. Pattern Recognit. Artif. Intell.* **2006**, *20*, 377–391. [CrossRef]
5. Luvizon, D.C.; Picard, D.; Tabia, H. 2d/3d pose estimation and action recognition using multitask deep learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
6. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
7. Choutas, V.; Weinzaepfel, P.; Revaud, J.; Schmid, C. Potion: Pose motion representation for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.

8. Liu, X.; Xia, T.; Wang, J.; Yang, Y.; Zhou, F.; Lin, Y. Fully convolutional attention networks for fine-grained recognition. *arXiv* **2016**, arXiv:1603.06765.
9. Zhang, X.; Xiong, H.; Zhou, W.; Lin, W.; Tian, Q. Picking deep filter responses for fine-grained image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
10. Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; Zhang, Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
11. Kim, D.J.; Choi, J.; Oh, T.H.; Yoon, Y.; Kweon, I.S. Disjoint multi-task learning between heterogeneous human-centric tasks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 12–15 March 2018.
12. Ranjan, R.; Patel, V.M.; Chellappa, R. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 121–135. [CrossRef] [PubMed]
13. Huang, G.; Chen, D.; Li, T.; Wu, F.; van der Maaten, L.; Weinberger, K.Q. Multi-scale dense networks for resource efficient image classification. *arXiv* **2017**, arXiv:1703.09844.
14. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
15. Yao, B.; Jiang, X.; Khosla, A.; Lin, A.L.; Guibas, L.; Fei-Fei, L. Human action recognition by learning bases of action attributes and parts. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.
16. Gupta, S.; Malik, J. Visual Semantic Role Labeling. *arXiv* **2015**, arXiv:1505.04474.
17. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
18. Cho, J.; Lee, M.; Oh, S. Single image 3D human pose estimation using a procrustean normal distribution mixture model and model transformation. *Comput. Vis. Image Underst.* **2017**, *155*, 150–161. [CrossRef]
19. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.
20. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
21. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
22. Cha, G.; Lee, M.; Cho, J.; Oh, S. Deep pose consensus networks. *Comput. Vis. Image Underst.* **2019**, *182*, 64–70. [CrossRef]
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
24. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; Zitnick, C..L. Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
25. Guo, G.; Lai, A. A survey on still image based human action recognition. *Pattern Recognit.* **2014**, *47*, 3343–3361. [CrossRef]
26. Khan, F.S.; Xu, J.; Van De Weijer, J.; Bagdanov, A.D.; Anwer, R.M.; Lopez, A.M. Recognizing actions through action-specific person detection. *IEEE Trans. Image Process.* **2015**, *24*, 4422–4432. [CrossRef] [PubMed]
27. Zhao, Z.; Ma, H.; You, S. Single image action recognition using semantic body part actions. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
28. Zhang, Y.; Cheng, L.; Wu, J.; Cai, J.; Do, M.N.; Lu, J. Action recognition in still images with minimum annotation efforts. *IEEE Trans. Image Process.* **2016**, *25*, 5479–5490. [CrossRef] [PubMed]
29. Zhao, Z.; Ma, H.; Chen, X. Semantic parts based top-down pyramid for action recognition. *Pattern Recognit. Lett.* **2016**, *84*, 134–141. [CrossRef]

30. Safaei, M.; Foroosh, H. Still Image Action Recognition by Predicting Spatial-Temporal Pixel Evolution. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Waikoloa Village, HI, USA, 8–10 January 2019.
31. Chao, Y.W.; Liu, Y.; Liu, X.; Zeng, H.; Deng, J. Learning to detect human-object interactions. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 12–15 March 2018.
32. Wan, B.; Zhou, D.; Liu, Y.; Li, R.; He, X. Pose-Aware Multi-Level Feature Network for Human Object Interaction Detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
33. Dubey, A.; Gupta, O.; Guo, P.; Raskar, R.; Farrell, R.; Naik, N. Pairwise confusion for fine-grained visual classification. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
34. Fukui, H.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. Attention branch network: Learning of attention mechanism for visual explanation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
35. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
36. Dahl, G.E.; Sainath, T.N.; Hinton, G.E. Improving deep neural networks for LVCSR using rectified linear units and dropout. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013.
37. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
38. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
39. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
40. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
41. Lan, X.; Zhu, X.; Gong, S. Knowledge distillation by on-the-fly native ensemble. In Proceedings of the International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018.
42. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
43. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.