

Person Recognition in Social Media Photos

Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele

Abstract—People nowadays share large parts of their personal lives through social media. Being able to automatically recognise people in personal photos may greatly enhance user convenience by easing photo album organisation. For human identification task, however, traditional focus of computer vision has been face recognition and pedestrian re-identification. Person recognition in social media photos sets new challenges for computer vision, including non-cooperative subjects (e.g. backward viewpoints, unusual poses) and great changes in appearance. To tackle this problem, we build a simple person recognition framework that leverages convnet features from multiple image regions (head, body, etc.). We propose new recognition scenarios that focus on the time and appearance gap between training and testing samples. We present an in-depth analysis of the importance of different features according to time and viewpoint generalisability. In the process, we verify that our simple approach achieves the state of the art result on the PIPA [1] benchmark, arguably the largest social media based benchmark for person recognition to date with diverse poses, viewpoints, social groups, and events.

Compared the conference version of the paper [2], this paper additionally presents (1) analysis of a face recogniser (DeepID2+ [3]), (2) new method `naei12` that combines the conference version method `naei1` and DeepID2+ to achieve state of the art results even compared to post-conference works, (3) discussion of related work since the conference version, (4) additional analysis including the head viewpoint-wise breakdown of performance, and (5) results on the open-world setup.

Index Terms—Computer vision, person recognition, social media.

1 INTRODUCTION

WITH the advent of social media and the shift of image capturing mode from digital cameras to smartphones and life-logging devices, users share massive amounts of personal photos online these days. Being able to recognise people in such photos would benefit the users by easing photo album organisation. Recognising people in natural environments poses interesting challenges; people may be focused on their activities with the face not visible, or can change clothing or hairstyle. These challenges are largely new – traditional focus of computer vision research for human identification has been face recognition (frontal, fully visible faces) or pedestrian re-identification (no clothing changes, standing pose).

Intuitively, the ability to recognise faces in the wild [3], [4] is still an important ingredient. However, when people are engaged in an activity (i.e. not posing) their faces become only partially visible (non-frontal, occluded) or simply fully invisible (back-view). Therefore, additional information is required to reliably recognize people. We explore other cues that include (1) body of a person that contains information about the shape and appearance; (2) human attributes such as gender and age; and (3) scene context. See Figure 1 for a list of examples that require increasing number of contextual cues for successful recognition.

This paper presents an in-depth analysis of the person recognition task in the social media type of photos: given a few annotated training images per person, who is this person in the test image? The main contributions of the paper are summarised as follows:

- S. Oh, M. Fritz, and B. Schiele are with the Computer Vision and Multimodal Computing Group, Max Planck Institute for Informatics, Saarbrücken, Germany. E-mail: {joon,mfritz,schiele}@mpi-inf.mpg.de.
- R. Benenson was with the Computer Vision and Multimodal Computing Group, Max Planck Institute for Informatics, Saarbrücken, Germany when this work was done; he is currently with Google, Zürich, Switzerland. E-mail: rodrigo.benenson@gmail.com.

Manuscript received ??; revised ??.



Fig. 1: In social media photos, depending on face occlusion or pose, different cues may be effective. For example, the surfer in the third column is not recognised using only head and body cues due to unusual pose. However, she is successfully recognised when additional attribute cues are considered.

- Propose realistic and challenging person recognition scenarios on the PIPA benchmark (§2).
- Provide a detailed analysis of the informativeness of different cues, in particular of a face recognition module DeepID2+ [3] (§3).
- Verify that our journal version final model `naei12` achieves the new state of the art performance on PIPA (§4).
- Analyse the contribution of cues according to the amount of appearance and viewpoint changes (§5).
- Discuss the performance of our methods under the open-world recognition setup (§B, Appendix)
- Code and data are open source: available at <https://goo.gl/DKuh1Y>.

1.1 Related work

Data type

The bulk of previous work on person recognition focuses on faces. The Labeled Faces in the Wild (LFW) [4] has been a great testbed for a host of works on the face identification and verification outside the lab setting. The benchmark has nearly saturated in the recent years, attributing to the deep features [3], [5], [6], [7], [8], [9], [10] trained on large scale face databases that outperform the traditional methods involving sophisticated classifiers based on hand-crafted features and metric learning approaches [11], [12], [13], [14]. However, LFW is not representative for the social media photos: the data consists mainly of unoccluded frontal faces (face detections) and has a bias towards public figures. Indeed, more recent benchmarks have introduced more difficult types of data. IARPA Janus Benchmark A (IJB-A) [15] includes faces with profile viewpoints, but is still limited to public figures.

Not only face, but body has also been explored as a cue for human identification. For example, pedestrian re-identification (re-id) tackles the problem of matching pedestrian detections in different camera views. Standard benchmarks include VIPeR [16], CAVIAR [17], CUHK [18], and Caltech Roadside Pedestrians [19], with active line of research that previously focused on devising good hand-crafted features [20], [21], [22], and now focusing more on developing effective convnet architectures [23], [24], [25], [26], [27], [28], [29], [30]. However, typical re-id benchmarks do not fully cover the social media setup in three aspects: (1) subjects mostly appear in the standing pose, (2) resolution is low, and (3) the matching is only evaluated across a short time span.

Human identification in natural, everyday environment was first covered by the ‘‘Gallagher collection person dataset’’ [31]. However, the dataset is small (~ 600 images, 32 identities) compared to the size of a typical social media account, and again only the frontal faces are annotated. MegaFace [32], [33] is perhaps the largest known open source face database over social media photos. However, MegaFace does not contain any back-view subject (pruned by a face detector) and the per-account statistics (e.g. number of photos per account) is not preserved due to data processing steps. We build our paper upon the PIPA dataset [1], also crawled from Flickr and fairly large in scale ($\sim 40k$ images, $\sim 2k$ identities), with diverse appearances and subjects with all viewpoints and occlusion levels. Heads are annotated with bounding boxes each with an identity tag. We describe PIPA in greater detail in §2.

Recognition tasks

There exist multiple tasks related to person recognition [34] differing mainly in the amount of training and testing data. Face and surveillance re-identification is most commonly done via verification: *given one reference image (gallery) and one test image (probe), do they show the same person?* [4], [35]. In this paper, we consider two recognition tasks. (1) closed world identification: *given a single test image (probe), who is this person among the identities that are among the training identities (gallery set)?* (2) Open world recognition [32] (§B, Appendix): *given a single test image (probe), is this person among the training identities (gallery set)? If so, who?*

Other related tasks are, face clustering [7], [36], finding important people [37], or associating names in text to faces in images [38], [39].

Prior work with the same data type and task

Since the introduction of the PIPA dataset [1], multiple works have proposed different methods for solving the person recognition problem in social media photos. Zhang et al. proposed the Pose Invariant Person Recognition (PIPER) [1], obtaining promising results by combining three ingredients: DeepFace [5] (face recognition module trained on a large private dataset), poselets [40] (pose estimation module trained with 2k images and 19 keypoint annotations), and convnet features trained on detected poselets [41], [42].

Oh et al. [2], the conference version of this paper, have proposed a simple model `naeil` that extracts AlexNet cues from multiple *fixed* image regions. In particular, it does not require data-heavy DeepFace or time-costly poselets, while achieving a slightly better recognition performance than PIPER.

There have been many follow-up works since then. Kumar et al. [43] have improved the performance by normalising the body pose using pose estimation. Li et al. [44] considered exploiting people co-occurrence statistics. Liu et al. [45] have proposed to train a person embedding in a metric space instead of training a classifier on a fixed set of identities, thereby making the model more adaptable to unseen identities. Some works have exploited the photo-album metadata, allowing the model to reason over different photos [46], [47].

In this journal version, we build `naeil2` from `naeil` and DeepID2+ [3] to achieve the state of the art result among the published work on PIPA. We provide additional analysis of cues according to time and viewpoint changes.

2 DATASET AND EXPERIMENTAL SETUP

Dataset

The PIPA dataset (‘‘People In Photo Albums’’) [1] is, to the best of our knowledge, the first dataset to annotate people’s identities even when they are pictured from the back. The annotators labelled instances that can be considered hard even for humans (see qualitative examples in figure 15, 16). PIPA features 37 107 Flickr personal photo album images (Creative Commons license), with 63 188 head bounding boxes of 2 356 identities. The head bounding boxes are tight around the skull, including the face and hair; occluded heads are hallucinated by the annotators. The dataset is partitioned into *train*, *val*, *test*, and *leftover* sets, with rough ratio 45 : 15 : 20 : 20 percent of the annotated heads. The leftover set is not used in this paper. Up to annotation errors, neither identities nor photo albums by the same uploader are shared among these sets.

Task

At test time, the system is given a photo and ground truth head bounding box corresponding to the test instance (probe). The task is to choose the identity of the test instance among a given set of identities (gallery set, 200~500 identities) each with ~ 10 training samples.

In Appendix §B, we evaluate the methods when the test instance may be a background person (e.g. bystanders – no training image given). The system is then also required to determine if the given instance is among the seen identities (gallery set).

Protocol

We follow the PIPA protocol in [1] for data utilisation and model evaluation. The *train* set is used for convnet feature training. The *test* set contains the examples for the test identities. For each identity, the samples are divided into $test_0$ and $test_1$. For evaluation, we perform a two-fold cross validation by training on one of the splits and testing on the other. The *val* set is likewise split into val_0 and val_1 , and is used for exploring different models and tuning hyperparameters.

Evaluation

We use the recognition rate (or accuracy), the rate of correct identity predictions among the test instances. For every experiment, we average two recognition rates obtained from the (training, testing) pairs (val_0, val_1) and (val_1, val_0) – analogously for *test*.

2.1 Splits

We consider four different ways of splitting the training and testing samples ($val_{0/1}$ and $test_{0/1}$) for each identity, aiming to evaluate different level of generalisation ability. The first one is from a prior work, and we introduce three new ones. Refer to table 1 for data statistics and figure 3 for visualisation.

Original split \mathcal{O} [1]

The Original split shares many similar examples per identity across the split – e.g. photos taken in a row. The Original split is thus easy - even nearest neighbour on raw RGB pixels works (§4.1). In order to evaluate the ability to generalise across long-term appearance changes, we introduce three new splits below.

Album split \mathcal{A} [2]

The Album split divides training and test samples for each identity according to the photo album metadata. Each split takes the albums while trying to match the number of samples per identity as well as the total number of samples across the splits. A few albums are shared between the splits in order to match the number of samples. Since the Flickr albums are user-defined and do not always strictly cluster events and occasions, the split may not be perfect.

Time split \mathcal{T} [2]

The Time split divides the samples according to the time the photo was taken. For each identity, the samples are sorted according to their “photo-taken-date” metadata, and then divided according to the newest versus oldest basis. The instances without time metadata are distributed evenly. This split evaluates the temporal generalisation of the recogniser. However, the “photo-taken-date” metadata is very noisy with lots of missing data.

Day split \mathcal{D} [2]

The Day split divides the instances via visual inspection to ensure the firm “appearance change” across the splits. We define two criteria for division: (1) a firm evidence of date change such as {change of season, continent, event, co-occurring people} and/or (2) visible changes in {hairstyle, make-up, head or body wear}. We discard identities for whom such a division is not possible. After division, for each identity we randomly discard samples from the larger split until the sizes match. If the smaller split has ≤ 4 instances, we discard the identity altogether. The Day split enables clean experiments for evaluating the generalisation performance across strong appearance and event changes.

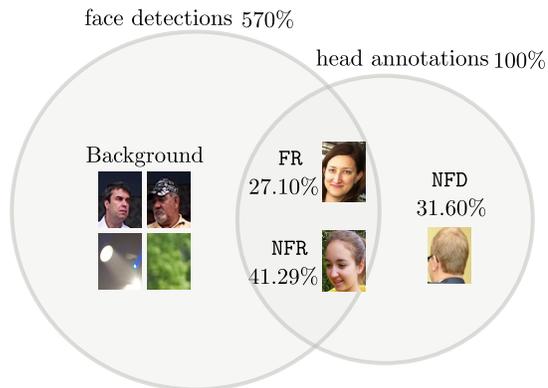


Fig. 2: Face detections and head annotations in PIPA. The matches are determined by overlap (intersection over union). For matched faces (heads), the detector DPM component gives the orientation information (frontal versus non-frontal).

| | | <i>val</i> | | | | <i>test</i> | | | |
|-------|----------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| | | \mathcal{O} | \mathcal{A} | \mathcal{T} | \mathcal{D} | \mathcal{O} | \mathcal{A} | \mathcal{T} | \mathcal{D} |
| spl.0 | instance | 4820 | 4859 | 4818 | 1076 | 6443 | 6497 | 6441 | 2484 |
| | identity | 366 | 366 | 366 | 65 | 581 | 581 | 581 | 199 |
| spl.1 | instance | 4820 | 4783 | 4824 | 1076 | 6443 | 6389 | 6445 | 2485 |
| | identity | 366 | 366 | 366 | 65 | 581 | 581 | 581 | 199 |

TABLE 1: Split statistics for *val* and *test* sets. Total number of instances and identities for each split is shown.

2.2 Face detection

Instances in PIPA are annotated by humans around their heads (tight around skull). We additionally compute face detections over PIPA for three purposes: (1) to compare the amount of identity information in head versus face (§3), (2) to obtain head orientation information for further analysis (§5), and (3) to simulate the scenario without ground truth head box at test time (Appendix §B). We use the open source DPM face detector [48].

Given a set of detected faces (above certain detection score threshold) and the ground truth heads, the match is made according to the overlap (intersection over union). For matched heads, the corresponding face detections tell us which DPM component is fired, thereby allowing us to infer the head orientation (frontal or side view). See Appendix §A for further details.

Using the DPM component, we partition instances in PIPA as follows: (1) detected and frontal (FR, 41.29%), (2) detected and non-frontal (NFR, 27.10%), and (3) no face detected (NFD, 31.60%). We denote detections without matching ground truth head as Background. See figure 2 for visualisation.

3 CUES FOR RECOGNITION

In this section, we investigate the cues for recognising people in social media photos. We begin with an overview of our model. Then, we experimentally answer the following questions: how informative are fixed body regions (no pose estimation) (§3.4)? How much does scene context help (§3.5)? Is it head or face (head minus hair and background) that is more informative (§3.6)? And how much do we gain by using extended data (§3.7 & §3.8)? How effective is a specialised face recogniser (§3.10)? Studies in this section are based exclusively on the *val* set.



Fig. 3: Visualisation of Original, Album, Time and Day splits for three identities (rows 1-3). Greater appearance gap is observed from Original to Day splits.

3.1 Model overview

At test time, given a ground truth head bounding box, we estimate five different regions depicted in figure 4. Each region is fed into one or more convnets to obtain a set of cues. The cues are concatenated to form a feature vector describing the instance. Throughout the paper we write $+$ to denote vector concatenation. Linear SVM classifiers are trained over this feature vector (one versus the rest). In our final system, except for DeepID2+ [3], all features are computed using the seventh layer (fc7) of AlexNet [41] pre-trained for ImageNet classification. The cues only differ amongst each other on the image area and the fine-tuning used (type of data or surrogate task) to alter the AlexNet, except for the DeepID2+ [3] feature.



Fig. 4: Regions considered for feature extraction: face f , head h , upper body u , full body b , and scene s . More than one cue can be extracted per region (e.g. h_1, h_2).

3.2 Image regions used

We choose five different image regions based on the ground truth head annotation (given at test time, see the protocol in §2). The head rectangle h corresponds to the ground truth annotation. The full body rectangle b is defined as $(3 \times \text{head width}, 6 \times \text{head height})$, with the head at the top centre of the full body. The upper body rectangle u is the upper-half of b . The scene region s is the whole image containing the head.

The face region f is obtained using the DPM face detector discussed in §2.2. For head boxes with no matching detection (e.g. back views and occluded faces), we regress the face area from the head using the face-head displacement statistics on the *train* set. Five respective image regions are illustrated in figure 4.

Note that the regions overlap with each other, and that depending on the person’s pose they might be completely off. For example, b for a lying person is likely to contain more background than the actual body.

3.3 Fine-tuning and parameters

Unless specified otherwise AlexNet is fine-tuned using the PIPA *train* set ($\sim 30k$ instances, $\sim 1.5k$ identities), cropped at five different image regions, with 300k mini-batch iterations (batch size 50). We refer to the base cue thus obtained as f , h , u , b , or s , depending on the cropped region. On the *val* set we found the

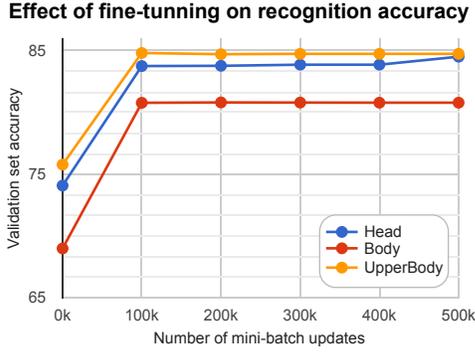


Fig. 5: PIPA *val* set performance of different cues versus the SGD iterations in fine-tuning.

| Cue | Accuracy |
|--------------|----------|
| Chance level | 1.04 |
| Scene (§3.5) | 27.06 |
| Body | 80.81 |
| Upper body | 84.76 |
| Head | 83.88 |
| Face (§3.6) | 74.45 |
| Zoom out | 74.45 |
| | 84.80 |
| | 90.65 |
| | 91.14 |
| | 91.16 |
| Zoom in | 27.06 |
| | 82.16 |
| | 86.39 |
| | 90.40 |
| | 91.16 |
| Head+body | 89.42 |
| Full person | 91.14 |
| Full image | 91.16 |

TABLE 2: PIPA *val* set accuracy of cues based on different image regions and their concatenations (+ means concatenation).

fine-tuning to provide a systematic ~ 10 percent points (pp) gain over the non-fine-tuned AlexNet (figure 5). We use the seventh layer (fc7) of AlexNet for each cue (4 096 dimensions).

We train for each identity a one-versus-all SVM classifier with the regularisation parameter $C = 1$; it turned out to be an insensitive parameter in our preliminary experiments. As an alternative, the naive nearest neighbour classifier has also been considered. However, on the *val* set the SVMs consistently outperforms the NNs by a ~ 10 pp margin.

3.4 How informative is each image region?

Table 2 shows the *val* set results of each region individually and in combination. Head h and upper body u are the strongest individual cues. Upper body is more reliable than the full body b because the lower body is commonly occluded or cut out of the frame, and thus is usually a distractor. Scene s is, unsurprisingly, the weakest individual cue, but it still useful information for person recognition (far above chance level). Importantly, we see that all cues complement each other, despite overlapping pixels. Overall, our features and combination strategy are effective.

3.5 Scene (s)

Other than a fine-tuned AlexNet we considered multiple feature types to encode the scene information. s_{gist} : using the

| Method | Accuracy |
|----------------------|----------|
| Gist | 21.56 |
| PlacesNet scores | 21.44 |
| raw PlacesNet | 27.37 |
| PlacesNet fine-tuned | 25.62 |
| raw AlexNet | 26.54 |
| AlexNet fine-tuned | 27.06 |

TABLE 3: PIPA *val* set accuracy of different scene cues. See descriptions in §3.5.

Gist descriptor [49] (512 dimensions). $s_{0 \text{ places}}$: instead of using AlexNet pre-trained on ImageNet, we consider an AlexNet (PlacesNet) pre-trained on 205 scene categories of the “Places Database” [50] (~ 2.5 million images). $s_{\text{places } 205}$: Instead of the 4 096 dimensions PlacesNet feature vector, we also consider using the score vector for each scene category (205 dimensions). s_0, s_3 : finally we consider using AlexNet in the same way as for body or head (with zero or 300k iterations of fine-tuning on the PIPA person recognition training set). $s_{3 \text{ places}}$: $s_{0 \text{ places}}$ fine-tuned for person recognition.

Results

Table 3 compares the different alternatives on the *val* set. The Gist descriptor s_{gist} performs only slightly below the convnet options (we also tried the 4 608 dimensional version of Gist, obtaining worse results). Using the raw (and longer) feature vector of $s_{0 \text{ places}}$ is better than the class scores of $s_{\text{places } 205}$. Interestingly, in this context pre-training for places classification is better than pre-training for objects classification ($s_{0 \text{ places}}$ versus s_0). After fine-tuning s_3 reaches a similar performance as $s_{0 \text{ places}}$.

Experiments trying different combinations indicate that there is little complementarity between these features. Since there is not a large difference between $s_{0 \text{ places}}$ and s_3 , for the sake of simplicity we use s_3 as our scene cue s in all other experiments.

Conclusion

Scene s by itself, albeit weak, can obtain results far above the chance level. After fine-tuning, scene recognition as pre-training surrogate task [50] does not provide a clear gain over (ImageNet) object recognition.

3.6 Head (h) or face (f)?

A large portion of work on face recognition focuses on the face region specifically. In the context of photo albums, we aim to quantify how much information is available in the head versus the face region. As discussed in §2.2, we obtain the face regions f from the DPM face detector [48].

Results

There is a large gap of ~ 10 percent points performance between f and h in table 2 highlighting the importance of including the hair and background around the face.

Conclusion

Using h is more effective than f , but f result still shows a fair performance. As with other body cues, there is a complementarity between h and f ; we suggest to use them together.

| | Method | Accuracy |
|-------------------|--|----------|
| More data (§3.7) | h | 83.88 |
| | $h + h_{\text{cacad}}$ | 84.88 |
| | $h + h_{\text{casia}}$ | 86.08 |
| | $h + h_{\text{casia}} + h_{\text{cacad}}$ | 86.26 |
| Attributes (§3.8) | h_{pipallm} | 74.63 |
| | h_{pipall} | 81.74 |
| | $h + h_{\text{pipall}}$ | 85.00 |
| | u_{peta5} | 77.50 |
| | $u + u_{\text{peta5}}$ | 85.18 |
| | $A = h_{\text{pipall}} + u_{\text{peta5}}$ | 86.17 |
| | $h + u$ | 85.77 |
| naeil (§3.9) | $h + u + A$ | 90.12 |
| | naeil [2] | 91.70 |

TABLE 4: PIPA *val* set accuracy of different cues based on extended data. See §3.7, §3.8, and §3.9 for details.

3.7 Additional training data ($h_{\text{cacad}}, h_{\text{casia}}$)

It is well known that deep learning architectures benefit from additional data. DeepFace [5] used by PIPER [1] is trained over $4.4 \cdot 10^6$ faces of $4 \cdot 10^3$ persons (the private SFC dataset [5]). In comparison our cues are trained over ImageNet and PIPA’s $29 \cdot 10^3$ faces over $1.4 \cdot 10^3$ persons. To measure the effect of training on larger data we consider fine-tuning using two open source face recognition datasets: CASIA-WebFace (CASIA) [51] and the “Cross-Age Reference Coding Dataset” (CACD) [52].

CASIA contains $0.5 \cdot 10^6$ images of $10.5 \cdot 10^3$ persons (mainly actors and public figures). When fine-tuning AlexNet over these identities (using the head area h), we obtain the h_{casia} cue.

CACD contains $160 \cdot 10^3$ faces of $2 \cdot 10^3$ persons with varying ages. Although smaller in total number of images than CASIA, CACD features greater number of samples per identity ($\sim 2\times$). The h_{cacad} cue is built via the same procedure as h_{casia} .

Results

See the top part of table 4 for the results. $h + h_{\text{cacad}}$ and $h + h_{\text{casia}}$ improve over h (1.0 and 2.2 pp, respectively). Extra convnet training data seems to help. However, due to the mismatch in data distribution, h_{cacad} and h_{casia} on their own are about ~ 5 pp worse than h .

Conclusion

Extra convnet training data helps, even if they are from different type of photos.

3.8 Attributes ($h_{\text{pipall}}, u_{\text{peta5}}$)

Albeit overall appearance might change day to day, one could expect that stable, long term attributes provide means for recognition. We build attribute cues by fine-tuning AlexNet features not for the person recognition task (like for all other cues), but rather for the attribute prediction surrogate task. We consider two sets attributes, one on the head region and the other on the upper body region.

We have annotated identities in the PIPA *train* and *val* sets ($1409 + 366$ in total) with five long term attributes: age, gender, glasses, hair colour, and hair length (see table 5 for details). We build h_{pipall} by fine-tuning AlexNet features for the task of head attribute prediction.

For fine-tuning the attribute cue h_{pipall} , we consider two approaches: training a single network for all attributes as a multi-label classification problem with the sigmoid cross entropy loss,

| Attribute | Classes | Criteria |
|------------|-------------|--|
| Age | Infant | Not walking (due to young age) |
| | Child | Not fully grown body size |
| | Young Adult | Fully grown & Age < 45 |
| | Middle Age | $45 \leq \text{Age} \leq 60$ |
| | Senior | Age > 60 |
| Gender | Female | Female looking |
| | Male | Male looking |
| Glasses | None | No eyewear |
| | Glasses | Transparent glasses |
| | Sunglasses | Glasses with eye occlusion |
| Haircolour | Black | Black |
| | White | Any hint of whiteness |
| | Others | Neither of the above |
| Hairlength | No hair | Absolutely no hair on the scalp |
| | Less hair | Hairless for $> \frac{1}{2}$ upper scalp |
| | Short hair | When straightened, < 10 cm |
| | Med hair | When straightened, < chin level |
| | Long hair | When straightened, > chin level |

TABLE 5: PIPA attributes details.

or tuning one network per attribute separately and concatenating the feature vectors. The results on the *val* set indicate the latter (h_{pipall}) performs better than the former (h_{pipallm}).

For the upper body attribute features, we use the “PETA pedestrian attribute dataset” [53]. The dataset originally has 105 attributes annotations for $19 \cdot 10^3$ full-body pedestrian images. We chose the five long-term attributes for our study: gender, age (young adult, adult), black hair, and short hair (details in table 5). We choose to use the upper-body u rather than the full body b for attribute prediction – the crops are much less noisy. We train the AlexNet feature on upper body of PETA images with the attribute prediction task to obtain the cue u_{peta5} .

Results

See results in table 4. Both PIPA (h_{pipall}) and PETA (u_{peta5}) annotations behave similarly (~ 1 pp gain over h and u), and show complementary (~ 5 pp gain over $h+u$). Amongst the attributes considered, gender contributes the most to improve recognition accuracy (for both attributes datasets).

Conclusion

Adding attribute information improves the performance.

3.9 Conference version final model (naeil) [2]

The final model in the conference version of this paper combines five vanilla regional cues ($P_s = P+s$), two head cues trained with extra data ($h_{\text{cacad}}, h_{\text{casia}}$), and ten attribute cues ($h_{\text{pipall}}, u_{\text{peta5}}$), resulting in 17 cues in total. We name this method *naeil* [2]¹.

Results

See table 4 for the results. *naeil*, by combining all the cues considered naively, achieves the best result 91.70% on the *val* set.

Conclusion

Cues considered thus far are complementary, and the combined model *naeil* is effective.

1. “naeil”, 내일, means “tomorrow” and sounds like “nail”.

3.10 DeepID2+ face recognition module (h_{deepid}) [3]

Face recognition performance have improved significantly in recent years with better architectures and larger open source datasets [3], [4], [5], [7], [8], [9], [10], [54], [55]. In this section, we study how much face recognition helps in person recognition. While DeepFace [5] used by the PIPER [1] would have enabled more direct comparison against the PIPER, it is not publicly available. We thus choose the DeepID2+ face recogniser [3]. Face recognition technology is still improving quickly, and larger and larger face datasets are being released – the analysis in this section would be an underestimate of current and future face recognisers.

The DeepID2+ network is a siamese neural network that takes 25 different crops of head as input, with the joint verification-identification loss. The training is based on large databases consisting of CelebFaces+ [56], WDRRef [57], and LFW [4] – totalling $2.9 \cdot 10^5$ faces of $1.2 \cdot 10^4$ persons. At test time, it ensembles the predictions from the 25 crop regions obtained by facial landmark detections. The resulting output is 1 024 dimensional head feature that we denote as h_{deepid} .

Since the DeepID2+ pipeline begins with the facial landmark detection, the DeepID2+ features are not available for instances with e.g. occluded or backward orientation heads. As a result, only 52 709 out of 63 188 instances (83.42%) have the DeepID2+ features available, and we use vectors of zeros as features for the rest.

Results - Original split

See table 6 for the *val* set results for h_{deepid} and related combinations. h_{deepid} in itself is weak (68.46%) compared to the vanilla head feature h , due to the missing features for the back-views. However, when combined with h , the performance reaches 85.86% by exploiting information from strong DeepID2+ face features and the viewpoint robust h features.

Since the feature dimensions are not homogeneous (4 096 versus 1 024), we try L_2 normalisation of h and h_{deepid} before concatenation ($h \oplus h_{\text{deepid}}$). This gives a further 3% boost (88.74%) – better than $h + h_{\text{cacd}} + h_{\text{casia}}$, previous best model on the head region (86.26%).

Results - Album, Time and Day splits

Table 6 also shows results for the Album, Time, and Day splits on the *val* set. While the general head cue h degrades significantly on the Day split, h_{deepid} is a reliable cue with roughly the same level of recognition in all four splits (60~70%). This is not surprising, since face is largely invariant over time, compared to hair, clothing, and event.

On the other splits as well, the complementarity of h and h_{deepid} is guaranteed only when they are L_2 normalised before concatenation. The L_2 normalised concatenation $h \oplus h_{\text{deepid}}$ envelops the performance of individual cues on all splits.

Conclusion

DeepID2+, with face-specific architecture/loss and massive amount of training data, contributes highly useful information for the person recognition task. However, being only able to recognise face-visible instances, it needs to be combined with orientation-robust h to ensure the best performance. Unsurprisingly, having a specialised face recogniser helps more in the setup with larger appearance gap between training and testing samples (Album, Time, and Day splits). Better face recognisers will further improve the results in the future.

| Method | Original | Album | Time | Day |
|------------------------------|----------|-------|-------|-------|
| h | 83.88 | 77.90 | 70.38 | 40.71 |
| h_{deepid} | 68.46 | 66.91 | 64.16 | 60.46 |
| $h + h_{\text{deepid}}$ | 85.86 | 80.54 | 73.31 | 47.86 |
| $h \oplus h_{\text{deepid}}$ | 88.74 | 85.72 | 80.88 | 66.91 |
| $naeil$ [2] | 91.70 | 86.37 | 80.66 | 49.21 |
| $naeil + h_{\text{deepid}}$ | 92.11 | 86.77 | 81.08 | 51.02 |
| $naeil2$ | 93.42 | 89.95 | 85.87 | 70.58 |

TABLE 6: PIPA *val* set accuracy of methods involving h_{deepid} . The optimal combination weights are $\lambda^* = [0.60 \ 1.05 \ 1.00 \ 1.50]$ for Original, Album, Time, and Day splits, respectively. \oplus means L_2 normalisation before concatenation.

3.11 Combining $naeil$ with h_{deepid} ($naeil2$)

We build the final model of the journal version, namely the $naeil2$ by combining $naeil$ and h_{deepid} . As seen in §3.10, naive concatenation is likely to fail due to even larger difference in dimensionality ($4\,096 \times 17 = 69\,632$ versus 1 024). We consider L_2 normalisation of $naeil$ and h_{deepid} , and then performing a weighted concatenation.

$$naeil \oplus_{\lambda} h_{\text{deepid}} = \frac{naeil}{\|naeil\|_2} + \lambda \cdot \frac{h_{\text{deepid}}}{\|h_{\text{deepid}}\|_2}, \quad (1)$$

where, $\lambda > 0$ is a parameter and $+$ denotes a concatenation.

Optimisation of λ on *val* set

λ determines how much relative weight to be given to h_{deepid} . As we have seen in §3.10, the amount of additional contribution from h_{deepid} is different for each split. In this section, we find λ^* , the optimal values for λ , for each split over the *val* set. The resulting combination of $naeil$ and h_{deepid} is our final method, $naeil2$. λ^* is searched on the equi-distanced points $\{0, 0.05, 0.1, \dots, 3\}$.

See figure 6 for the *val* set performance of $naeil \oplus_{\lambda} h_{\text{deepid}}$ with varying values of λ . The optimal weights are found at $\lambda^* = [0.60 \ 1.05 \ 1.00 \ 1.50]$ for Original, Album, Time, and Day splits, respectively. The relative importance of h_{deepid} is greater on splits with larger appearance changes. For each split, we denote $naeil2$ as the combination $naeil$ and h_{deepid} based on the optimal weights.

Note that the performance curve is rather stable for $\lambda \geq 1.5$ in all splits. In practice, when the expected amount of appearance changes of subjects are unknown, our advice would be to choose $\lambda \approx 1.5$. Finally, we remark that the weighted sum can also be done for the 17 cues in $naeil$; finding the optimal cue weights is left as a future work.

Results

See table 6 for the results of combining $naeil$ and h_{deepid} . Naively concatenated, $naeil + h_{\text{deepid}}$ performs worse than h_{deepid} on the Day split (51.02% vs 60.46%). However, the weighted combination $naeil2$ achieves the best performance on all four splits.

Conclusion

When combining $naeil$ and h_{deepid} , a weighted combination is desirable, and the resulting final model $naeil2$ beats all the previously considered models on all four splits.

| | Method | Special modules | | General features | | Original | Album | Time | Day |
|-------|------------------------|-----------------|----------------|------------------|----------|--------------|--------------|--------------|--------------|
| | | Face rec. | Pose est. | Data | Arch. | | | | |
| | Chance level | \times | \times | — | — | 0.78 | 0.89 | 0.78 | 1.97 |
| Head | h_{rgb} | \times | \times | — | — | 33.77 | 27.19 | 16.91 | 6.78 |
| | h | \times | \times | I+P | Alex | 76.42 | 67.48 | 57.05 | 36.48 |
| | $h+h_{casia}+h_{cacd}$ | \times | \times | I+P+CC | Alex | 80.32 | 72.82 | 63.18 | 45.45 |
| | h_{deepid} | DeepID2+ [3] | \times | — | — | 68.06 | 65.49 | 60.69 | 61.49 |
| | $h \oplus h_{deepid}$ | DeepID2+ [3] | \times | I+P | Alex | 85.94 | 81.95 | 75.85 | 66.00 |
| | DeepFace [1] | DeepFace [5] | \times | — | — | 46.66 | — | — | — |
| Body | b | \times | \times | I+P | Alex | 69.63 | 59.29 | 44.92 | 20.38 |
| | $h+b$ | \times | \times | I+P | Alex | 83.36 | 73.97 | 63.03 | 38.15 |
| | $P = f+h+u+b$ | \times | \times | I+P | Alex | 85.33 | 76.49 | 66.55 | 42.17 |
| | GlobalModel [1] | \times | \times | I+P | Alex | 67.60 | — | — | — |
| | PIPER [1] | DeepFace [5] | Poselets [40] | I+P | Alex | 83.05 | — | — | — |
| | Pose [43] | \times | Pose group | I+P+V | Alex | 89.05 | 82.37 | 74.84 | 56.73 |
| | COCO [45] | \times | Part det. [58] | I+P | Goog,Res | 92.78 | 83.53 | 77.68 | 61.73 |
| Image | $P_s = P+s$ | \times | \times | I+P | Alex | 85.71 | 76.68 | 66.55 | 42.31 |
| | $naeil = P_s + E$ [2] | \times | \times | I+P+E | Alex | 86.78 | 78.72 | 69.29 | 46.54 |
| | Contextual [47] | DeepID [56] | \times | I+P | Alex | 88.75 | 83.33 | 77.00 | 59.35 |
| | $naeil2$ (this paper) | DeepID2+ [3] | \times | I+P+E | Alex | 90.42 | 86.30 | 80.74 | 70.58 |

TABLE 7: PIPA *test* set accuracy (%) of the proposed method and prior arts on the four splits. For each method, we indicate any face recognition or pose estimation module included, and the data and convnet architecture for other features.

Cues on extended data $E = h_{casia} + h_{cacd} + h_{pipa11} + u_{peta5}$.

\oplus means concatenation after L_2 normalisation.

In the data column, I indicates ImageNet [42] and P indicates PIPA *train* set. CC means CACD [52]+CASIA [51] and E means CC+PETA [53]. V indicates the VGGFace dataset [8].

In the architecture column, (Alex,Goog,Res) refers to (AlexNet [41],GoogleNetv3 [59],ResNet50 [60]).

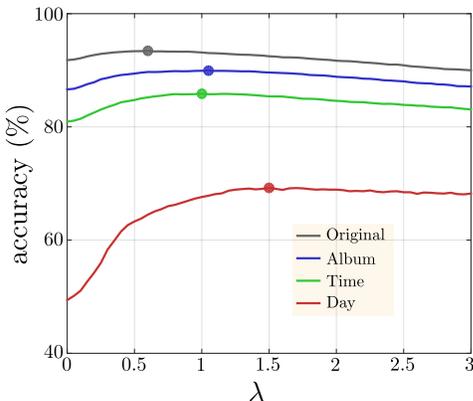


Fig. 6: PIPA *val* set accuracy of $naeil \oplus_{\lambda} h_{deepid}$ for varying values of λ . Round dots denote the maximal *val* accuracy.

4 PIPA TEST SET RESULTS AND COMPARISON

In this section, we measure the performance of our final model and key intermediate results on the PIPA *test* set, and compare against the prior arts. See table 7 for a summary.

4.1 Baselines

We consider two baselines for measuring the inherent difficulty of the task. First baseline is the “chance level” classifier, which does not see the image content and simply picks the most commonly occurring class. It provides the lower bound for any recognition method, and gives a sense of how large the gallery set is.

Our second baseline is the raw RGB nearest neighbour classifier h_{rgb} . It uses the raw downsized (40×40 pixels) and blurred RGB head crop as the feature. The identity of the Euclidean distance nearest neighbour training image is predicted at test time. By design, h_{rgb} is only able to recognize near identical head crops across the $test_{0/1}$ splits.

Results

See results for “chance level” and h_{rgb} in table 7. While the “chance level” performance is low ($\leq 2\%$ in all splits), we observe that h_{rgb} performs unreasonably well on the Original split (33.77%). This shows that the Original splits share many nearly identical person instances across the split, and the task is very easy. On the harder splits, we see that the h_{rgb} performance diminishes, reaching only 6.78% on the Day split. Recognition on the Day split is thus far less trivial – simply taking advantage of pixel value similarity would not work.

Conclusion

Although the gallery set is large enough, the task can be made arbitrarily easy by sharing many similar instances across the splits (Original split). We have remedied the issue by introducing three more challenging splits (Album, Time, and Day) on which the naive RGB baseline (h_{rgb}) no longer works (§2.1).

4.2 Methods based on head

We consider our four intermediate models (h , $h+h_{casia}+h_{cacd}$, h_{deepid} , $h \oplus h_{deepid}$) and a prior work DeepFace [1], [5].

We observe the same trend as described in the previous sections on the *val* set (§3.6, 3.10). Here, we focus on the comparison against DeepFace [5]. Even without a specialised face module, h already performs better than DeepFace (76.42% versus 46.66%, Original split). We believe this is for two reasons: (1) DeepFace only takes face regions as input, leaving out valuable hair and background information (§3.6), (2) DeepFace only makes predictions on 52% of the instances where the face can be registered. Note that h_{deepid} also do not always make prediction due to the failure to estimate the pose (17% failure on PIPA), but performs better than DeepFace in the considered scenario (68.06% versus 46.66%, Original split).

4.3 Methods based on body

We consider three of our intermediate models (b , $h+b$, $P = f+h+u+b$) and four prior arts (GlobalModel [1], PIPER [1], Pose [43], COCO [45]). Pose [43] and COCO [45] methods appeared after the publication of the conference version of this paper [2]. See table 7 for the results.

Our body cue b and Zhang et al.’s GlobalModel [1] are the same methods implemented independently. Unsurprisingly, they perform similarly (69.63% versus 67.60%, Original split).

Our $h+b$ method is the minimal system matching Zhang et al.’s PIPER [1] (83.36% versus 83.05%, Original split). The feature vector of $h+b$ is about 50 times smaller than PIPER, and do not make use of face recogniser or pose estimator.

In fact, PIPER captures the head region via one of its poselets. Thus, $h+b$ extracts cues from a subset of PIPER’s “GlobalModel+Poselets” [1], but performs better (83.36% versus 78.79%, Original split).

Methods since the conference version [2]

Pose by Kumar et al. [43] uses extra keypoint annotations on the PIPA *train* set to generate pose clusters, and train separate models for each pose cluster (PSM, pose-specific models). By performing a form of pose normalisation they have improved the results significantly: 2.27 pp and 10.19 pp over *naeil* on Original and Day splits, respectively.

COCO by Liu et al. [45] proposes a novel metric learning loss for the person recognition task. Metric learning gives an edge over classifier-based methods by enabling recognition of unseen identities without re-training. They further use Faster-RCNN detectors [58] to localise face and body more accurately. The final performance is arguably good in all four splits, compared to Pose [43] or *naeil* [2]. However, one should note that the face, body, upper body, and full body features in COCO are based on GoogleNetv3 [59] and ResNet50 [60] – the numbers are not fully comparable to all the other methods that are largely based on AlexNet.

4.4 Methods based on full image

We consider our two intermediate models ($P_s = P+s$, *naeil* = P_s+E) and Contextual [47], a method which appeared after the conference version of this paper [2].

Our *naeil* performs better than PIPER [1] (86.78% versus 83.05%, Original split), while having a 6 times smaller feature vector and not relying on face recogniser or pose estimator.

Methods since the conference version [2]

Contextual by Li et al. [47] makes use of person co-occurrence statistics to improve the results. It performs 1.97 pp and 12.81 pp better than *naeil* on Original and Day splits, respectively. However, one should note that Contextual employs a face recogniser DeepID [56]. We have found that a specialised face recogniser improves the recognition quality greatly on the Day split (§3.10).

4.5 Our final model *naeil2*

naeil2 is a weighted combination of *naeil* and h_{deepid} (see §3.11 for details). Observe that by attaching a face recogniser module on *naeil*, we achieve the best performance on Album, Time, and Day splits. In particular, on the Day split, *naeil2*

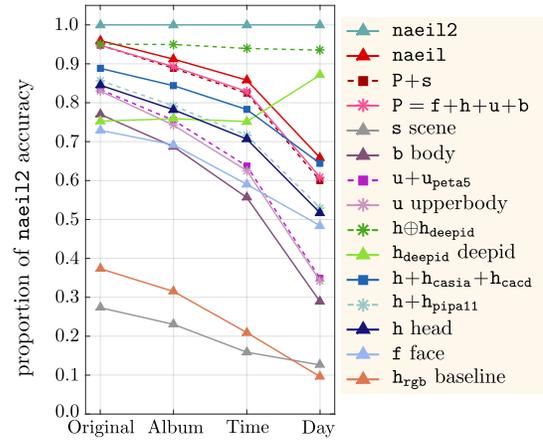


Fig. 7: PIPA *test* set relative accuracy of various methods in the four splits, against the final system *naeil2*.

makes a 8.85 pp boost over the second best method COCO [45] (table 7). On the Original split, COCO performs better (2.36 pp gap), but note that COCO uses more advanced feature representations (GoogleNet and ResNet).

Since *naeil2* and COCO focus on orthogonal techniques, they can be combined to yield even better performances.

4.6 Computational cost

We report computational times for some pipelines in our method. The feature training takes 2-3 days on a single GPU machine. The SVM training takes 42 seconds for h (4096 dim) and 1118 seconds for *naeil* on the Original split (581 classes, 6443 samples). Note that this corresponds to a realistic user scenario in a photo sharing service where ~ 500 identities are known to the user and the average number of photos per identity is ~ 10 .

5 ANALYSIS

In this section, we provide a deeper analysis of individual cues towards the final performance. In particular, we measure how contributions from individual cues (e.g. face and scene) change when either the system has to generalise across time or head viewpoint. We study the performance as a function of the number of training samples per identity, and examine the distribution of identities according to their recognisability.

5.1 Contribution of individual cues

We measure the contribution of individual cues towards the final system *naeil2* (§3.11) by dividing the accuracy for each intermediate method by the performance of *naeil2*. We report results in the four splits in order to determine which cues contribute more when there are larger time gap between training and testing samples and vice versa.

Results

See figure 7 for the relative performances in four splits. The cues based more on context (e.g. b and s) see greater drop from the Original to Day split, whereas cues focused on face f and head h regions tend to drop less. Intuitively, this is due to the greater changes in clothing and events in the Day split.

On the other hand, h_{deepid} increases in its relative contribution from Original to Day split, nearly explaining 90% of $naeil2$ in the Day split. h_{deepid} provides valuable invariant face feature especially when the time gap is great. However, on the Original split h_{deepid} only reaches about 75% of $naeil2$. Head orientation robust $naeil$ should be added to attain the best performance.

Conclusion

Cues involving context are stronger in the Original split; cues around face, especially the h_{deepid} , are robust in the Day split. Combining both types of cues yields the best performance over all considered time/appearance changes.

5.2 Performance by viewpoint

We study the impact of test instance viewpoint on the proposed systems. Cues relying on face are less likely to be robust to occluded faces, while body or context cues will be robust against viewpoint changes. We measure the performance of models on the head orientation partitions defined by a DPM head detector (see §2.2): frontal FR, non-frontal NFR, and no face detected NFD. NFD subset is a proxy for back-view and occluded-face instances.

Results

Figure 8 shows the accuracy of methods on the three head orientation subsets for the Original and Day splits. All the considered methods show worse performance from frontal FR to non-frontal NFR and no face detected NFD subsets. However, in the Original split, $naeil2$ still robustly predicts the identities even for the NFD subset ($\sim 80\%$ accuracy). On the Day split, $naeil2$ also do struggle on the NFD subset ($\sim 20\%$ accuracy). Recognition of NFD instances under the Day split constitutes the main remaining challenge of person recognition.

In order to measure contributions from individual cues in different head orientation subsets, we report the relative performance against the final model $naeil2$ in figure 9. The results are reported on the Original and Day splits. Generally, cues based on more context (e.g. b and s) are more robust when face is not visible than the face specific cues (e.g. f and h). Note that h_{deepid} performance drops significantly in NDET, while $naeil$ generally improves its relative performance in harder viewpoints. $naeil2$ envelops the performance of the individual cues in all orientation subsets.

Conclusion

$naeil$ is more viewpoint robust than h_{deepid} , making a contrast to the time-robustness analysis (§5.1). The combined model $naeil2$ takes the best of both worlds. The remaining challenge for person recognition lies on the no face detected NFD instances under the Day split. Perhaps image or social media metadata could be utilised (e.g. camera statistics, time and GPS location, social media friendship graph).

5.3 Generalisation across viewpoints

Here, we investigate the viewpoint generalisability of our models. For example, we challenge the system to identify a person from the back, having only shown frontal face samples.

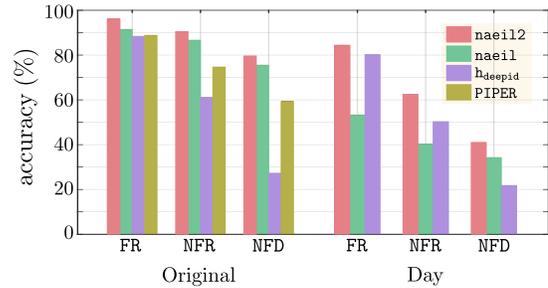


Fig. 8: PIPA *test* set accuracy of methods on the frontal (FR), non-frontal (NFR), and no face detected (NFD) subsets. Left: Original split, right: Day split.

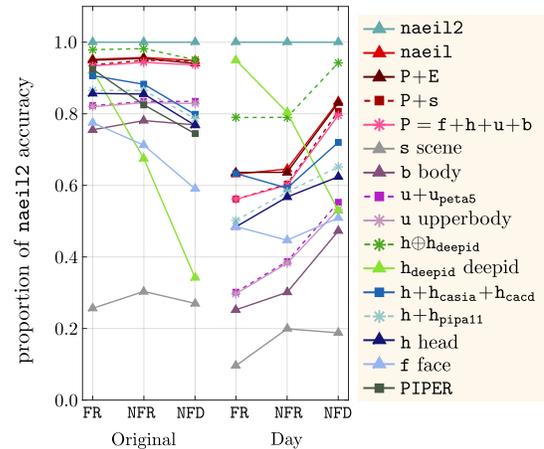


Fig. 9: PIPA *test* set relative accuracy of frontal (FR), non-frontal (NFR), and non-detection (NDET) head orientations, relative to the final model $naeil2$. Left: Original split, right: Day split.

Results

Figure 10 shows the accuracies of the methods, when they are trained either only on the frontal subset FR (left plot) or only on the no face detected subset NFD (right plot). When trained on FR, $naeil2$ has difficulties generalising to the NFD subset (FR versus NFD performance is $\sim 95\%$ to $\sim 40\%$ in Original; $\sim 85\%$ to $\sim 35\%$ in Day). However, the absolute performance is still far above the random chance (see §4.1), indicating that the learned identity representations are to a certain degree generalisable. The $naeil$ features are more robust in this case than h_{deepid} , with less dramatic drop from FR to NFD.

When no face is given during training (training on NFD subset), identities are much harder to learn in general. The recognition performance is low even for no-generalisation case: $\sim 60\%$ and $\sim 30\%$ for Original and Day, respectively, when trained and tested on NFD.

Conclusion

$naeil2$ does generalise marginally across viewpoints, largely attributing to the $naeil$ features. It seems quite hard to learn identity specific features (either generalisable or not) from back-views or occluded faces (NFD).

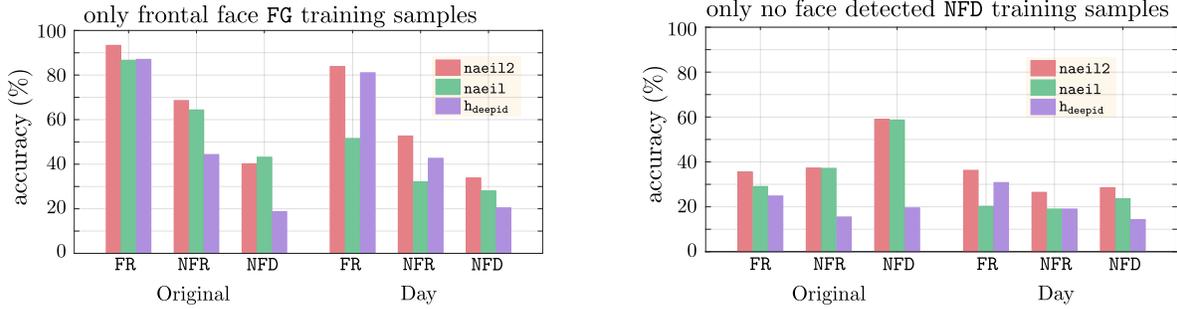


Fig. 10: PIPA *test* set performance when the identity classifier (SVM) is only trained on either frontal (FR, left) or no face detected (NFD, right) subset. Related scenario: a robot has only seen frontal views of people; who is this person shown from the back view?

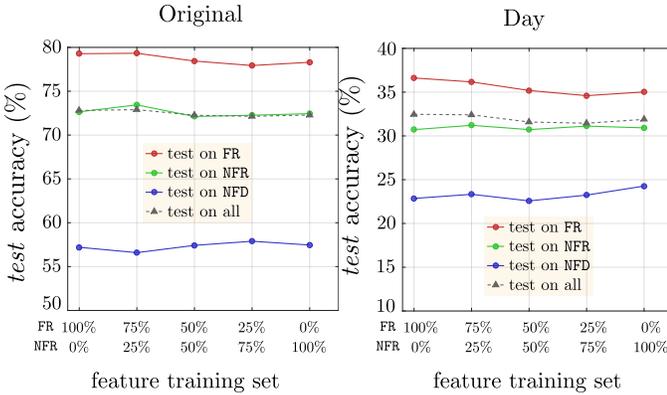


Fig. 11: Train the feature h with different mixtures of frontal FR and non-frontal NFR heads. The viewpoint wise performance is shown for the Original (left) and Day (right) splits.

5.4 Viewpoint distribution does not matter for feature training

We examine the effect of the ratio of head orientations in the feature training set on the quality of the head feature h . We fix the number of training examples that consists only of frontal FR and non-frontal faces NFR, while varying their ratio.

One would hypothesize that the maximal viewpoint robustness of the feature is achieved at a balanced mixture of FR and NFR for each person; also that h trained with FR (NFR) subset is relatively strong at predicting FR (NFR) subset (respectively).

Results

Figure 11 shows the performance of h trained with various FR to NFR ratios on FR, NFR, and NFD subsets. Contrary to the hypothesis, changing the distribution of head orientations in the feature training has $< 3\%$ effect on their performances across all viewpoint subsets in both Original and Day splits.

Conclusion

No extra care is needed to control the distribution of head orientations in the feature training set to improve the head feature h . Features on larger image regions (e.g. u and b) are expected to be even less affected by the viewpoint distribution.

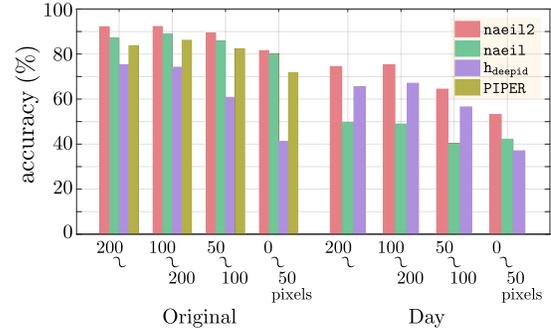


Fig. 12: PIPA *test* set accuracy of systems at different levels of input resolution. Resolution is measured in terms of the head height (pixels).

5.5 Input resolution

This section provides analysis on the impact of input resolution. We aim to identify methods that are robust in different range of resolutions.

Results

Figure 12 shows the performance with respect to the input resolution (head height in pixels). The final model $naeil2$ is robust against low input resolutions, reaching $\sim 80\%$ even for instances with < 50 pixel heads on Original split. On the day split, $naeil2$ is less robust on low resolution examples ($\sim 55\%$).

Component-wise, note that $naeil$ performance is nearly invariant to the resolution level. $naeil$ tends to be more robust for low resolution input than the h_{deepid} as it is based on body and context features and do not need high resolution faces.

Conclusion

For low resolution input $naeil$ should be exploited, while for high resolution input h_{deepid} should be exploited. If unsure, $naeil2$ is a good choice – it envelops the performance of both in all resolution levels.

5.6 Number of training samples

We are interested in two questions: (1) if we had more samples per identity, would person recognition be solved with the current method? (2) how many examples per identity are enough to gather substantial amount of information about a person? To investigate the questions, we measure the performance of methods

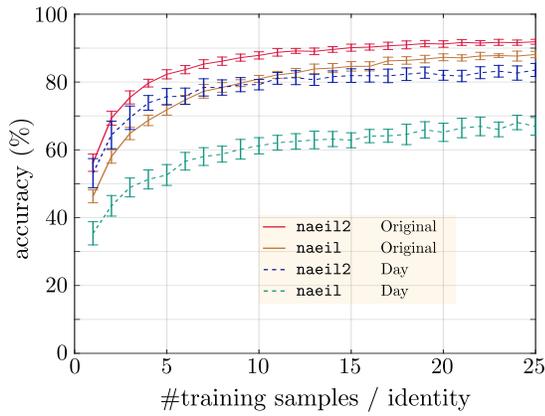


Fig. 13: Recognition accuracy at different number of training samples per identity. Error bars indicate ± 1 standard deviation from the mean.

at different number of training samples per identity. We perform 10 independent runs per data point with fixed number of training examples per identity (subset is uniformly sampled at each run).

Results

Figure 13 shows the trend of recognition performances of methods with respect to different levels of training sample size. `naeil2` saturates after 10 \sim 15 training examples per person in Original and Day splits, reaching $\sim 92\%$ and $\sim 83\%$, respectively, at 25 examples per identity. At the lower end, we observe that 1 example per identity is already enough to recognise a person far above the chance level ($\sim 67\%$ and $\sim 35\%$ on Original and Day, respectively).

Conclusion

Adding a few times more examples per person will not push the performance to 100%. Methodological advances are required to fully solve the problem. On the other hand, the methods already collect substantial amount of identity information only from single sample per person (far above chance level).

5.7 Distribution of per-identity accuracy

Finally, we study how much proportion of the identities are easy to recognise and how many are hopeless. We study this by computing the distribution of identities according to their per-identity recognition accuracies.

Results

Figure 14 shows the per identity accuracy for each identity in a descending order for each considered method. On the Original split, `naeil2` gives 100% accuracy for 185 out of the 581 test identities, whereas there was only one identity where the method totally fails. On the other hand, on the Day split there are 11 out of the 199 test identities for whom `naeil2` achieves 100% accuracy and 12 identities with zero accuracy. In particular, `naeil2` greatly improves the per-identity accuracy distribution over `naeil`, which gives zero prediction for 40 identities.

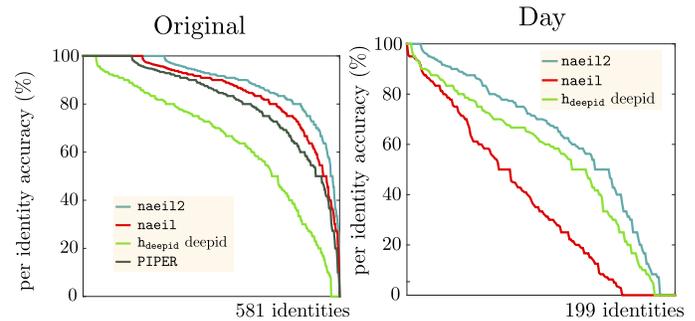


Fig. 14: Per identity accuracy of on the Original and Day splits. The identities are sorted according to the per identity accuracy for each method separately.

Conclusion

In the Original split, `naeil2` is doing well on many of the identities already. In the Day split, the `hdeepid` feature has greatly improved the per-identity performances, but `naeil2` still misses some identities. It is left as future work to focus on the hard identities.

6 CONCLUSION

We have analysed the problem of person recognition in social media photos where people may appear with occluded faces, in diverse poses, and in various social events. We have investigated efficacy of various cues, including the face recogniser DeepID2+ [3], and their time and head viewpoint generalisability. For better analysis, we have contributed additional splits on PIPA [1] that simulate different amount of time gap between training and testing samples.

We have made four major conclusions. (1) Cues based on face and head are robust across time (§5.1). (2) Cues based on context are robust across head viewpoints (§5.2). (3) The final model `naeil2`, a combination of face and context cues, is robust across both time and viewpoint and achieves a ~ 9 pp improvement over a recent state of the art on the challenging Day split (§4.5). (4) Better convnet architectures and face recognisers will improve the performance of the `naeil` and `naeil2` frameworks in the future §4.5).

The remaining challenges are mainly the large time gap and occluded face scenarios (§5.2). One possible direction is to exploit non-visual cues like GPS and time metadata, camera parameters, or social media album/friendship graphs. Code and data are publicly available at <https://goo.gl/DKuh1Y>.

ACKNOWLEDGMENTS

This research was supported by the German Research Foundation (DFG CRC 1223).

COPYRIGHT

© 2017 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.



Fig. 15: Success and failure cases on the Original split. Single images: test examples. Arrows point to the training samples for the predicted identities. Green and red crosses indicate correct and wrong predictions.



Fig. 16: Failure cases of `naeil2` and `PIPER` on the Original split. Single images: test examples. Arrows point to the training samples for the predicted identities. Green and red crosses indicate correct and wrong predictions. Typical hard cases are: 1) uniform clothing (top left), 2) babies (top right), 3) children (bottom left), and 4) annotation errors (bottom right).

APPENDIX A FACE DETECTION

For face detection we use the DPM detector from Mathias et al. [48]. This detector is trained on $\sim 15k$ faces from the AFLW database, and is composed of 6 components which give a rough indication of face orientation: $\pm 0^\circ$ (frontal), $\pm 45^\circ$ (diagonal left and right), and $\pm 90^\circ$ (side views). Figure 19 shows example face detections on the PIPA dataset. It shows detections, the estimated orientation, the regressed head bounding box, the corresponding ground truth head box, and some failure modes. Faces corresponding to $\pm 0^\circ$ are considered frontal (FR), and all others ($\pm 45^\circ$, $\pm 90^\circ$) are considered non-frontal (NFR). No ground truth is available to evaluate the face orientation estimation; except a few mistakes, the $\pm 0^\circ$ components seems a rather reliable estimators (while more confusion is observed between $\pm 45^\circ/\pm 90^\circ$).

APPENDIX B OPEN-WORLD RECOGNITION

In the main paper, we have focused on the scenario where the test instances are always from a closed world of gallery identities. However, for example when person detectors are used to localise instances, as opposed to head box annotations, the detected person may not be one of the gallery set. One may wonder how our person recognisers would perform when the test instance could be an unseen identity.

In this section, we study the task of “open-world person recognition”. The test identity may be either from a gallery set (training identities) or from a background set (unseen identities). We consider the scenario where test instances are given by a face detector [48] while the training instance locations have been annotated by humans.

Key challenge for our recognition system is to tell apart gallery identities from background faces, while simultaneously classifying the gallery identities. Obtained from a detector, the background faces may contain any person in the crowd or even non-faces. We will introduce a simple modification of our recognition systems’ test time algorithm to let them further make the gallery versus background prediction. We will then discuss the relevant metrics for our systems’ open-world performances.

B.1 Method

At test time, body part crops are inferred from the detected face region (\mathfrak{f}). First, \mathfrak{h} is regressed from \mathfrak{f} , using the PIPA *train* set statistics on the scaling and displacement transformation from \mathfrak{f} to \mathfrak{h} . All the other regions (\mathfrak{u} , \mathfrak{b} , \mathfrak{s}) are computed based on \mathfrak{h} in the same way as in §3.2 of main paper.

To measure if the inferred head region \mathfrak{h} is sound and compatible with the models trained on \mathfrak{h} (as well as \mathfrak{u} and \mathfrak{b}), we train the head model \mathfrak{h} on head annotations and test on the heads inferred from face detections. The recognition performance is 87.74%, while when trained and tested on the head annotations, the performance is 89.85%. We see a small drop, but not significant – the inferred regions to be largely compatible.

The gallery-background identity detection is done by thresholding the final SVM score output. Given a recognition system and test instance x , let $\mathcal{S}_k(x)$ be the SVM score for identity k . Then, we apply a thresholding parameter $\tau > 0$ to predict background if $\max_k \mathcal{S}_k(x) < \tau$, and predict the argmax_k gallery identity otherwise.

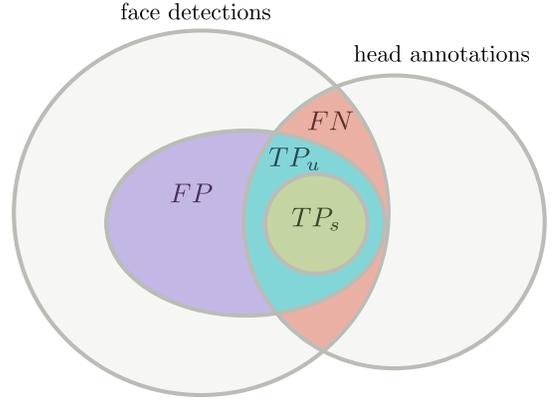


Fig. 17: Diagram of various subsets generated by a person recognition system in an open world setting (cf. Figure 2 of main paper). TP_s : sound true positive, TP_u : unsound true positive, FP : false positive, FN : false negative. See text for the definitions.

B.2 Evaluation metric

The evaluation metric should measure two aspects simultaneously: (1) ability to tell apart background identities, (2) ability to classify gallery identities. We first introduce a few terms to help defining the metrics. Refer to figure 17 for a visualisation. We say a detected test instance x is a “foreground prediction” if $\max_k \mathcal{S}_k(x) \geq \tau$. A foreground prediction is either a true positive (TP) or a false positive (FP), depending on whether x is a gallery identity or not. If x is a TP , it is either a sound true positive TP_s or an unsound true positive TP_u , depending on the classification result $\text{argmax}_k \mathcal{S}_k(x)$. A false negative (FN) is incurred if a gallery identity is predicted as background.

We first measure the system’s ability to screen background identities while at the same time classifying the gallery identities. The **recognition recall (RR)** at threshold τ is defined as follows

$$\text{RR}(\tau) = \frac{|TP_s|}{|\text{face det.} \cap \text{head anno.}|} = \frac{|TP_s|}{|TP \cup FN|}. \quad (2)$$

To factor out the performance of face detection, we constrain our evaluation to the intersection between face detections and head annotations (the denominator $TP \cup FN$). Note that the metric is a decreasing function of τ , and when $\tau \rightarrow -\infty$ the corresponding system is operating under the closed world assumption.

The system enjoys high RR when τ is decreased, but the system then predicts many background cases as foreground (FP). To quantify the trade-off we introduce a second metric: **false positive per image (FPPI)**. Given a threshold $\tau > 0$, FPPI is defined as

$$\text{FPPI}(\tau) = \frac{|FP_{\text{type1}}|}{|\text{images}|}, \quad (3)$$

measuring how many wrong foreground predictions the system makes per image. It is also a decreasing function of τ . When $\tau \rightarrow \infty$, the FPPI attains zero.

B.3 Results

Figure 18 shows the recognition rate (RR) versus false positive per image (FPPI) curves parametrised by τ . As $\tau \rightarrow \infty$, $\text{RR}(\tau)$ approaches the close world performance on the face detected subset ($\text{FR} \cup \text{NFR}$): 87.74% (Original) and 46.67% (Day) for *naeil*. In the open-world case, for example when the system

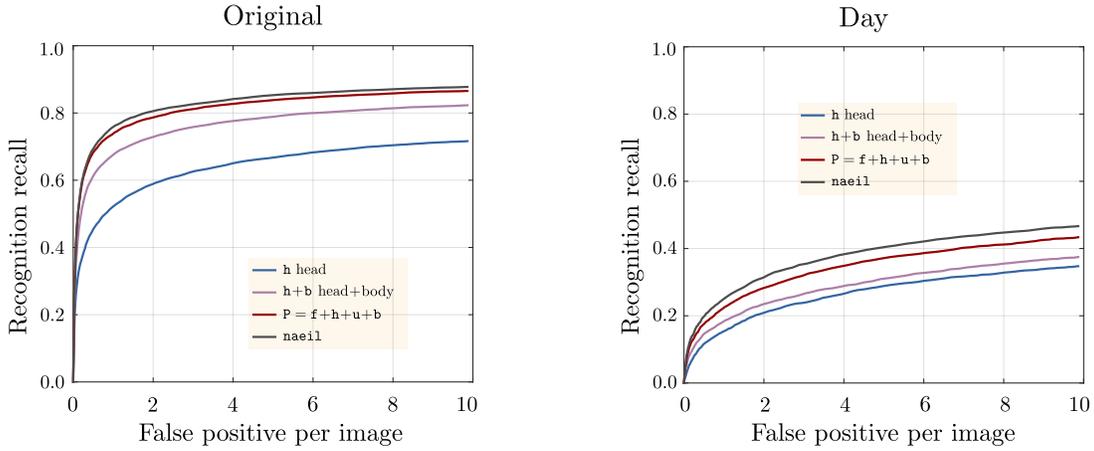


Fig. 18: Recognition recall (RR) versus false positive per image (FPPI) of our recognition systems in the open world setting. Curves are parametrised by τ – see text for details.

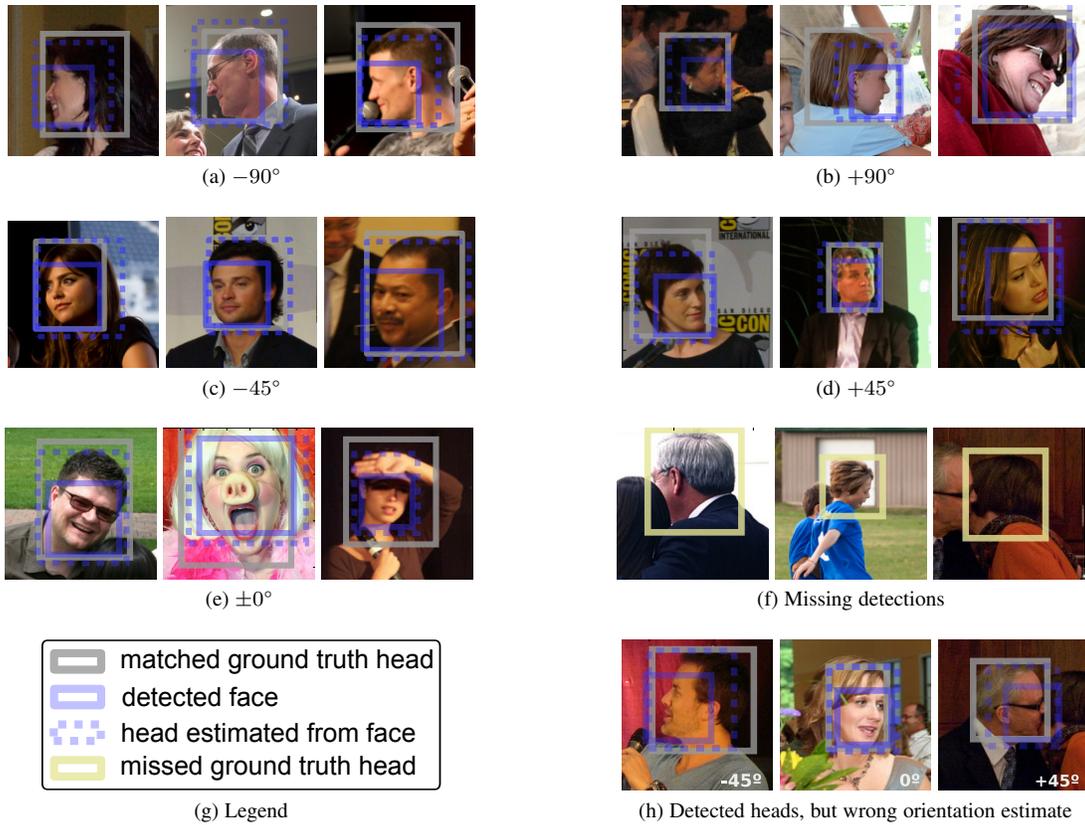


Fig. 19: Example results from the face detector (PIPA *val* set), and estimated head boxes.

makes one FPPI, the recognition recall for *naeil* is 76.25% (Original) and 25.29% (Day). Transitioning from the open world to close world, we see quite some drop, but one should note that the set of background face detections is more than $7\times$ greater than the foreground faces.

B.4 Conclusion

Although performance is not ideal, a simple SVM score thresholding scheme can make our systems work in the open world recognition scenario.

Note that the DeepID2+ [3] is not a public method, and so we cannot compute h_{deepid} features ourselves; we have not included the h_{deepid} or *naeil2* results in this section.

REFERENCES

- [1] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev, "Beyond frontal faces: Improving person recognition using multiple cues," in *CVPR*, 2015.
- [2] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele, "Person recognition in personal photo collections," in *ICCV*, 2015.
- [3] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," *arXiv*, 2014.
- [4] 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," UMass, Tech. Rep., 2007.
- [5] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014.
- [6] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of lfw benchmark or not?" *arXiv*, 2015.
- [7] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *arXiv*, 2015.
- [8] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.
- [9] J.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep cnn features," in *WACV*. IEEE, 2016, pp. 1–9.
- [10] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.
- [11] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *ICCV*, 2009.
- [12] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *CVPR*, 2013.
- [13] X. Cao, D. Wipf, F. Wen, and G. Duan, "A practical transfer learning algorithm for face verification," in *ICCV*, 2013.
- [14] C. Lu and X. Tang, "Surpassing human-level face verification performance on lfw with gaussianface," *arXiv*, 2014.
- [15] B. F. Klare, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," *algorithms*, vol. 13, p. 4, 2015.
- [16] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, vol. 3, no. 5. Citeseer, 2007.
- [17] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *BMVC*, vol. 1, no. 2. Citeseer, 2011, p. 6.
- [18] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *ACCV (1)*, 2012, pp. 31–44.
- [19] D. Hall and P. Perona, "Fine-grained classification of pedestrians in video: Benchmark and state of the art," in *CVPR*, 2015, pp. 5482–5491.
- [20] W. Li and X. Wang, "Locally aligned feature transforms across views," in *CVPR*, 2013.
- [21] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by saliency matching," in *ICCV*, 2013.
- [22] S. Bak, R. Kumar, and F. Bremond, "Brownian descriptor: A rich meta-feature for appearance matching," in *WACV*, 2014.
- [23] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.
- [24] D. Yi, Z. Lei, and S. Z. Li, "Deep metric learning for practical person re-identification," *arXiv*, 2014.
- [25] Y. Hu, D. Yi, S. Liao, Z. Lei, and S. Li, "Cross dataset person re-identification," in *ACCV, workshop*, 2014.
- [26] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," *Differences*, vol. 5, p. 25, 2015.
- [27] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *CVPR*, June 2016.
- [28] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, June 2016.
- [29] R. R. Viorator, B. Shuai, J. Lu, D. Xu, and G. Wang, *A Siamese Long Short-Term Memory Architecture for Human Re-identification*. Cham: Springer International Publishing, 2016, pp. 135–153. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46478-7_9
- [30] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *CVPR*, 2017.
- [31] A. Gallagher and T. Chen, "Clothing cosegmentation for recognizing people," in *CVPR*, 2008.
- [32] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *CVPR*, 2016, pp. 4873–4882.
- [33] A. Nech and I. Kemelmacher-Shlizerman, "Level playing field for million scale face recognition," in *CVPR*, 2017.
- [34] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person re-identification*. Springer, 2014.
- [35] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *IVC*, 2014.
- [36] J. Cui, F. Wen, R. Xiao, Y. Tian, and X. Tang, "Easyalbum: an interactive photo annotation system based on face clustering and re-ranking," in *SIGCHI*, 2007.
- [37] C. S. Mathialagan, A. C. Gallagher, and D. Batra, "Vip: Finding important people in images," *arXiv*, 2015.
- [38] M. Everingham, J. Sivic, and A. Zisserman, "Hello! my name is... buffy-automatic naming of characters in tv video," in *BMVC*, 2006.
- [39] —, "Taking the bite out of automated naming of characters in tv video," *IVC*, 2009.
- [40] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *ICCV*, 2009.
- [41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.
- [42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.
- [43] V. Kumar, A. Nambodiri, M. Paluri, and C. Jawahar, "Pose-aware person recognition," in *CVPR*, 2017.
- [44] 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," in *CVPR*, 2017.
- [45] Y. Liu, H. Li, and X. Wang, "Rethinking feature discrimination and polymerization for large-scale recognition," in *NIPS Workshop*, 2017.
- [46] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele, "Faceless person recognition; privacy implications in social media," in *ECCV*, 2016, to appear.
- [47] H. Li, J. Brandt, Z. Lin, X. Shen, and G. Hua, "A multi-level contextual model for person recognition in photo albums," in *CVPR*, June 2016.
- [48] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *ECCV*, 2014.
- [49] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, 2001.
- [50] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning Deep Features for Scene Recognition using Places Database." *NIPS*, 2014.
- [51] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv*, 2014.
- [52] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *ECCV*, 2014.
- [53] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *ACMMM*, 2014.
- [54] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *ICCV*, 2013.
- [55] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *arXiv*, 2015.
- [56] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *CVPR*, 2014, pp. 1891–1898.
- [57] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, *Bayesian Face Revisited: A Joint Formulation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 566–579. [Online]. Available: https://doi.org/10.1007/978-3-642-33712-3_41
- [58] R. G. J. S. Shaoqing Ren, Kaiming He, "Faster R-CNN: Towards real-time object detection with region proposal networks," *arXiv*, 2015.
- [59] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.
- [60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, June 2016.