# Region Pooling with Adaptive Feature Fusion for End-to-End Person Recognition

Vijay Kumar Anoop Namboodiri C. V. Jawahar
CVIT, IIIT Hyderabad, India

## Abstract

*Current approaches for person recognition train an ensemble of region specific convolutional neural networks for representation learning, and then adopt naive fusion strategies to combine their features or predictions during testing. In this paper, we propose an unified end-to-end architecture that generates a complete person representation based on pooling and aggregation of features from multiple body regions. Our network takes a person image and the predetermined locations of body regions as input, and generates common feature maps that are shared across all the regions. Multiple features corresponding to different regions are then pooled and combined with an aggregation block, where the adaptive weights required for aggregation are obtained through an attention mechanism. Evaluations on three person recognition datasets - PIPA, Soccer and Hannah show that a single model trained end-to-end is computationally faster, requires fewer parameters and achieves improved performance over separately trained models.*

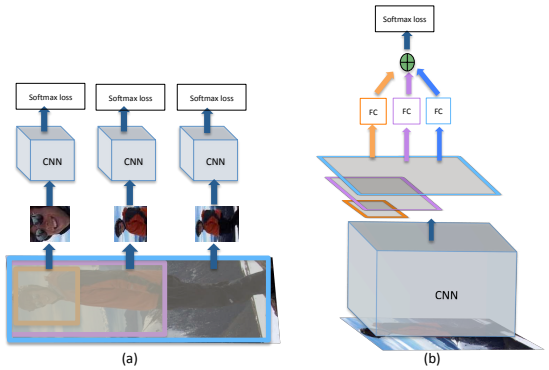


Figure 1: (a) Current approaches for person recognition follow an ad-hoc scheme in which features are extracted from several region-specific models and combined as two separate process. (b) Our proposed unified end-to-end trainable model that shares computations across multiple body regions and produces compact person representation by adaptively aggregating several pooled features.

## 1. Introduction

Person recognition in unconstrained real-world scenarios such as photo-albums, entertainment or surveillance videos is a challenging problem. People often appear in arbitrary poses and view points, and instances may be of low resolution, blurred or severely occluded. Often, the most discriminative facial region may not be visible completely. The state-of-the-art face recognition algorithms when applied in such scenarios typically under-perform and are not reliable [42]. It is therefore necessary to incorporate other complementary information in the form of body [16, 18, 23, 42], context [15, 19, 20] or attribute level cues [16] to improve the recognition performance.

Most existing approaches, including the ones mentioned above follow a multi-step process, which consists of feature extraction from multiple body regions followed by their aggregation at the feature or decision stage. Several body regions are regressed from head ground truth of an image and a convolutional neural network (CNN) is trained for each of these regions as shown in Figure 1(a). During testing, features are extracted from region specific CNNs and the results are combined.

There are two shortcomings with this approach. The first is related to the training of an ensemble of models. It is not only sub-optimal to train several region specific models that are aimed at the same task of identity prediction, but is also slower and require extremely large number of parameters. For instance, `PIPER` [42] and `naeil` [16] train 107 and 17 models with $\sim$6 billion and $\sim$1 billion parameters, respectively, corresponding to different body regions and attributes. Clearly, such a solution with multiple models is not suitable for memory-constrained applications. The other issue is related to the fusion of features or classifier scores obtained by multiple models. A stand-alone fusion scheme based on naive strategies such as concatenation [4, 16] or global weights estimated from validation sets [23, 42] may not be effective as it weighs each body feature across all the instances constantly irrespective of whether it is informative or not. In an ideal case, weights should be adaptive

according to the quality of the region (illumination, resolution, visibility, *etc*.) and its discriminative ability.

In this paper, we propose an end-to-end person recognition architecture (N2NPR) in which a single convolutional neural network is designed to produce a compact representation. Our architecture consists of three essential blocks as shown in Figure 1(b). The first block takes a complete person image as input and produces convolutional feature maps. These feature maps are *shared* across multiple body regions and hence require only a single forward pass to compute. In the second block, features are *pooled* from multiple locations of the shared feature maps according to fixed and predetermined region of interest (ROI) inputs to obtain an intermediate representation for each body region. Finally, the *aggregation* block maps the pooled features into a low dimensional space and combines them with adaptive weights produced by the attention module. The parameters of these modules can be trained in an end-to-end manner with other network parameters.

We evaluate our proposed approach on three challenging person recognition datasets namely PIPA [16], Hannah movie [28] and Soccer broadcast video [18]. Our results suggest that a single end-to-end model produces recognition performance that is significantly better than the previous methods with large ensemble of models, even outperforming the state-of-the-art approaches. We also show comparisons w.r.t parameter size and required computations, and highlight how our model is better suited for practical applications.

The salient features and contributions are the following:

- We propose an end-to-end architecture that treats that problem of person recognition as a whole to produce a compact representation.
- We propose a unified framework based on a ROI feature pooling and a novel aggregation scheme to combine features from multiple body regions.
- Our N2NPR model achieves state-of-the-art performance on three person recognition benchmarks with least memory and computations.

## 2. Related Work

**Face recognition and verification** are the widely studied problems related to person identification. There is a vast literature focused on hand-crafted features [2], metric learning [31], sparse representations [38] and current state-of-the-art deep representations [6, 30, 35, 37].

**Person recognition** that uses face together with additional cues is another popular direction employed in unconstrained settings with cluttered background such as photo albums and entertainment videos. In addition to face, various domain-specific cues such as meta-data, clothing, skin and hair [3, 32, 41], sub-titles [8], audio [36], camera pose and timestamps [9] and sport jersey numbers [5] are exploited to improve the performance. The problem has received increased attention in the recent years due to the creation of large scale person identification dataset PIPA [42]. Since its introduction, several approaches [16, 18, 19, 20, 23, 42] have been proposed for person recognition with additional cues.

Recent solutions have focused on CNN representation learning for different body regions and attributes. PIPER [42] trains as many as 107 models with poselet patches, naeil [16] uses features from 17 models optimized for identity and attribute prediction tasks. Kumar *et al*. [18] on the other hand learn pose-specific representations. Their individual pose models though are optimized jointly for multiple regions do not share any computations. A new loss function is proposed in [23] instead of traditional softmax loss that encourages large inter-class separations in the feature space. Several contextual relations that exist between instances in family albums are exploited in [15, 19, 20]. Compared to these approaches, ours is a more practical solution that uses a single end-to-end model and doesn't require any contextual information.

**Person re-identification** deals with matching pedestrians captured from non-overlapping camera views in video-surveillance applications. Existing works primarily focus on metric learning [11, 14] with hand-crafted [22, 25] or deep learning [1, 43] based features to handle variations related to view-point, pose and appearance.

**Deep learning** has achieved tremendous success in the recent years progressing immensely the "3R's of computer vision" [26]. Our work in particular draws inspiration from ROI pooling [10, 12, 29] and aggregation techniques employed in image-set recognition [24, 39, 40]. We however apply these two independent ideas to obtain person representation in an end-to-end manner. The adaptive weights are used to obtain image-set representations by aggregating individual image features. In these approaches, a scoring module generates the weights that are indicative of the visual or content quality of the individual images in the set. This is particularly useful in video face recognition where the impact of noisy predictions made by blurred or poor quality faces are reduced with adaptive weights. The attention schemes have also been applied to sequential tasks [33] and multi-modal feature fusion [27].

We finally note that our approach has some resemblance to bilinear pooling technique [21] employed in fine-grained visual recognition where the features from multiple regions of an object are pooled and converted to a single vector representation.

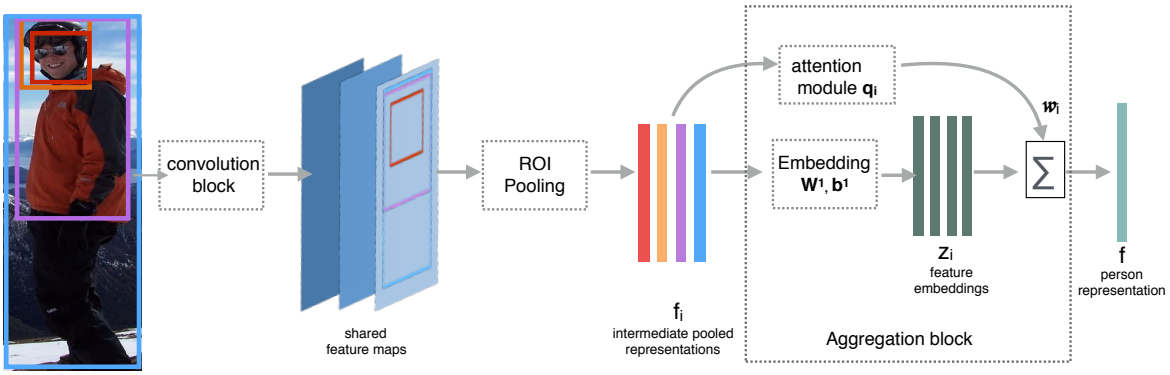


Figure 2: Our proposed end-to-end person recognition architecture. The input person image along with locations of body regions are passed through a convolution block to obtain shared feature maps. Multiple features are then pooled from various body locations, converted into compact embeddings and aggregated through an attention mechanism.

## 3. End-To-End Person Recognition

Given a full person image, we are interested in correctly predicting the person's identity by combining multiple cues such as head, upper body *etc*. As person recognition is usually evaluated in open-set protocol where the instances available during training are disjoint from gallery and probe instances, CNNs are primarily optimized for representation learning. However, training a separate CNN for every region as done in previous works [4, 15, 16, 19, 23, 42] is not scalable when more regions have to be considered.

In our work, we focus on representation learning for entire person image in a unified framework. Our objective is to encode information from multiple body regions in an efficient manner to produce a compact and powerful representation with limited memory and computations.

### 3.1. Architecture

Our proposed end-to-end person recognition (`N2NPR`) architecture is shown in Figure 2. The input to our network consists of full body image of a person and pre-determined bounding boxes (§ 3.2) of regions considered for identification. It consists of three blocks namely: a *convolution* block (§ 3.3) that produces shared convolution feature maps for the input image, a ROI *pooling* layer (§ 3.4) that generates intermediate representations for each region, and an *aggregation* block (§ 3.5) that produces region embeddings and combines them with adaptive weights so as to produce the final representation.

### 3.2. Body Regions

An important consideration in person recognition literature is the choice of body regions. Some of the choices include poselets [42], predefined important body regions either regressed from head [16, 18, 19] or predicted by object detectors [23]. In this work, we consider *face*, *head*, *upper body* and *body*, and compute these regions roughly from the head ground-truths similar to [16]. Given head co-ordinates $(x_h, y_h, w_h, h_h)$, we obtain the co-ordinates corresponding to face, upper-body and body regions as $(x_h + 0.1l, y_h + 0.3l, 0.8l, 0.7l)$, $(x_h - 0.5l, y_h, 2l, 3l)$ and $(x_h - 0.5l, y_h, 2l, 4l)$, respectively with $l = min(w_h, h_h)$.

### 3.3. Convolution Block

The convolutional block which forms the backbone of our network accepts the person image and produces shared feature maps through a series of spatial convolution, relu and pooling operations. It can be built upon any modern and powerful deep CNNs such as Inception [34], ResNet [13]), *etc*. To enable fair comparison with previous works that employed varied architectures (AlexNet [17] in [18, 42], Inception in [4], a combination of AlexNet and DeepFace [35] in [42], Inception and ResNet in [23]), we consider two variants for the convolutional block - a simpler (AlexNet) and more advanced (Inception) architecture. We however make certain modifications for our purpose that are detailed later.

### 3.4. ROI Pooling

Based on the observation that body regions considered above are highly overlapping (*i.e*. upper body also contains head and facial regions) and CNNs operating on these patches are optimized for the common goal of identity prediction, it would be more efficient to share computations across regions. It reduces the network parameters, training and testing times significantly as it avoids forward and backward pass for every patch. Such computation sharing scheme is already popular in object detection [10], classification [21] and segmentation tasks [12]. Our experiments demonstrate that region pooling is equally effective for full body person identification, and can achieve performance on par with separately trained models (§ 4.3).

Given shared convolutional feature maps and bounding

box locations of different regions, pooling layer simply takes the features inside the bounding box and applies adaptive average pooling to produce fixed size feature maps. The window size required for pooling is adjusted according to the ratio of each ROI region and the desired output size which is $6 \times 6$ for AlexNet and $1 \times 1$ for Inception. Following [10], we perform ROI pooling after the last convolution operation (`Conv_5` for AlexNet and `Mixed_7c` for Inception) of the convolutional block. For a person image of size ($W \times H$), the convolution block produces shared feature maps of size ($\sim W/17$, $\sim H/17$). After pooling, it results in intermediate representations $d_1$ of size 2048 and 9216 dimensions, respectively for Inception and AlexNet networks.

### 3.5. Adaptive Feature Aggregation

Once the representations are computed for different body regions, the next task is to combine them using a fusion strategy. As described earlier, there are two commonly employed approaches. The first is to combine features through naive pooling strategies such as concatenation, max or average pooling, and then to train an identity classifier [16, 18, 19] on the aggregated feature. The second option is to combine the prediction scores of several classifiers with global weights obtained from a validation set [23, 42].

Fusion with such global weights will result in noisier representation when instances have large variations in pose, occlusion, *etc*. Instead we consider adaptive weights to combine different features according to their quality such as resolution, pose, occlusion *etc*. For instance, a frontal head can be considered of relatively better quality and assigned high weight compared to other head instance that has poor resolution, occlusion or appears in back-view.

We propose an aggregation mechanism to combine intermediate region features through *attention* and *embedding* modules. These modules are used in conjunction with other blocks described earlier, and have additional differentiable parameters that can be learnt in an end-to-end manner.

Formally, given a set of $n$ intermediate representations $S = \{f_1, f_2, \ldots, f_n\}$, $f_i \in \mathcal{R}^{d_1}$ corresponding to $n$ body regions of a person image following ROI pooling operation, the final representation $f$ is obtained as

$$f = \frac{\sum_{i=1}^{n} w_i \alpha(z_i | f_i, \theta_i)}{\sum_{i=1}^{n} w_i} \tag{1}$$

where $w_i$ are the attention weights and $\alpha$ is the embedding function with parameter $\theta_i$. The **embedding** function $\alpha$ essentially projects the intermediate features $f_i$ to an embedding space $z_i$ through two operations:

$$e_i = W_i f_i + b_i, \tag{2}$$

$$z_i = \frac{e_i}{||e_i||_2} \tag{3}$$

Figure 3: Examples showing adaptive weights of instances obtained during training on PIPA. Notice how weights shown below for head, face, upper body and body corresponds to discriminative ability of the image patches.

where $\theta_i = (b_i, W_i)$, $b_i \in \mathcal{R}^{d_2}$, $W_i \in \mathcal{R}^{d_1 \times d_2}$ are the region specific parameters. The main purpose of the embedding module is to transform shared features into a compact embeddings with unit-norm. The parameters ($W_i$, $b_i$) can be viewed as region-specific "local experts" that learn the mapping from shared feature maps to more discriminative region representations.

Similarly, the **attention** module takes the features $f_i$ and produces the scalar weights $w_i$ ([0-1]) required for aggregations by computing the dot product between $f_i$ and parameter $q_i$. This can be denoted as

$$w_i = \sigma(q_i^T f_i), \tag{4}$$

where $q_i \in \mathcal{R}^{d_1}$ is the attention parameter that decide the quality of each body region $i$.

The weights $w_i$ obtained for a few images during training are shown in Figure 3. It is clear from the examples that the weights correspond to the discriminative ability of the body regions. Whenever a particular region is occluded or not clearly visible, the corresponding weights are small. For images that have visible head region with near frontal pose, the head and face weights are larger compared to other regions. Also, upper body and body regions tend to get large values only when the head is less informative (e.g., $2nd$, $4th$ and $8th$ images in the first row). As expected, the weights of face and head region are highly correlated and so are the upper body and body weights as they contain large overlapping regions.

Figure 4: Images from PIPA (top), movie gallery and probe set (middle: left and right) and soccer (bottom) datasets.

# 4. Experiments and Results

## 4.1. Datasets

We consider three challenging person recognition datasets based on family albums, movie and sport videos. Few images from these datasets are shown in Figure 4.

**People in photo-albums (PIPA)** [42] consists of photo-albums created from user-uploaded photos in Flickr website. The images capture day-to-day lives and important family events of people, as a result contain cluttered background, pose, lighting and resolution variations. It consists of a total 37,107 photos containing 63,188 instances belonging to 2,356 users, which are further divided into four sets namely train, validation, test and left-over splits. While the train split is primarily used for training CNNs, recognition accuracies are reported on test set which consists of 6,443 gallery ($\texttt{fold}_0$) and probe ($\texttt{fold}_1$) instances belonging to 581 different subjects. We also evaluate on additional splits proposed by Oh *et al.* [16] that contain gallery and probe instances from different albums, time and day.

**Hannah movie dataset** [18, 28] consists of face bounding boxes of people appearing in full length movie "Hannah and Her Sisters" movie [28]. Although originally created for face detection, Kumar *et al*. [18] created a person recognition protocol by creating another gallery set from IMDB images. We follow the same protocol. The IMDB gallery set consists of 2,385 images belonging to 26 prominent actors appearing in the movie and the probe set consists of total 159,458 instances belonging to 41 actors. Unlike PIPA, the dataset has additional challenges due to motion blur, age and domain mismatch between IMDB and movie instances.

**Soccer dataset** [18] consists of broadcast video frames from World cup 2014 final game played between Argentina and Germany. The dataset consists of 37 replay clips with an average duration of 30 seconds. The soccer instances exhibit heavy occlusion and body deformations due to the fast movement of players during the game. It consists of 28 subjects namely 13 Germany players, 14 Argentina players and a referee. We use the same gallery/probe set created by [18] for evaluation.The gallery set consist of 10 clips totaling 19,813 instances and the probe set consists of 27 clips totaling 51,051 instances. While the above datasets provide head boxes, soccer dataset provides fully body ground-truths. We therefore obtain head boxes with [7] to create a consistent recognition setting.

## 4.2. Implementation

As mentioned earlier, we conduct our experiments with AlexNet [17] and Inception-v3 [34] backbones whose results are reported with a suffix $\texttt{-A}$ and $\texttt{-I}$. Unless otherwise mentioned, we consider identical training parameters for both the variants. We consider person crops of size $600 \times 300$ and ensure that head has a dimension of $150 \times 150$. The dimension of embedding space $d_2$ is fixed to 1024 for all our experiments, including models trained on individual body regions. We initialize the models with ImageNet weights and pre-train on a subset of VGGFace2 [6] dataset to avoid overfitting. The pre-training dataset consists of 863K samples obtained by randomly sampling a maximum of 100 images for each class in the VGGFace2 dataset. We note that similar external data is used in the previous works [16, 18, 23]. We finally fine-tune the model on PIPA trainset consisting of 29,223 instances. We augment the dataset by random horizontal flipping during training.

Our implementation is based on PyTorch. We optimize for softmax loss using stochastic gradient descent with a batch size of 25. We set momentum to 0.9 and weight decay to 0.0005. The learning rate is set to $5\text{e}{-4}$ and $1\text{e}{-3}$ for Inception and AlexNet backbones, respectively and decreased by a factor of 5 after every 4 epochs. Models are trained for 10 and 20 epochs during pre-training and fine-tuning, respectively. During testing, we extract deep features from the original image and its horizontal flipped image and concatenate the feature for all the models. We train an SVM classifier with gallery features with parameter $c$ set to 1.

## 4.3. Ablation Study

We conduct ablation studies on PIPA test splits with Inception backbone and train four types of models - the baseline models (A) trained on individual region inputs as in previous works, models (B) trained for individuals regions with ROI pooling of person images, end-to-end models (C) and (D) trained without and with an embedding layer, respectively for various fusion schemes. For end-to-end models, we consider two other fusion techniques based on average ($\texttt{Avg}$) and max ($\texttt{max}$) pooling along with our proposed

| Type | Method | Feature | Original | Album | Time | Day |
|---|---|---|---|---|---|---|
| (A) | Head | h | 86.14 | 80.12 | 72.56 | 58.98 |
| | Face | f | 82.98 | 78.78 | 70.14 | 55.19 |
| | Upper body | u | 85.36 | 80.75 | 71.56 | 55.34 |
| | Body | b | 81.35 | 78.33 | 70.73 | 52.15 |
| | Fusion | h+f+u+b | 89.18 | 83.65 | 76.13 | 60.95 |
| | | hofoubo | 88.23 | 82.34 | 75.89 | 60.12 |
| (B) | RCNN | $h_p$ | 84.45 | 77.62 | 72.32 | 57.50 |
| | | $f_p$ | 81.16 | 76.25 | 71.85 | 55.93 |
| | | $u_p$ | 85.83 | 80.17 | 70.44 | 54.11 |
| | | $b_p$ | 81.35 | 78.33 | 70.73 | 52.15 |
| (C) | RCNN+Avg | $h_p+f_p+u_p+b_p$ | 85.62 | 81.04 | 73.77 | 58.75 |
| | RCNN+max | $h_p \circ f_p \circ u_p \circ b_p$ | 85.12 | 80.85 | 72.61 | 57.26 |
| | RCNN+Att | $h_p \oplus f_p \oplus u_p \oplus b_p$ | 87.35 | 82.98 | 75.59 | 60.45 |
| (D) | RCNN+Emb+Avg | $h_p+f_p+u_p+b_p$ | 87.90 | 83.04 | 76.44 | 59.19 |
| | RCNN+Emb+max | $h_p \circ f_p \circ u_p \circ b_p$ | 85.39 | 80.93 | 74.04 | 60.08 |
| | RCNN+Emb+Att (N2NPR-I) | $h_p \oplus f_p \oplus u_p \oplus b_p$ | **89.68** | **84.75** | **79.25** | **63.74** |

Table 1: Ablation study showing the recognition accuracy (%) of various training and fusion techniques on PIPA test splits. The subscript $p$ refers to ROI pooled features and symbols +, $\circ$, $\oplus$ denote average, max and adaptive fusion, respectively.

aggregation technique (Att). We did not consider concatenation as it consumes more memory during training.

The results are shown in Table 1. As observed already in previous works, we see significant improvement in overall performance after combining features from multiple body regions, irrespective of the training and fusion technique. We further observe that the individual region features (B) obtained through ROI pooling have similar performance compared to features (A) obtained by models trained on region (head, upper body, *etc*.) patches. End-to-end models have performance on par with the ensemble of models (A) even with naive fusion strategies. Among the three fusion strategies, max pooling produces the least performance since it throws away information. The proposed technique with adaptive weights outperforms the global weights of average pooling. Finally we observe that embedding layer that learns region-specific mappings improve the performance for all the fusion techniques.

### 4.4. Comparisons with State-of-the-arts

While our primary objective is to show the effectiveness of end-to-end person recognition over separately trained models, we provide comparisons with recent state-of-the-arts in Table 2 for completeness. We note that the key ingredients in these algorithms differ in their choice of architecture, use of contextual information, external data and loss function. As a result, we mention system details of various algorithms to make comparisons more meaningful.

We make the following observations from the table. Approaches [15, 19, 20] modeling the contextual relations among instances achieve good performance even with fewer models. However, this is not the objective of our work. Amongst approaches that focus on learning representations, results are reported majorly with AlexNet [16, 18, 42] and Inception [4, 23] architectures. With AlexNet backbone, our results (N2NPR-A) are significantly better than PIPER and naeil that train 107 and 17 models, and slightly worse than PSM which trains 8 pose-specific models.

We achieve the best results except original split with Inception backbone. Our single end-to-end model outperforms [4, 23] that train 4 models for each body region. Finally, we note that irrespective of the architecture choice, our models require significantly less parameters compared to the the previous approaches and hence is better suited for practical applications.

### 4.5. Results on Video Datasets

We now show results on newly introduced Hannah and soccer datasets for which we make comparison with two previously reported results naeil and PSM [18]. We follow the same protocol as [18], and report both frame and track level accuracies. The track predictions are obtained by majority voting of the labels. The results are shown in Table 4 and Table 5 for movie and soccer dataset, respectively. Both of our models outperform the previous methods and achieve state-of-the-art results.

### 4.6. Training and Testing Time

N2NPR models are faster due to reduced number of forward and backward operations. Table 3 compares training (fine-tuning) and testing times of N2NPR-A with other approaches. We note that a similar conclusion can be drawn


Figure 5: Visualization of "quality" of body regions for images from PIPA day split according to adaptive weights ($w_i$). Each row from top to bottom is obtained by sorting face, head, upper body and body weights, respectively. When face and head regions are not visible, low weights (leftmost images in the first two rows) are assigned. Similarly, low weights are assigned to upper body and body whenever the illumination is poor and images are blurred (leftmost images in the last two rows).

| Method | Accuracy Original | Accuracy Album | Time | Day | Number of models | Architecture / parameters | Total parameters | Context |
|---|---|---|---|---|---|---|---|---|
| Li *et al*. [20] | 84.93 | 78.25 | 66.43 | 43.73 | 2 | VGG (130M) | ∼260M | ✓ |
| Li *et al*. [19] | 88.78 | 83.33 | 77.00 | 59.35 | 2 | deepid (15M) + $\mathcal{A}$ (60M) | ∼75M | ✓ |
| RANet [15] | 89.73 | **85.33** | **80.42** | **67.16** | 4 | $\mathcal{RN}$ (25M) | ∼100M | ✓ |
| PIPER [42] | 83.05 | - | - | - | 107 | $\mathcal{A}$ (60M) | ∼6B | ✗ |
| naeil [16] | 86.78 | 78.72 | 69.29 | 46.61 | 17 | $\mathcal{A}$ (60M) | ∼1B | ✗ |
| PSM [18] | 89.21 | 82.73 | 74.84 | 56.73 | $8 \times 2$ | $\mathcal{A}$ (60M) | ∼1B | ✗ |
| N2NPR-A | 87.23 | 80.98 | 71.52 | 50.23 | 1 | $\mathcal{A}$ (60M) | ∼60M | ✗ |
| Liu *et al*. [23] | **92.78** | 83.53 | 77.68 | 61.73 | $3+1$ | $\mathcal{IN}$ (20M) + $\mathcal{RN}$ (45M) | ∼105M | ✗ |
| Zoom-RNN [4] | 90.88 | 84.40 | 76.44 | 56.92 | 4 | $\mathcal{IN}$ (20M) | ∼80M | ✗ |
| N2NPR-I | 89.68 | 84.75 | 79.25 | 63.74 | 1 | $\mathcal{IN}$ (20M) | ∼20M | ✗ |

Table 2: Detailed comparison of approaches that evaluated previously on PIPA dataset. We compare the approaches in terms of architecture choice, use of context, number of models and parameters. Our end-to-end model N2NPR-I achieves best results on three splits without using any context. $\mathcal{A}, \mathcal{IN}, \mathcal{RN}$ denote AlexNet, Inception and ResNet-50 architectures. M and B indicate million and billion respectively. The numbers in **bold** and with underline indicate top result reported on the dataset and our result.

for N2NPR-I. For all the other approaches that require model ensemble training, we computed the total time based on time required for training/testing each region. For previous approaches which use images of size $224 \times 224$, it takes $\sim$1.5 milli-seconds (ms) for one image patch with AlexNet architecture while our input image $600 \times 300$ takes $\sim$3.5ms for four body regions for one forward pass. Our approach which benefits from computation sharing across regions runs faster than the previous approaches in both training and testing stages.

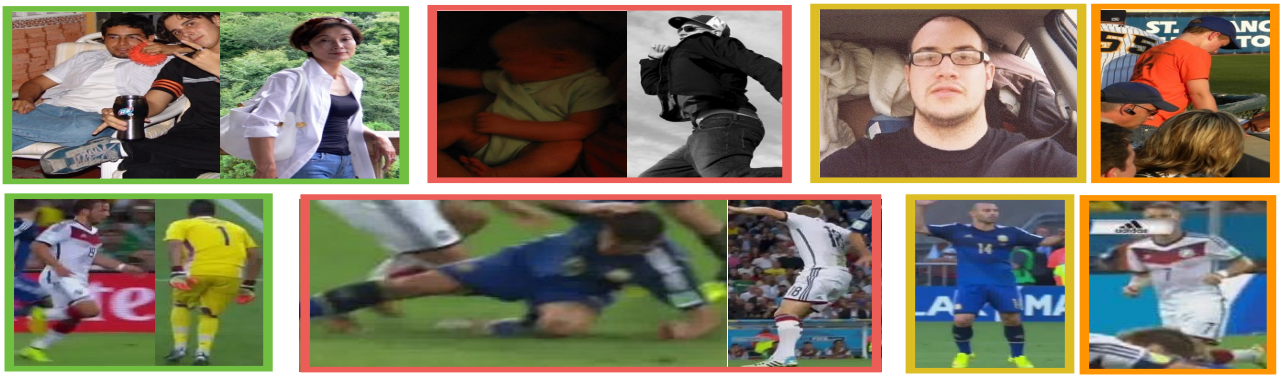

Figure 6: Success and failure cases on (top) PIPA and (bottom) soccer datasets. The images in green and red are our success and failure cases, respectively. Images in yellow and orange are the success cases where the improvement is obtained due to high weights assigned to head and upper body regions.

| Method | Train time (hrs) | Train speedup | Test* time (ms) | Test speedup |
|---|---|---|---|---|
| PIPER [42] | 642 | 64x | 160 | 32x |
| naeil [16] | 102 | 10x | 25 | 5x |
| PSM [18] | 84 | 8x | 20 | 4x |
| Liu *et al*. [23] | 40 | 4x | 60 | 12x |
| N2NPR-A | 10 | 1x | 5 | 1x |

Table 3: Run time comparison of various approaches. *Time measured on a GTX 1080 Ti GPU with 10 CPUs

| Method | Accuracy w/o tracks | Accuracy with tracks |
|---|---|---|
| naeil [16] | 31.41 | 37.57 |
| PSM [18] | 40.95 | 44.46 |
| N2NPR-A | **41.66** | **47.75** |
| Head ($h_p$) | 49.80 | 60.63 |
| Face ($f_p$) | 40.53 | 44.82 |
| Upper body ($u_p$) | 46.84 | 52.96 |
| Body ($b_p$) | 48.45 | 58.45 |
| N2NPR-I | **56.97** | **67.46** |

Table 4: Recognition performance (%) of various approaches on the Hannah dataset with AlexNet (top) and Inception (bottom) backbone.

### 4.7. Qualitative Results

We visualize how the weights obtained by attention mechanism are indicative of the "quality" of body regions in Figure 5. We sort the scores in ascending order for face, head, upper body and body regions, and show their corresponding person images. It is clear from Figure 5 that whenever face and head is not visible, low weights are assigned to them. Similarly, upper body and body regions are assigned lower weights when the illumination is poor or images are

| Method | Accuracy w/o tracks | Accuracy with tracks |
|---|---|---|
| naeil [16] | 20.15 | 23.77 |
| PSM [18] | 20.48 | 24.31 |
| N2NPR-A | **22.62** | **29.07** |
| Head ($h_p$) | 22.14 | 35.04 |
| Face ($f_p$) | 24.26 | 39.32 |
| Upper body ($u_p$) | 26.34 | 37.32 |
| Body ($b_p$) | 25.59 | 37.43 |
| N2NPR-I | **31.88** | **45.45** |

Table 5: Recognition performance (%) of various approaches on the Soccer dataset with AlexNet (top) and Inception (bottom) architecture.

blurred. We finally show few success and failure cases of our approach in Figure 6. The failure cases shown here correspond to severe occlusions and blur while the successful cases show the effectiveness of the adaptive weights to focus on the discriminative aspects of a test image.

## 5. Conclusion

In this work, we propose a unified network architecture for person recognition from multiple body regions. Our approach operates on full person images and produces shared activation maps from which several features are pooled and aggregated in an end-to-end manner. The fusion weights produced by the attention modules correspond to the quality of body region and thus produces better representations in unconstrained scenarios. Our end-to-end approach outperforms previous approaches on various person recognition benchmarks with least memory and computations, and and hence is suited for practical applications.

## References


[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 2

[2] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *PAMI*, 2006. 2

[3] D. Anguelov, K.-c. Lee, S. B. Gokturk, and B. Sumengen. Contextual identity recognition in personal photo albums. In *CVPR*, 2007. 2

[4] S. M. Azar, S. Azami, M. G. Atigh, M. Javadi, and A. Nickabadi. Zoom-rnn: A novel method for person recognition using recurrent neural networks. *arXiv preprint arXiv:1809.09189*, 2018. 1, 3, 6, 7

[5] M. Bertini, A. Del Bimbo, and W. Nunziati. Player identification in soccer videos. In *SIGMM workshop on Multimedia information retrieval*, 2005. 2

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018. 2, 5

[7] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018. 5

[8] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automated naming of characters in tv video. *Image and Vision Computing*, 2009. 2

[9] R. Garg, S. M. Seitz, D. Ramanan, and N. Snavely. Where's waldo: Matching people in images of crowds. In *CVPR*, 2011. 2

[10] R. Girshick. Fast R-CNN. In *ICCV*, 2015. 2, 3, 4

[11] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*. Springer, 2014. 2

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 2, 3

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[14] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012. 2

[15] Q. Huang, Y. Xiong, and D. Lin. Unifying identification and context learning for person recognition. In *CVPR*, 2018. 1, 2, 3, 6, 7

[16] S. Joon Oh, R. Benenson, M. Fritz, and B. Schiele. Person recognition in personal photo collections. In *ICCV*, 2015. 1, 2, 3, 4, 5, 6, 7, 8

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3, 5

[18] V. Kumar, A. Namboodiri, M. Paluri, and C. Jawahar. Pose-aware person recognition. In *CVPR*, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[19] H. Li, J. Brandt, Z. Lin, X. Shen, and G. Hua. A multi-level contextual model for person recognition in photo albums. In *CVPR*, 2016. 1, 2, 3, 4, 6, 7

[20] Y. Li, G. Lin, B. Zhuang, L. Liu, C. Shen, and A. v. d. Hengel. Sequential person recognition in photo albums with a recurrent network. *CVPR*, 2017. 1, 2, 6, 7

[21] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear CNN models for fine-grained visual recognition. In *ICCV*, 2015. 2, 3

[22] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important? In *ECCV*, 2012. 2

[23] Y. Liu, H. Li, and X. Wang. Rethinking feature discrimination and polymerization for large-scale recognition. In *NIPS*, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[24] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017. 2

[25] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV*, 2012. 2

[26] J. Malik, P. Arbeláez, J. Carreira, K. Fragkiadaki, R. Girshick, G. Gkioxari, S. Gupta, B. Hariharan, A. Kar, and S. Tulsiani. The three R's of computer vision: Recognition, Reconstruction and Reorganization. *PR Letters*, 2016. 2

[27] A. Miech, I. Laptev, and J. Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 2

[28] A. Ozerov, J.-R. Vigouroux, L. Chevallier, and P. Pérez. On evaluating face tracks in movies. In *ICIP*, 2013. 2, 5

[29] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 2

[30] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2

[31] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *BMVC*, 2013. 2

[32] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *BMVC*, 2006. 2

[33] S. Sukhbaatar, J. Weston, R. Fergus, et al. End-to-end memory networks. In *NIPS*, 2015. 2

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 3, 5

[35] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 2, 3

[36] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. In *CVPR*, 2012. 2

[37] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 2

[38] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI*, 2009. 2

[39] W. Xie and A. Zisserman. Multicolumn networks for face recognition. *arXiv preprint arXiv:1807.09192*, 2018. 2

[40] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *CVPR*, 2017. 2

[41] R. B. Yeh, A. Paepcke, H. Garcia-Molina, and M. Naaman. Leveraging context to resolve identity in photo albums. In *JCDL*, 2005. 2

[42] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *CVPR*, 2015. 1, 2, 3, 4, 5, 6, 7, 8

[43] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 2