Pose-Aware Person Recognition

Vijay Kumar * Anoop Namboodiri * Manohar Paluri † C. V. Jawahar *
* CVIT, IIIT Hyderabad, India † Facebook AI Research

Abstract

Person recognition methods that use multiple body regions have shown significant improvements over traditional face-based recognition. One of the primary challenges in full-body person recognition is the extreme variation in pose and view point. In this work, (i) we present an approach that tackles pose variations utilizing multiple models that are trained on specific poses, and combined using pose-aware weights during testing. (ii) For learning a person representation, we propose a network that jointly optimizes a single loss over multiple body regions. (iii) Finally, we introduce new benchmarks to evaluate person recognition in diverse scenarios and show significant improvements over previously proposed approaches on all the benchmarks including the photo album setting of PIPA.

Figure 1: (a) Different body regions provide cues that help in recognition. (b) Each row shows the appearance of a person in different poses. Note that the distinguishing features of a person that help classification are different in each pose.

1. Introduction

People are ubiquitous in our images and videos. They appear in photographs, entertainment videos, sport broadcasts, and surveillance and authentication systems. This makes person recognition an important step towards automatic understanding of media content. Perhaps, the most straight-forward and popular way of recognizing people is through facial cues. Consequently, there is a large literature focused on face recognition [25, 56]. Current face recognition algorithms [33, 35, 39, 40] achieve impressive performance on verification and recognition benchmarks [20, 44], and are close to human-level performance.

Face based recognition approaches require that faces are visible in images, which is often not the case in many practical scenarios. For instance, in social media photos, movies and sport videos, faces may be occluded, be of low resolution, facing away from the camera or even cropped from the view (Figure 1(a)). Hence it becomes necessary to look *beyond faces* for additional identity cues. It has been recently demonstrated [23, 26, 53] that different body parts provide complementary information and significantly improve person recognition when used in conjunction with face.

One of the major challenges in person recognition or fine-grained object recognition in general is the pose and alignment of different object parts. The appearance of the same object changes drastically with different poses and view-points (rows of Figure 1(b)) causing a serious challenge for recognition. One way to overcome this problem is through pose normalization, where objects in different poses and view-points are transformed to a canonical pose [6, 10, 18, 40, 48, 57]. Another popular strategy is to model the appearance of objects in individual poses by learning view-specific representations [2, 22, 31, 52].

In this work, we aim to learn pose-aware representations for person recognition. While it is straight-forward to align objects such as faces, it is harder to align human body parts that exhibit large variations. Hence we design view-specific models to obtain pose-aware representations. We partition the space of human pose into finite clusters (columns of Figure 1(b)) each containing samples in a particular body orientation or view-point. We then learn multi-region convolutional neural network (convnet) representations for each view-point. However, unlike previous approaches that train a convnet for each body region, we jointly optimize the network over multiple body regions with a single identification loss. This provides additional flexibility to the network to make predictions based on a few informative body regions. This is in contrast to separate training which strictly enforces correct predictions from each body region. Dur-

ing testing, we obtain the identity predictions of a sample through a linear combination of classifier scores, each of which is trained using a pose-specific representation. The weights for combining the classifiers are obtained by a pose estimator that computes the likelihood of each view.

Our approach overcomes some of the limitations of the previously proposed approaches, PIPER [53] and naeil [23]. Although poselet-based representation of PIPER normalizes the pose; individual poselet patches [8] by themselves are not discriminative enough for recognition tasks, and under-perform compared to fixed body regions such as head and upper body. On the other hand, naeil learns a pose-agnostic representation using more informative body regions. Our framework is able to combine the best of both approaches by generating pose-specifc representations based on discriminative body regions, which are combined using pose-aware weights.

Another major contribution of the work is the rigorous evaluation of person recognition. Current approaches [23, 26, 53] have solely focused on the photo albums scenario, reporting primarily on PIPA dataset [53]. However, this setting is very limited due to the similar appearance of people in albums, clothing and scene cues. To create a more challenging evaluation, we consider three different scenarios of photo albums, movies and sports and show significant performance improvements with our proposed approach. The datasets are available at http://cvit.iiit.ac.in/research/projects/personrecognition

2. Related Work

Person recognition has been attempted in multiple settings, each assuming the availability of specific types of information regarding the subjects to be recognized.

Face recognition is by far the most widely studied form of person recognition. The area has witnessed great progress with several techniques proposed to solve the problem, varying from hand-crafted feature design [4, 34, 45], metric learning [17, 36], sparse representations [46, 54] to state-of-the-art deep representations [33, 35, 40].

Person re-identification is the task of matching pedestrians captured in non-overlapping camera views; a primary requirement in video-surveillance applications. Most popular existing works employ metric learning [16, 19, 49] using hand-crafted [13, 28, 30] or data-driven [3, 27, 55] features to achieve invariance with respect to view-point, pose and photometric transformations. The approaches in [42, 49] also optimized a joint architecture with siamese loss on non-overlapping body regions for re-identification.

Pose normalization and **multi-view representation** are the two common approaches in dealing with object pose variations. Frontalization [18, 40, 48, 57] is a pose normalization scheme used commonly in face recognition, where faces in arbitrary poses are transformed to a canonical pose before recogniton. Pose-normalization is also applied to the similar problem of fine-grained bird classification [6, 51]. Unlike rigid objects such as faces, it is difficult to align human body parts due to large deformations. Hence we follow a multi-view representation approach where the objects are modeled independently in different views. This has also been employed in face recognition [2, 31], where training faces are grouped into different poses and pose-aware CNN representations are learnt for each group.

Person recognition with multiple body cues is the problem of interest in this work. We make direct comparisons with the recent efforts that use multiple body cues: PIPER [53], naeil [23] and Li *et al*. [26]. PIPER uses a complex pipeline with 109 classifiers, each predicting identities based on different body part representations. These include one representation based on Deep Face [40] architecture trained on millions of images, one AlexNet [24] trained on the full body and 107 AlexNets trained on poselet patches, the latter two using PIPA [53] trainset. On the other hand, naeil is based on fixed body regions such as face, head and body along with scene and human attribute cues trained using four different datasets, namely PIPA, CASIA [50], CACD [9] and PETA [11]. While poselets (used in PIPER) normalizes the pose, they are less discriminative compared to fixed body regions employed by naeil. We combine the strengths of both approaches using pose-aware representations based on fixed body regions.

Person identification using context is another popular direction of work, where domain-specific information is exploited. Li *et al*. [26] focus on person recognition in photo-albums exploiting context at multiple levels. They propose a transductive approach where a spectral embedding of the training and test examples is used to find the nearest neighbors of a test sample. An online classifier is then trained to classify each test sample. They also exploited photo meta-data such as time-stamp and the co-occurrence of people to improve the performance. In [5, 47], meta-data and clothing information are exploited to identify people in photo collections. Similarly in [15], the authors use timestamp, camera-pose, and people co-occurrence to find all the instances of a specific person from a community-contributed set of photos of a crowded public event. Sivic *et al*. [38] improve the recall by modeling the appearance of cloth, hair, skin of people in repeated shots of the same scene.

People identification in videos may use cues such as sub-title [12] or appearance models [14, 37], in addition to clothing, audio, face [41]. Similarly, a combination of jersey, face identification and contextual constraints are used to identify players in broadcast videos [7, 29].

We focus on the generic person recognition problem similar to [23, 53] that work in diverse settings without using any domain level information and demonstrate the effectiveness of the pose-aware models in different scenarios.

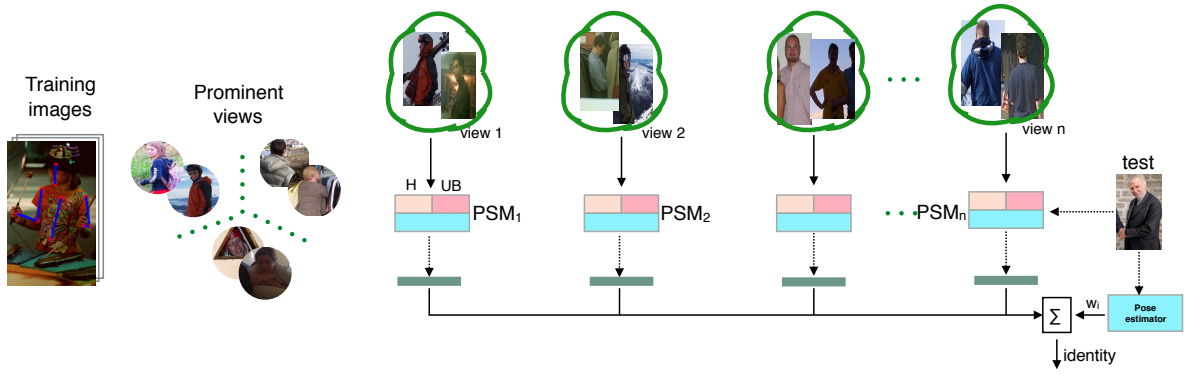

Figure 2: **Overview of our approach:** The database is partitioned into a set of prominent views (poses) based on keypoints. A *PSM* is trained for each pose based on multiple body regions. During testing, predictions from multiple classifiers, each based on a particular *PSM* representation, are obtained and combined using pose-aware weights provided by the pose estimator.

3. Pose-Aware Person Recognition

The primary challenge in person recognition is the variation in pose[1] of the subjects. The appearance of the body parts change significantly with pose. We aim to tackle this by learning pose-specific models (*PSM*s), where each *PSM* focuses on specific discriminative features that are relevant to a particular pose. We fuse the information from different *PSM*s to make an identity prediction.

Our proposed framework is shown in Figure 2. Given a database of training images with identity labels and keypoints, we cluster the images into a set of prominent views (poses) based on keypoint features (§ 3.1). A pose estimator is learned on these clusters for view classification. For learning person representation in each view, a *PSM* is then trained for identity recognition that makes use of multiple body regions (§ 3.2). We train multiple linear classifiers that predict the identities based on *PSM* representations (§ 3.3). Given an input image x, we first compute the pose-specific identity scores, $s_i(y, x)$, each based on the i^{th} *PSM* representation. The final score for each identity y is a linear combination of the pose-specific scores.

$$s(y, x) = \sum_i w_i s_i(y, x), \tag{1}$$

where w_is are the pose-aware weights predicted by the pose estimator (§ 3.1). To allow robustness to rare views with limited training examples, we also incorporate a `base` model in the above equation similar to [53] which is trained on the entire train set, whose scores and weights are referred as $s_o(y, x)$ and w_o respectively. The predicted label of the sample is computed as: $\arg\max_y s(y, x)$.

Our framework differs from `PIPER` in two aspects. First, our pose-aware weights are specific to each instance as opposed to the `PIPER`, which uses fixed weights computed from a validation set. Second, `PIPER` extracts features from a single model for a given localized poselet patch, however, we extract features from different pose models but combine them softly using the pose weights. This allows multiple *PSM*s that are very near in pose space (*e.g.* a semi-left and left) to contribute during the prediction.

[1] In this work, we use the terms pose and view interchangeably. We use these terms to refer the overall orientation of the body with respect to the camera and not the location of keypoints within the body.

3.1. Learning Prominent Views

To facilitate pose-aware representations, we partition the training images into prominent views using body keypoints. Although people exhibit large variations in arm and leg positions, we consider only the informative regions such as head and torso. We construct a 24-D feature for pose clustering using 14 key points and visibility annotations as shown in Figure 3. It consists of -

1. 10-D *orientation feature* based on the relative location of different body parts computed as $[cos(\theta_1), \ldots, cos(\theta_8), sign(x_6 - x_3), sign(x_7 - x_4)]$, where $\theta_i, i = \{1, 2, 3, \ldots, 8\}$ denote the angle between the line joining two key points and the x-axis. For example, θ_6 is the angle between head midpoint and right shoulder. The last two elements distinguish front and back views, and are based on the sign of x-coordinate differences of left and right points of shoulder and elbow, respectively.
2. 14-D *visibility feature*, where each element is either 1 or 0 depending upon whether the corresponding keypoint is visible or not. This provides strong pose cues as certain body parts are not visible in particular views.

To identify meaningful views, we apply k-means algorithm to cluster the images based on the above features. We first obtain a large number (30) of highly similar groups, which are then hierarchically merged to obtain seven prominent views. Figure 1(b) shows an example from each of these views for two different people. The views from left to

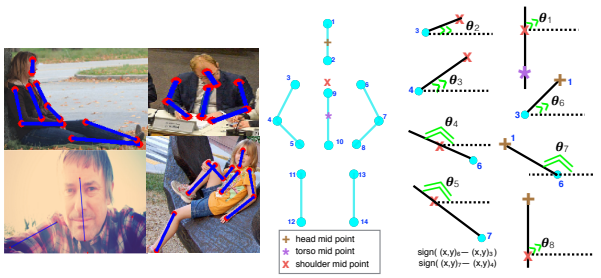

Figure 3: We use body keypoints (left) to learn prominent views using a set of features (right) based on orientation of informative body parts and keypoints.

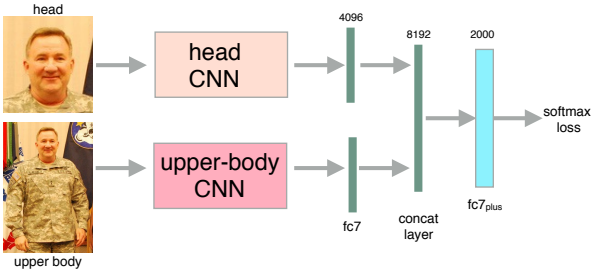

Figure 4: **PSM:** Our network consists of two AlexNets for head and upper body with a single output layer. The last fully connected layer of the two regions are concatenated and passed to an joint hidden layer with 2000 nodes.

right are: `frontal`, `semi-left`, `left`, `semi-right`, `right`, `back` and `partial` views, respectively. The `partial` cluster contains images where only the head and possibly part of shoulders are visible.

Once we obtain the prominent views or poses, we train an pose-estimator based on AlexNet that takes full body image as input and computes the likelihood of each pose. During testing, the pose likelihood estimated from the pose-estimator provide the pose-aware weights w_i, in Eqn. 1. We noticed from our experiments that, it is critical to l_2-normalize the weights to obtain the improved performance.

3.2. Learning a PSM

To train a pose-specific model (*PSM*), we select the training samples that belong to a specific pose cluster. We consider the head and upper body regions as these are the most informative cues for recognition [23, 26]. Given a head at location (l_x, l_y) with dimensions (δ_x, δ_y), we estimate the upper body to be a box at location $(l_x - 0.5\alpha, l_y)$ of dimensions $(2\alpha, 4\alpha)$ where $\alpha = min(\delta_x, \delta_y)$.

Given different body parts, one possibility is to train independent convnets on each of these regions [23, 26, 53]. However, discriminative body regions that help in recognition may vary across training instances. For example, Figure 1(a)-4 contains an occluded face region and is less informative. Similarly, upper body may be less informative in some other instances. If such noisy or less informative regions influence the optimization process, it may reduce the generalization ability of the networks.

We propose an approach to improve the generalization ability by allowing the network to selectively focus on informative body regions during the training process. The idea is to optimize both the head and upper body networks jointly over a single loss function. Our *PSM* contains two AlexNets corresponding to the head and upper-body regions (see Figure 4). The final `fc7` layers of each region are concatenated and passed to a joint hidden layer ($\texttt{fc7}_{\texttt{plus}}$) with 2000 nodes before the classification layer. This provides more flexibility to the network to make the predictions based on one region even if the other region is noisy or less informative. As we show in our experiments (Table 5), the joint training approach performs better than separate training of regions.

3.3. Identity Prediction with PSMs

We derive multiple features from each *PSM* and train classifiers on these feature vectors. The primary feature vector ($\mathcal{F}$) consists of the sixth and seventh layers of the head (`h`) and upper body (`u`), and the joint fully connected layer ($\mathcal{F}$: $<\texttt{fc6}_\texttt{h}, \texttt{fc7}_\texttt{h}, \texttt{fc6}_\texttt{u}, \texttt{fc7}_\texttt{u}, \texttt{fc7}_\texttt{plus}>$). In addition to $\mathcal{F}$, we define two additional feature vectors solely based on the head and upper body layers - $\mathcal{F}_h$: $<\texttt{fc6}_\texttt{h}, \texttt{fc7}_\texttt{h}>$ and $\mathcal{F}_u$: $<\texttt{fc6}_\texttt{u}, \texttt{fc7}_\texttt{u}>$. We train linear SVM classifiers on each of the above feature vectors to obtain the identity predictions. The pose-specific identity score, $s_i(y, x)$, is simply the sum of the three SVM classifier outputs.

$$s_i(y, x) = \sum_{f \in \{\mathcal{F}, \mathcal{F}_h, \mathcal{F}_u\}} P_i(y|f; x), \qquad (2)$$

where $P_i(y|f; x)$ is the class y score of the sample x predicted by the classifier trained on the feature f in i-th view.

4. Experiments and Results

4.1. Datasets and Setup

We select three datasets from the domain of photo-albums, movies and sport broadcast videos as shown in Figure 9. Each of these settings have their own set of advantages and challenges as summarized in Table 1. To the best of our knowledge, this is the first work that evaluates person recognition in such diverse scenarios.

4.1.1 Photo Album Dataset

PIPA [53] consists of 37,107 photos containing 63,188 instances of 2,356 identities collected from user-uploaded photos in Flickr. The dataset consists of four splits with an approximate ratio of 45:15:20:20. The larger split is

Figure 5: Few images from (top) PIPA, (middle) Hannah and (bottom) Soccer datasets.

	PIPA [53]	Hannah [32]	Soccer
Train instances	6,443	2,385	19,813
Train subjects	581	26	28
Test instances	6,443	202,178	51,051
Test subjects	581	41	28
Annotations	Head	Face	Body
Domain variation	No	Yes	No
Clothing	Yes	No	No
Age gap	No	Yes	No
Head resolution	High	Medium	Low
Motion blur	No	Moderate	Severe
Deformation	Less	Moderate	Severe

Table 1: Comparison of the datasets in terms of statistics, annotations, merits and challenges.

primarily used to train convnets, second split to optimize parameters during validation and the third split to evaluate recognition algorithms. The evaluation set is further divided into two equal subsets, each with 6,443 instances belonging to 581 subjects for training and testing the classifiers. We follow PIPA experimental protocol and train the classifiers on one fold and test on another fold, and vice-versa. We also conduct experiments on challenging splits introduced by Oh *et al.* [23] based on album, time and day information.

4.1.2 Hannah Movie Dataset

We consider "Hannah and Her Sisters" dataset [32] to recognize the actors appearing in the movie. The dataset consists of 153,833 movie frames containing 245 shots and 2,002 tracks with 202,178 face bounding boxes. We regress the face annotations to get the rough estimate of head. There are a total of 254 labels of which 41 are the named subjects. The remaining labels are the unnamed characters (boy1, girl1, *etc*.) and miscellaneous regions (crowd, *etc*.).

To consider a more practical recognition setting, we create another dataset for classifier training using IMDB photos. For each named character, we collect photos from actor's profile in IMDB[1] and annotate the head bounding boxes. Of the 41 named characters, only 26 prominent actors had profiles in IMDB. The IMDB train set consists of 2,385 images belonging to 26 prominent actors appeared in the movie. There are a total of 159,458 instances belonging to these 26 actors in the test set. There is a significant age variation between train and test instances since the Hannah instances are created from a particular year (1986) while the IMDB photos are captured over a long period of time.

4.1.3 Soccer Dataset

We create soccer dataset from the broadcast video of World cup 2014 final match played between Argentina and Germany. We considered only replay clips as these capture the important events of the match. We further filtered the replay clips to retain only those clips that are shot in close-up and medium views. We used VATIC toolbox [43] to annotate the players in videos. Our soccer dataset consists of 37 video clips with an average duration of 30 secs. It consists of 28 subjects with 13 players from Germany team, 14 players from Argentina team, and a referee.

Unlike PIPA, we marked full-body bounding boxes for each player since head is not visible or out-of-view in many instances, and it also is difficult to estimate the bounding boxes of different body regions from head, due to large deformations. We followed PIPA annotation protocol and labeled the players regardless of their pose, resolution and visibility. We annotated the players to generate continuous tracks even in the presence of severe occlusion. Whenever it is difficult to recognize the players, we relied on additional clues such as hair, shoes, jersey number and accessories. However, we do not rely on any of these domain-specific cues in this work. For evaluation purposes, we randomly select 10 clips into training and remaining 27 clips into testing. This resulted in 19,813 instances in training set and 51,051 instances in testing set.

4.2. Results and Analysis

For all our experiments, we use the *PSM* models trained on larger set of PIPA consisting of 29,223 instances. We annotate PIPA train instances with keypoint locations to learn prominent views as discussed in § 3.1. The number of instances in each view after pose clustering is shown in Figure 6(a). We train a separate *PSM* on `frontal`, `semi-left`, `left`, `semi-right`, `right`, `back` and `partial` views. The `base` model is trained on the entire PIPA train set. We augment each view by horizontal flipping of the instances from its symmetrically opposite view. For instance, images in left view are flipped and augmented to right view. We use Caffe library [21] for our implementation. For optimization, we use stochastic gradient descent

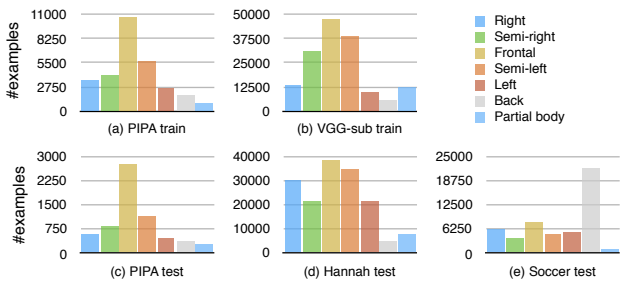

Figure 6: Pose statistics of different datasets.

with a batch size of 50 and momentum coefficient of 0.9. The learning rate is initially set to 0.001, which is decreased by a factor of 10 after every 50,000 iterations. We train the networks for a total of 300,000 iterations. The parameter C is set to 1 for training SVM and the base weight w_0 to 1.

We noticed that the *PSM*s trained on view samples lead to over-fitting. To overcome this, we used a subset of VGGFace dataset [33] for initializing the networks. We extended the face annotations and selected only those instances that have full body in the VGG images. The number of examples used to initialize *PSM*s are shown in Figure 6(b). We make two important points regarding the additional data. First, the extra data (∼160K) we considered is much smaller compared to `PIPER` (∼4M faces for training DeepFace [40]) and `naeil` (∼500K from four different datasets). Second, the proposed improvement is primarily due to the pose-aware combination strategy and not the ensemble of different view-specific models. This is discussed below in the Ablation study (III).

Overall performance: We compare the performance of various approaches on PIPA test splits in Table 2, including both baseline and contextual results of Li *et al*. [26] to provide a comprehensive review. When the contextual information is not considered, our approach outperforms all the previous approaches achieving an 89.05% on the original split. On the Hannah dataset, our approach outperforms `naeil` by a large margin as shown in Table 3. Finally, we show the results on the newly created player recognition in soccer in Table 4.

We notice on all the datasets that use of multiple body regions helps in recognition strengthening the motivation behind `PIPER` and `naeil` algorithms. We also show the results by merging track labels on Hannah and Soccer datasets to understand their impact on recognition. We reassign the frame labels based on simple majority voting of all the frames in a track. The results suggest that track information if available should be used to improve the performance.

The accuracies on Hannah and Soccer datasets are much lower compared to PIPA owing to lower resolution, motion blur, heavy occlusions, and age variations. We also observe a little improvement over `naeil` on soccer dataset due to

Method	Original	Album	Time	Day
`PIPER` [53]	83.05	-	-	-
`naeil` [23]	86.78	78.72	69.29	46.61
Li *et al*. w/o context [26]	83.86	78.23	70.29	56.40
Li *et al*. with context [26]	88.75	**83.33**	**77.00**	**59.35**
Our approach	**89.05**	82.37	74.84	56.73

Table 2: Performance comparison (%) of various approaches on different PIPA splits.

Method	Accuracy without tracks	Accuracy with tracks
Head (`H`)	27.52	31.91
Face (`F`)	26.53	31.55
Upper body (`U`)	16.49	17.72
Separate training of `H` and `U`	31.86	36.10
Joint training of `H` and `U`	32.92	37.74
`naeil` [23]	31.41	37.57
Our approach	**40.95**	**44.46**

Table 3: Recognition performance (%) of various approaches on Hannah movie dataset using IMDB dictionary.

unusual poses (kicking, falling, *etc*.) of the subjects. A large majority of these images are predicted as "back-view" by the pose-estimator (see Figure 6(e)). Since the performance of our `back` view model is poor due to limited training data, we noticed only minimal improvement.

Ablation study (I): We analyze the effectiveness of different features and joint optimization strategy with `base` model in Table 5. The use of both $\texttt{fc}_6$ and $\texttt{fc}_7$ features improve the performance for all the body regions. The head ($\mathcal{F}_h$) and upper body ($\mathcal{F}_u$) features obtained through joint training outperform the head ($\texttt{h}_2$) and upper body ($\texttt{u}_2$) features obtained through separate training by almost two percent points. Similarly, the concatenation of head and upper body features ($\mathcal{F}$) through joint training perform better than separate training ($\mathcal{S}_1$). Finally, the combination of three classifiers ($s_0(y, x)$) from head, upper-body and joint features further bring the performance improvement. We note that, our single `base` model with joint-training strategy and combination of classifiers itself outperforms `naeil`, which reports an accuracy of 86.78% with 17 models.

Ablation study (II): We conduct experiments to measure the performance of pose-specific *PSM* models using PIPA test set. In each experiment, we consider only those examples that are in i-th pose and select half of them randomly for classifier training and remaining half into testing. We extract $\mathcal{F}_u$ feature for these examples using *PSM* and `base` models. Figure 8 shows the performance of each model in recognizing examples from different views. It shows that for `frontal`, `semi-left` and `semi-right` examples, corresponding *PSM* models outperform other models in-

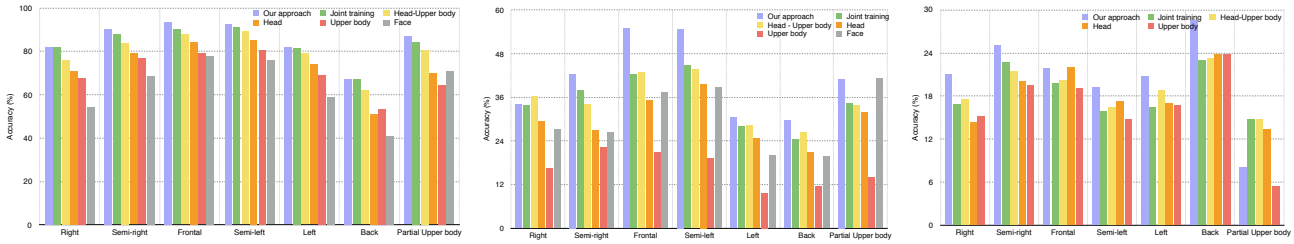

Figure 7: Pose-wise recognition performance on PIPA (left), Hannah (middle) and Soccer (right) datasets.

Method	Accuracy without tracks	Accuracy with tracks
Head (H)	17.68	20.54
Upper body (U)	18.01	19.76
Separate training of H and U	17.62	20.68
Joint training of H and U	18.35	20.18
naeil [23]	19.45	23.77
Our approach	**20.15**	**24.31**

Table 4: Performance comparison (%) on Soccer dataset.

	Feature		Accuracy
Face (F)	$\texttt{fc7}_\texttt{f}$		66.83
	$[\texttt{fc6}_\texttt{f}\ \texttt{fc7}_\texttt{f}]$		70.40
Head (H)	$\texttt{fc7}_\texttt{h}$	$\ldots(\texttt{h}_1)$	76.81
	$[\texttt{fc6}_\texttt{h}\ \texttt{fc7}_\texttt{h}]$	$\ldots(\texttt{h}_2)$	79.54
Upper body (U)	$\texttt{fc7}_\texttt{u}$	$\ldots(\texttt{u}_1)$	72.26
	$[\texttt{fc6}_\texttt{u}\ \texttt{fc7}_\texttt{u}]$	$\ldots(\texttt{u}_2)$	75.19
Separate training of H and U	$[\texttt{h}_1\ \texttt{u}_1]$		82.90
	$[\texttt{h}_2\ \texttt{u}_2]$	$\ldots(\mathcal{S}_1)$	84.01
Joint training of H and U	$\texttt{fc7}_\texttt{plus}$		85.98
	$[\texttt{fc6}_\texttt{h}\ \texttt{fc7}_\texttt{h}]$	$\ldots(\mathcal{F}_h)$	82.22
	$[\texttt{fc6}_\texttt{u}\ \texttt{fc7}_\texttt{u}]$	$\ldots(\mathcal{F}_u)$	77.62
	$[\texttt{fc7}_\texttt{h}\ \texttt{fc7}_\texttt{u}]$		85.05
	$[\texttt{j}_1\ \texttt{fc7}_\texttt{h}\ \texttt{fc7}_\texttt{u}]$		85.22
	$[\texttt{fc7}_\texttt{plus}\ \texttt{fc6}_\texttt{h}\ \texttt{fc6}_\texttt{u}]$		86.10
	$[\texttt{fc7}_\texttt{plus}\ \mathcal{F}_h\ \mathcal{F}_u]$	$\ldots(\mathcal{F})$	86.27
	$s_0(y,x)$		86.96

Table 5: Performance (%) of different features obtained from separate and joint training of regions on PIPA test set.

cluding the base model. This show that the person representations obtained from pose-specifc models are more robust than the pose-agnostic representations. However, for extreme profile views, we noticed that base model performed better than the corresponding profile models. We attribute this to the non-availability of enough profile images while training the *PSM*. For this reason we include the base model along with *PSM*s to bring more robustness when handling rear and non-prominent view images.

Ablation study (III): Our approach uses two kinds of

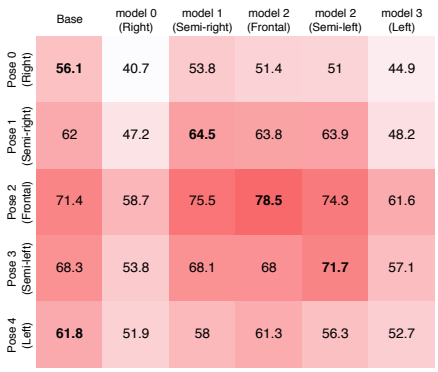

Figure 8: Effectiveness of *PSM*s: Each row shows the performance of test examples in a particular pose represented using the different *PSM*s.

Fusion type	(I)	(II)
Average pooling	83.57%	87.78%
Max pooling	80.54%	85.51%
Elementwise multiplication	81.71%	86.32%
Concatenation	**84.44%**	87.62%
Pose-aware weights	–	**89.05%**

Table 6: Comparison of different fusion schemes for combining (i) features during joint training (using `frontal` *PSM*) and (ii) pose-aware classifier scores during testing.

information pooling, one during *PSM* training and another for combining classifiers. For joint-training, we show the effect of different head ($\texttt{fc7}_\texttt{h}$) and upper body ($\texttt{fc7}_\texttt{u}$) combination strategies in Table 6 (I). The simple concatenation of $\texttt{fc7}_\texttt{h}$ and $\texttt{fc7}_\texttt{u}$ worked better, and hence considered. Similarly, we tried multiple strategies to combine the classifiers during testing. As can be seen in Table 6 (II), pose-aware weighting outperform other strategies including average pooling of the ensemble of classifiers.

Pose-wise recognition performance: The statistics of different poses is given in Figure 6 (bottom) for different datasets. Frontal images dominate PIPA due to which algorithms already achieve high performance (>80%). Hannah consists of different poses in similar proportion while the

Figure 9: Success and failure cases on (top) PIPA and (bottom) Hannah. First five columns show the success cases of our approach where the improvement is primarily due to the specific-pose model. Green and yellow boxes indicate the success and failure result of `naeil` respectively. Last column in red shows the failure cases for both our approach and `naeil`.

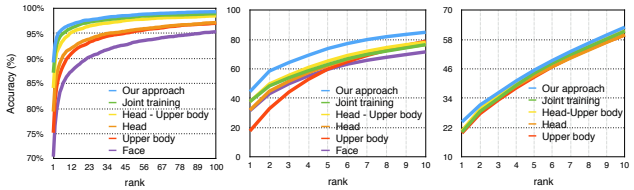

Figure 10: CMC curves of various approaches on PIPA (left), Hannah (middle) and Soccer (right) datasets.

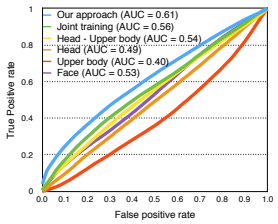

Figure 11: ROC curves of various approaches in rejecting unknown Hannah instances based on normalized prediction scores.

soccer dataset contains majority of back view images. Consequently, we observe a low performance on these datasets. In Figure 7, we show pose-wise recognition performance. PIPA and Hannah have a similar trend in which frontal and semi-profile images are recognized with greater accuracies, while profile and back-views with less accuracy. The upper body seems to be less informative in case of Hannah as the clothing is completely different between Hannah and IMDB. The proportion of back views that are correctly recognized is slightly better in soccer setting due to large number of back view images in the classifier train set.

Rank-n identification rates: The Cumulative matching Characteristic (CMC) curves are shown in Fig 10. Our approach achieves rank-10 accuracies of 96.56%, 84.83% and 63.2% on PIPA, Hannah and Soccer datasets respectively. On Hannah, we noticed a big difference of 12% between rank-1 and rank-2 performance. The performance gap between different approaches tend to reduce with higher rank.

Handling unknown instances: In the movie scenario, the test set has 41 ground truth labels while there are only 26 subjects in the trainset. Therefore, recognition algorithms should have the ability to reject such unknown instances. To achieve this, we l_2-normalized the predicted class scores and considered the maximum score as a confidence measure. The confidence score obtained on pose-aware representations are more robust in rejecting unknown, the performance being measured using ROC curve in Figure 11.

Computational complexity: The number of features extracted from each *PSM* is 18,384 ($4096 \times 4 + 2000$). With 7 pose-aware models and a base model, our total feature dimension is $(18,384 \times 8)$ which is $\sim$3 times smaller than `PIPER` (4096×109) and $\sim$2 times larger than `naeil` (4096×17). For memory critical applications, $\texttt{fc7}_{\text{plus}}$ alone can be used as feature. We achieve an accuracy of 87.01% on PIPA with $\texttt{fc7}_{\text{plus}}$ still outperforming `naeil` with a feature dimension of just 16,000 (2000×8).

5. Conclusion

We show that learning a pose-specific person representation helps to better capture the discriminative features in different poses. A pose-aware fusion strategy is proposed to combine the classifiers using weights obtained from a pose estimator. The person representations obtained using a joint optimization strategy is shown to be more powerful compared to separate training of body regions. We achieve state-of-the-art results on three different datasets from photo-albums, movies and sport domains.

References

[1] Hannah and her sisters (1986), full cast and crew. http://www.imdb.com/title/tt0091167/fullcredits.

[2] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan, et al. Face recognition using deep multi-pose representations. In *WACV*, 2016.

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

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

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

[6] T. Berg and P. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013.

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

[8] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[9] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European Conference on Computer Vision*, 2014.

[10] G. Chéron, I. Laptev, and C. Schmid. P-CNN: Pose-based CNN Features for Action Recognition. In *ICCV*, 2015.

[11] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *MM*, 2014.

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

[13] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[14] V. Gandhi and R. Ronfard. Detecting and naming actors in movies using generative appearance models. In *CVPR*, 2013.

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

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

[17] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009.

[18] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2015.

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

[20] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007.

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[22] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[23] S. Joon Oh, R. Benenson, M. Fritz, and B. Schiele. Person recognition in personal photo collections. In *ICCV*, 2015.

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

[25] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua. Labeled faces in the wild: A survey. In *Advances in Face Detection and Facial Image Analysis*, 2016.

[26] 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.

[27] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

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

[29] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy. Learning to track and identify players from broadcast sports videos. *PAMI*, 2013.

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

[31] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *CVPR*, 2016.

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

[33] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *BMVC*, 2015.

[34] N. Pinto, J. J. DiCarlo, and D. D. Cox. How far can you get with a modern face recognition test set using only simple features? In *CVPR*, 2009.

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

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

[37] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?" Learning person specific classifiers from video. In *CVPR*. IEEE, 2009.

[38] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *BMVC*, volume 2, page 3, 2006.

[39] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, pages 1891–1898, 2014.

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

[41] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. knock! knock! who is it? probabilistic person identification in tv-series. In *CVPR*, 2012.

[42] E. Ustinova, Y. Ganin, and V. Lempitsky. Multi-region bilinear convolutional neural networks for person re-identification. *arXiv preprint arXiv:1512.05300*, 2015.

[43] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *IJCV*, 2012.

[44] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011.

[45] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Workshop on faces in 'real-life' images: Detection, alignment, and recognition*, 2008.

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

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

[48] D. Yi, Z. Lei, and S. Z. Li. Towards pose robust face recognition. In *CVPR*, 2013.

[49] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *ICPR*, 2014.

[50] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[51] N. Zhang, R. Farrell, and T. Darrell. Pose pooling kernels for sub-category recognition. In *CVPR*, 2012.

[52] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*, 2014.

[53] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *CVPR*, 2015.

[54] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *CVPR*, 2010.

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

[56] W.-Y. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM computing surveys*, 2003.

[57] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, 2015.

Supplementary

This supplementary document provides additional qualitative and quantitative results that provide further insights into the results discussed in the main paper. A more detailed description of the datasets is given in § I and the pose clusters discussed in § 3.1 of the main paper are visualized in § II. Additional experimental results and visualizations that show the effectiveness of different components of our framework are given in § III and § IV, respectively.

I. Datasets

I.1. IMDB

We created the IMDB database to train the actor classifier in the movie scenario. The scenario is different from most of the person recognition in that the test set contains a single movie with lesser variation in appearance between multiple instances of an actor in terms of age, style of clothing, etc. We assume that there are no labeled images within the movie and hence training data is not a part of the movie. To create a training set, the images are collected from the IMDB profile[2] of actors appearing in the movie, which are then manually cropped and annotated. Few images from IMDB database are shown in Figure 12. We relied on text tags associated with photos for annotation whenever the photos contain multiple confusing identities. Apart from illumination, resolution, and pose variations, there is a large age variations among IMDB instances. In addition, there is a large domain contrast between IMDB and Hannah test set in terms of lighting, camera and imaging conditions. This creates a more challenging setting to match identites between IMDB and Hannah instances.

I.2. Soccer

Soccer is another scenario where there are a significant number of frames in which the face is not visible and the subjects are often occluded by other players. We show more examples from our soccer dataset in Figure 14. In many instances, head is largely occluded, and in back-view unlike PIPA and Hannah instances, which contain visible head and torso regions. Also, soccer instances exhibit large body deformations, are of low resolution with significant blur. The soccer dataset therefore offers different kinds of challenges for recognition that are not seen in PIPA and Hannah.

II. Pose clusters

We obtain a set of prominent views to facilitate pose-specific representations as discussed in § 3.1. To achieve this, we annotated 14 body keypoints for 29,223 PIPA train instances which are then used for clustering. More

[2] http://www.imdb.com/title/tt0091167/fullcredits

Figure 12: **IMDB:** Each row shows few images of an actor from the dataset. We used IMDB dataset to train classifiers for actor recognition in the Hannah movie.

Figure 13: **Pose clusters:** Each row from top to bottom shows people from PIPA with particular body orientation clustered using orientation and keypoint visibility features.

Figure 14: Images from soccer dataset. It offers a challenging person recognition scenario due to low resolution, high occlusion, deformation and motion blur exhibited by soccer instances.

examples of our pose clusters are shown in Figure 13. Each row from top to bottom contain images from `right`, `semi-right`, `frontal`, `semi-left`, `left`, `back` and `partial body` views. The orientation and keypoint visibility features produced tight clusters containing images with particular body orientation. The last cluster captures the instances with partial upper body such as head or shoulder, etc, in the images that are commonly seen in social media photos and movies. While we considered seven prominent views in this work, we note that generating a large number of views can be helpful, provided there are enough training samples in each cluster to train the convnets.

III. Quantitative Results and Analysis

We provide more insightful results that help to understand merits and challenges of different recognition settings that are considered.

Recognition per subject: Figure 15 shows the number of images for each actor in IMDB and Hannah test sets along with their individual recognition performances. We observe that, for those subjects with sufficiently large number of training instances (*Michael Caine, Barbara Harshey, Woody Allen, Julia Louis-Dreyfus, and Mia Farrow*), the performance is high as expected. For subjects with less than 20 training instances, the performance is very low.

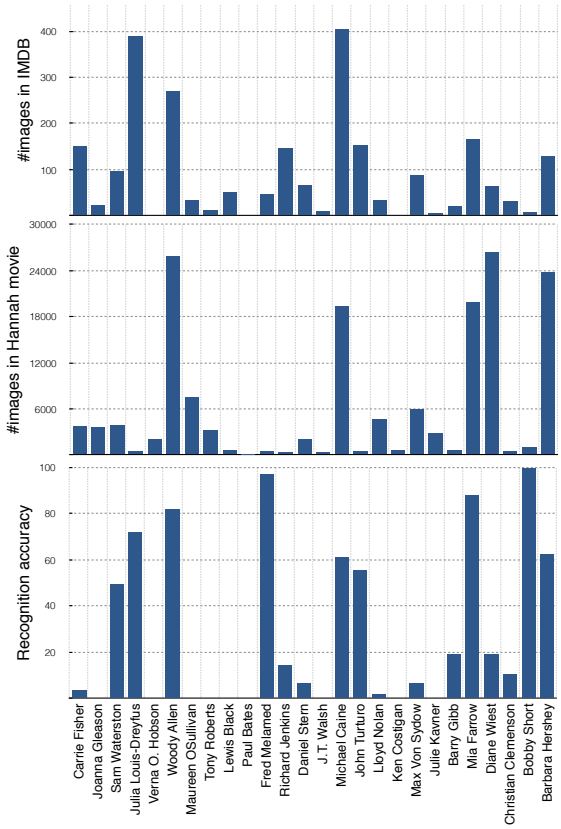

Figure 15: Number of images for each actor in (top) IMDB and (middle) Hannah movie test set. We show the (bottom) recognition performance of each actor on the test set.

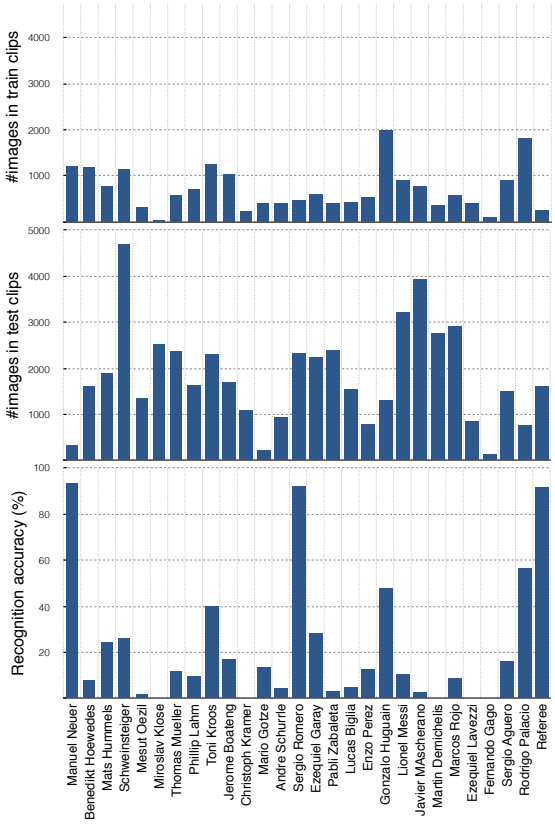

Figure 16: Number of images for each player in the training (top) and test (middle) split of Soccer dataset. We also show the (bottom) recognition performance of each player.

However, whenever there is a large difference in age between train and test instances (*Carrie Fisher, Dianne West, Richard Jenkins*), the performance is poor despite having enough training examples.

Similarly, we show the statistics of soccer players along with their individual performances in Figure 16. We see a similar trend of high performance for subjects (*Gonzalo Huguain* and *Rodrigo Palacio*) with sufficient training instances. We also observe a near 100% accuracy for goal keepers (*Manuel Neuer* and *Sergio Romero*) and the referee due to clothing cues, which are discussed next.

Recognition performance of top subjects: We compare the recognition performance of various approaches on 5 most occurring movie and soccer subjects in Figure 17 and Figure 18, respectively. Our approach reaches an accuracy of 61.17% on top actors, which is significantly better than `naeil`. Note that the overall performance of `naeil` with 17 models is comparable to head and upper body. Unlike photo-albums, clues such as scene and human attributes like age, glasses, and hair color are less useful in the movie setting. For actors with less change in appearance over time (*Michael Caine* and *Woody Allen*: See row three and five in Figure 12), face is found to be extremely informative and robust compared to head.

On the soccer dataset, the overall performance is poor for all the approaches. This suggest to develop better representations that are able to recognize people at a distance.

How informative is clothing? Though it is intuitively obvious that clothing helps in recognition, a qualitative evaluation is not done previously. We perform such a study using the soccer dataset. We show the performance of different approaches on three subjects (*Manuel Neuer*, *Sergio Romero* and *Referee*) with unique clothing in Figure 19. The first two subjects are the goal keepers of the Germany and Argentina, respectively.

As seen in Figure 19, upper body region, which is often less informative compared to head, outperforms head by a large margin due to clothing. The concatenation of head and upper body obtained through separate training is worse than upper body feature alone. On the other hand, the concatenation of features using jointly trained model is more robust and performs much better as it provide more flexibility to focus on selective regions. Finally, the overall performance of pose aware models and `naeil` are identical.

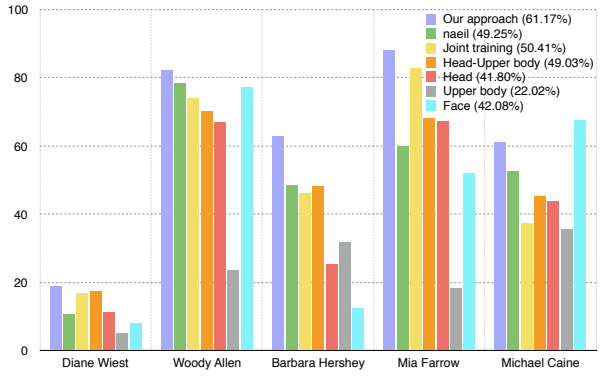

Figure 17: Recognition performance of five lead actors in Hannah dataset.

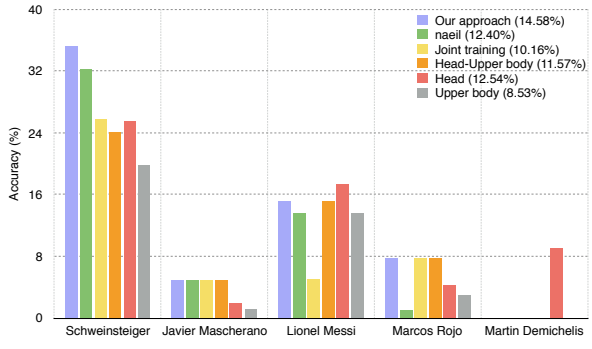

Figure 18: Recognition performance of five most occurring players in Soccer dataset.

It is interesting to note that, convnets that are trained for identity recognition can distinguish clothing without any explicit modeling or hand-crafted features [29].

Confusion between identities: We show the recognition confusion matrix for Hannah and Soccer datasets in Figure 21 and Figure 22 respectively, with and without tracking. We notice two important points related to gender and clothing. As seen from Figure 21, female subjects are mostly getting confused with female subjects, and similarly the male subjects are confused with male subjects. In Figure 22, we notice that players from each team are mislabeled with the members from the same team. These studies show the effectiveness of convnets in capturing human attributes without any explicit training. Finally, majority voting over a track helps to produce consistent predictions.

Domain gap: To understand the effect of domain contrast between train and test instances, we conduct an experiment adding different number of Hannah instances per subject to the IMDB training gallery. The results are shown in Figure 20. As seen from the graph, the addition of even a few instances from the test domain results in a very large improvement in the recognition performance.

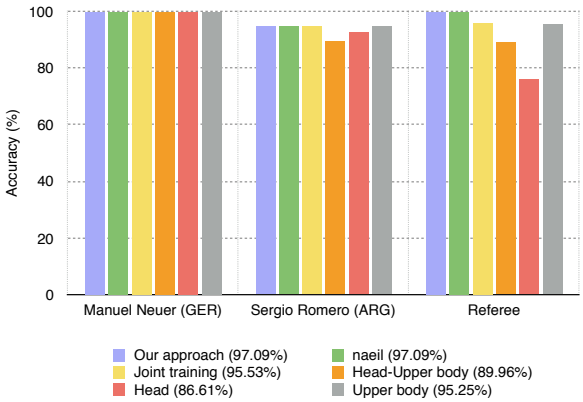

Figure 19: Effect of clothing on recognition.

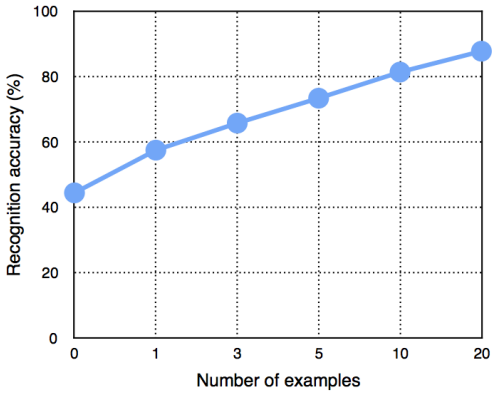

Figure 20: Recognition performance of Hannah movie set using IMDB plus samples from the Hannah test set.

IV. Qualitative Results

We show some qualitative results in Figures 23 to 27. Figure 23 shows the success and failure cases of joint training and separate training of body regions. We notice an over-influence of clothing while using separately trained and concatenated regional features, compared to the jointly training features. In Figure 24, we show the effectiveness of using multiple classifiers from each *PSM*. As seen in the figure, the concatenated head and upper body features ($\mathcal{F}$) may predict incorrect labels even when one (or two) of these features predict correctly, due to the over influence of less informative body region. Combining these three features is found to be more robust.

We show the top scoring predictions obtained from each pose-specific *PSM* in Figure 25. It clearly shows how each *PSM* helps in the prediction of instances in that particular pose when the base model is unable to predict correctly. Finally, we show the success and failure cases of our approach on Hannah and Soccer datasets in Figure 26 and Figure 27 respectively, and compare with the `naeil`.

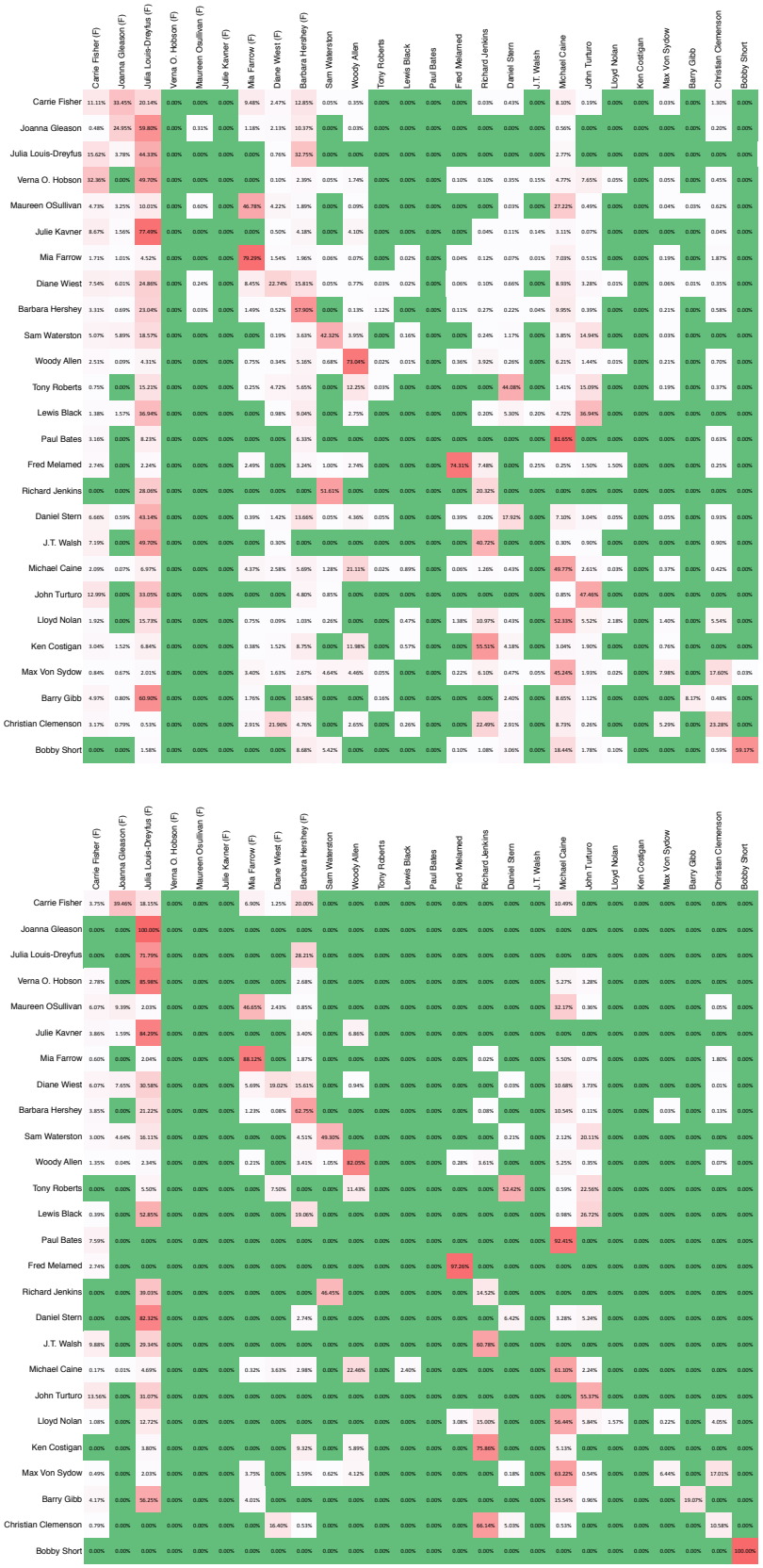

Figure 21: **Confusion matrix** on Hannah dataset (top) with and (bottom) without tracking.

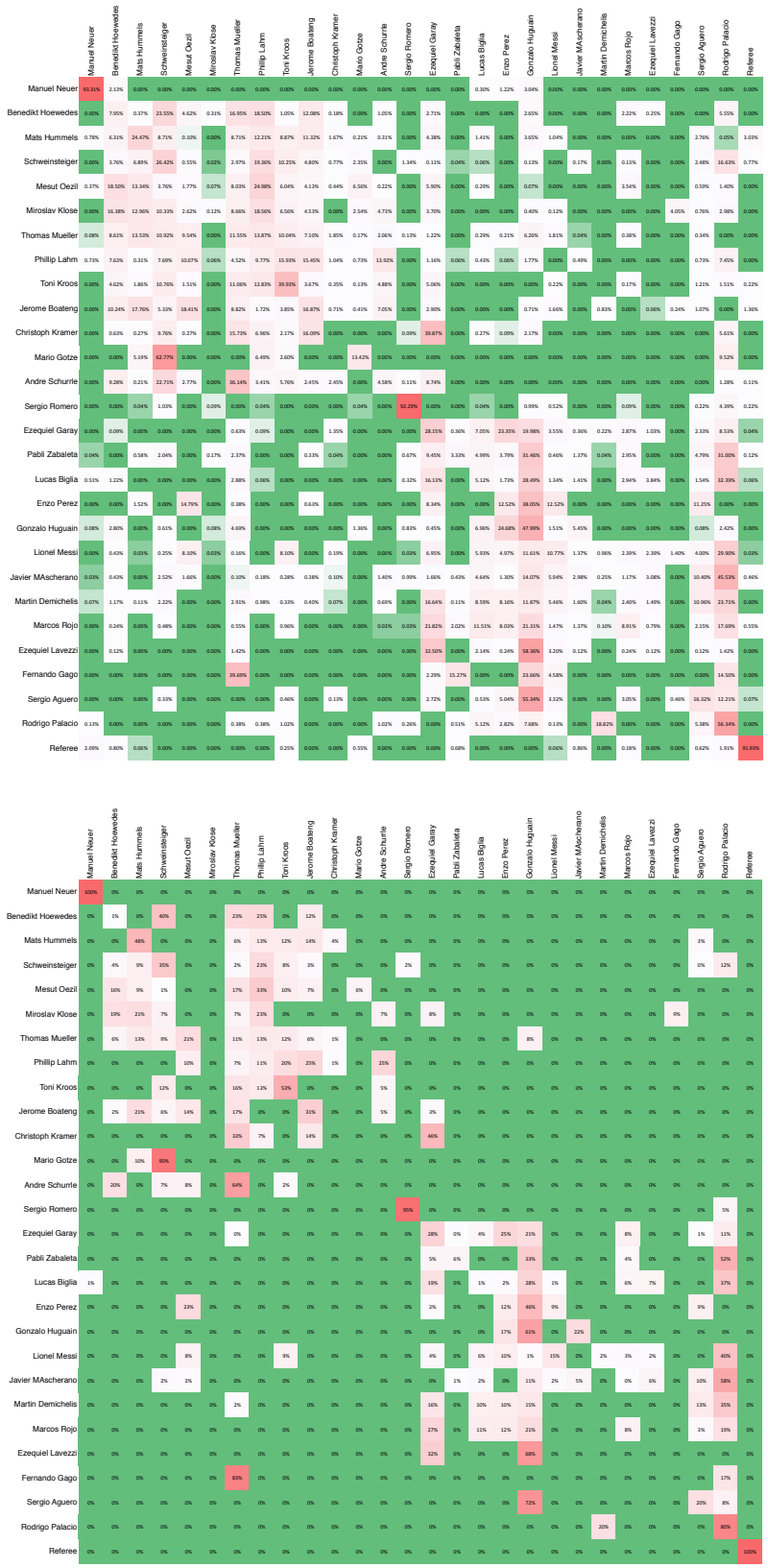

Figure 22: **Confusion matrix** on Soccer dataset (top) with and (bottom) without tracking.

Figure 23: Success and failure cases of separate and joint training of body regions on PIPA dataset. Column one shows the test images and the column two shows the training images belonging to the predicted subject. (Left) shows the success and failure case of joint training (JT) and separate training (ST), respectively and the reverse is shown in (right).

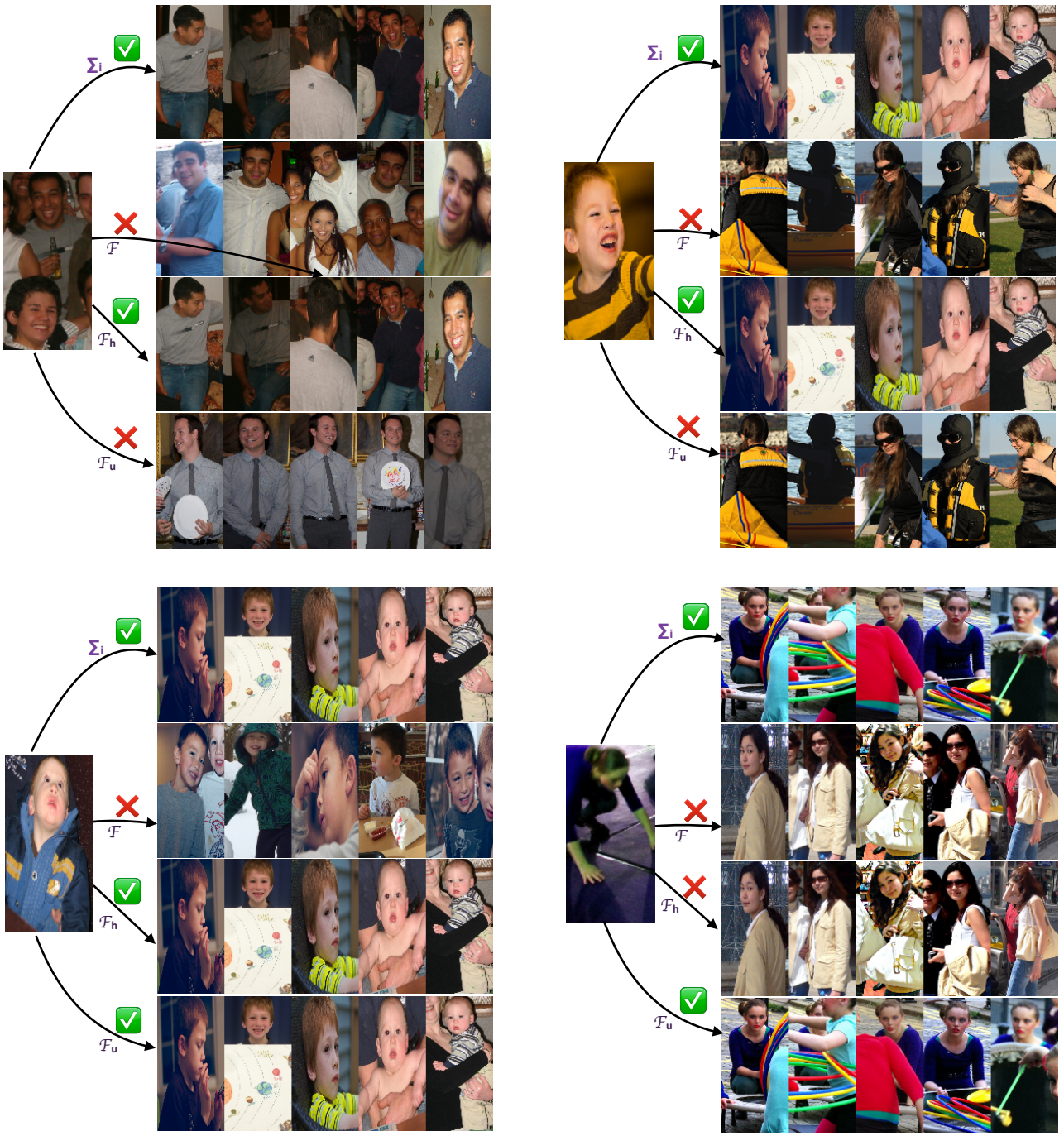

Figure 24: **Effectiveness of multiple classifiers from each *PSM*:** Column one shows the PIPA test images and the column two shows the training images belonging to the predicted subject using different approaches. The four approaches considered are the classifiers trained on head ($\mathcal{F}_h$) and upper body ($\mathcal{F}_u$) features, a classifier trained on concatenated head and upper body ($\mathcal{F}$) feature, and linear combination of three classifiers ($\sum_i$) trained on these features. It clearly shows that it is advantageous to consider individual classifiers trained on regional features and their combination for improved performance.

Figure 25: **Success cases of pose-specific models (*PSM*s) on PIPA dataset.** Each row shows the success predictions of our approach where the improvement is obtained primarily due to the specific-pose model *i.e.*, `base` model wrongly predicts but `base` + correct `PSM` predicts correctly. Green and yellow boxes indicate the success and failure result of `naeil` respectively.

Figure 26: Comparison of our approach with `naeil` on Hannah dataset. Column one shows the test images and the column two shows the training images belonging to the predicted subject. (Left) in green shows the success case of our approach and the failure case of `naeil`. (Right) in red shows the failure case of our approach and the success case of `naeil`.

Figure 27: Comparison of our approach with `naeil` on Soccer dataset. Column one shows the test images and the column two shows the training images belonging to the predicted subject. (Left) in green shows the success case of our approach and the failure case of `naeil`. (Right) in red shows the failure case of our approach and the success case of `naeil`.