Clothing Change Aware Person Identification

Jia Xue*[1] Zibo Meng*[2] Karthik Katipally[3] Haibo Wang[3] Kees van Zon[3]

[1]Department of Electrical and Computer Engineering, Rutgers University
[2]Department of Computer Science & engineering, University of South Carolina
[3]Philips Research North America, Cambridge

jia.xue@rutgers.edu, mengz@email.sc.edu, {karthik.katipally, haibo.wang,
kees.van.zon}@philips.com

Abstract

We develop a person identification approach - Clothing Change Aware Network (CCAN) for the task of clothing assisted person identification. CCAN concerns approaches that go beyond face recognition and particularly tackles the role of clothing to identification. Person identification is a rather challenging task when clothing appears changed under complex background information. With a pair of two person images as input, CCAN simultaneously performs a verification task to detect change in clothing and an identification task to predict person identity. When clothing from the pair of input images is detected to be different, CCAN automatically understates clothing information while emphasizing face, and vice versa. In practice, CCAN outperforms the way of equally stacking face and full body context features, and shows leading results on the People in Photo Album (PIPA) dataset.

1. Introduction

Person identification is a key task for many applications, such as access control, video surveillance, abnormal event detection and criminal identification. For person identification, face information plays a crucial role [18, 22, 23, 27] when near-frontal faces can be clearly captured by a camera. The typical workflow of a face recognition system consists of face detection, frontalization and similarity retrieval [23]. With the wide usage of deep convolutional neural networks (CNNs), 1:1 face verification and 1:N (N<1000) recognition are believed to be well-addressed and ready for certain commercial applications [22]. However, it remains challenging when frontal face images are not feasible or N is large. In this case, full-body recognition becomes complementary [4]. While early efforts tend to use full-body features, recent work shows that explicitly modeling local attributes greatly boosts performance [1, 11, 15, 31].

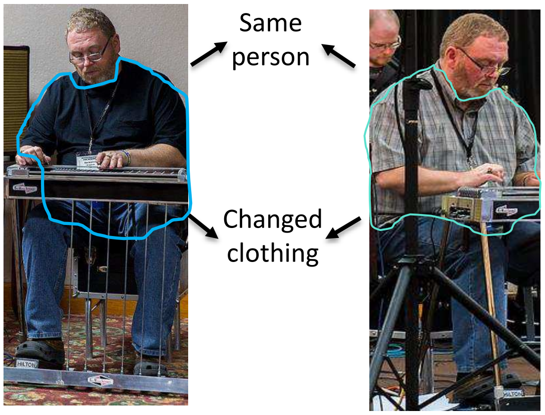

Figure 1: Face and body context are often concatenated for person identification. However, complications arise when people change their clothing. By explicitly modeling clothing change, we aim to improve identification accuracy especially when a person dresses differently.

In this paper, we address a specific type of person identification problem: dynamically registering a person in a uncontrolled scenario, and later on identifying the person in another uncontrolled scenario. The time interval between the registration and the identification stage can range from minutes to hours. Since both the registration and identification scenarios are uncontrolled, many subject and environmental challenges remain there, e.g., face visibility, body

* indicates equal contribution. This work was conducted when Jia, Zibo and Karthik were in their internship program at Philips Research North America, Cambridge, MA.

pose variation, illuminating, and partial occlusion. Face recognition alone, due to the uncontrolled visibility of face, is inadequate, and complementary full-body recognition becomes a must. For convenience we name the addressed problem *instant person identification*.

Such an instant identification task can be needed in many applications. A typical example is in hospitals. There is a recent trend to monitor the physiological status of patients via cameras [17, 25]. Instant person identification is critical for setups aimed at monitoring multiple patients simultaneously [24]. Similar application scenarios are hotels, banks or restaurants. In these scenarios, persons, typically customers, first go to a reception desk for registration. Here a camera is installed to capture photos of the customers as gallery set. Later on, these persons will stay in an area waiting for their service. With another camera capturing probe photos, instant person identification can be performed, as a basis for tasks like physiological status monitoring, emotion recognition, and abnormal event detection.

Note that instant person recognition can be treated as a sub-problem of person re-identification. It addresses instant appearance changes such as clothing and pose but excludes ones caused by aging. Moreover, we particularly focus on application scenarios where facial images are clear. Therefore, our work is more similar to the photo-level recognition problem [1, 11, 13, 15, 16, 31] and not the standard re-ID problem in video surveillance [32].

To handle the problem, intra-person variations such as lighting, pose and clothing must be alleviated so that inter-person differences can be enlarged. Convolutional neural networks (CNN) is shown to be able to well model deviations such as face angle, body pose and illumination [10, 13, 29]. However, clothing change is rarely addressed. In the aforementioned scenarios, clothing is actually changed very often. For example, a patient/customer often takes a jacket, hat or scarf off or puts one on for the reason of temperature difference between outdoor and indoor environments. When equally leveraging face and body information, which is typically used in literature [6], different clothing features tend to mislead the recognition result. In this case, face should play a more critical role while clothing should be understated. Recent efforts address the relative importance of face and body attributes by directly learning weights from training data [31]. However, this needs large training data and still does not model clothing change explicitly. The existing way to detect clothing change needs a clear segmentation of clothing from the rest of body [6], which is challenging in itself.

This paper presents a principled way, called Clothing Change Aware Network (CCAN), to explicitly model clothing change for person identification. CCAN takes a pair of features of two person photos as input. When clothing from the input pair is detected to be different, CCAN automatically adjusts the relative weights of face and body for identification. In this way, CCAN handles the intuition that clothing context should be understated when a person is found to have clothing changed. At the core of CCAN is a 3-layer neural network, which simultaneously performs a verification task to detect clothing change, and an identification task to predict person identity. The two tasks are coupled in the network in this way: on one hand, when clothing is changed, other unchanged contexts should be more employed; otherwise, clothing and other attributes should both be picked. In practice, CCAN outperforms the case of simply stacking face and body features in all experiments, and shows top results on the public People in Photo Album (PIPA) dataset [31].

Ahead of applying CCAN, we apply two other deep convolutional models to explicitly convert face and body images into feature vectors, respectively. For face converting, we apply the model suggested in [26] with slight modifications. For body converting, we randomly crop patches out of body image to feed into a single ResNet-50 network [9]. Compared to others [13, 15, 31], the random cropping eliminates the need of semantic attribute detection and reduces the number of needed deep models.

2. Related Work

Face recognition Before deep learning became popular, the three top-ranked Commercial Off The Shelf face recognition systems correctly matched probed faces against a large collection of 1.6 million identities with an 82%-92% accuracy rate [20]. However, when tested on a 1:N identification benchmark constructed using the LFW dataset [2], the rank-1 accuracy of the best system dropped to about 56%, even though the gallery has only a couple of thousand identities. In 2014, Facebook published a milestone paper that applied deep learning to face recognition [23] for the first time. The reported 1:1 verification accuracy on the LFW dataset reached 97.35%, 27% higher than the best counterparts. Since then, various deep learning models have been developed to address the problem [19, 22, 26, 28, 30]. On the standard LFW tests, the latest results have outperformed human beings [14]. It is widely believed that the 1:1 verification and the 1:N identification with $N<1000$ have been well solved. The remaining challenges are large-N identification and face recognition in the wild. The recent trend is to use more training data and develop lighter models for front-end applications. The public MegaFace [18] and MS-Cele-1M [8] challenges are the current leading datasets for large-scale face recognition benchmarking.

Person identification The work most relevant to ours is person recognition in photo albums [1, 11, 13, 15, 16, 31]. These daily life photos have rich variations such as age, pose, cluttered background, lighting and clothing, which together make identity recognition challenging. Anguelov

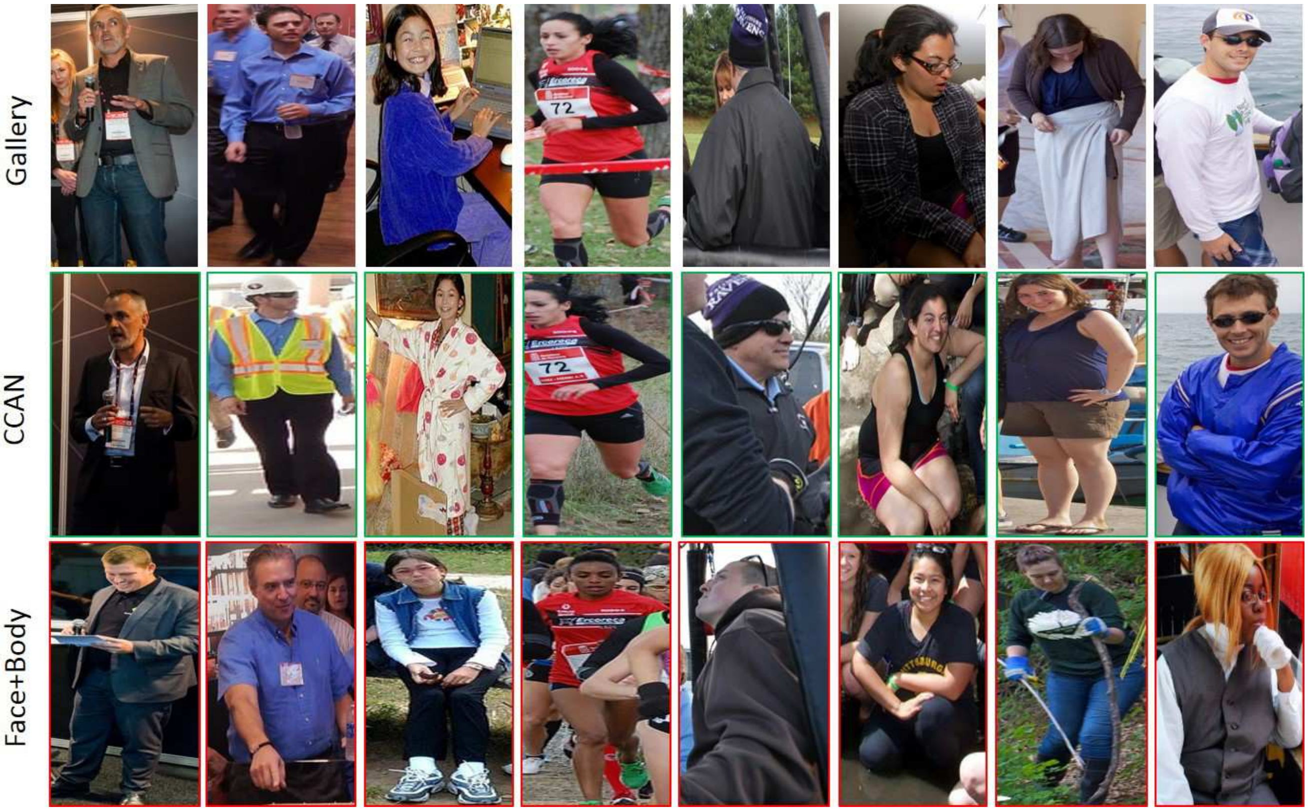

Figure 2: Some PIPA identities that CCAN predicts correctly while Face+Body (equally concatenating face and body features but without detecting clothing change, detailed in the experimental section) makes mistakes. Most of the cases have clothing changed. Without bewaring of the change, Face+Body still heavily relies on body information, leading to these errors.

et al. [1] addresses the problem by fusing all contextual cues based on a Markov random field framework. Unfortunately, only small-scale benchmarking was performed at that time. Recently, Zhang et al. [31] introduced the large-scale PIPA dataset for this task. They fuse a comprehensive list of body parts from poselet [3] and prove that context cues beyond faces help improve person recognition accuracy. Using deep learning models to compensate for pose is also a key of their approach. Oh et al. [11] thoroughly investigate the roles of various cues, including different body regions, scene context and long-term attributes such as age and gender. Li et al. [15] leverage more visual contexts, not only in person-level but also the contexts among persons in the same photo. While these efforts tend to fuse various contexts, they do not emphasize critical ones particularly. Most recently, Kumar et al. [13] particularly address body poses in person recognition. They tackled the problem by learning multiple models at specific poses. As a result, they achieved top results among others on the PIPA dataset. The latest proposed congenerous cosine loss [16] so far performs best on PIPA. Our approach particularly addresses the cue of clothing. Our logic is intuitive - when clothing is changed, clothing-related cues should be less important for the recognition task; otherwise, clothing is equally important as other contexts. The intuition is well proven by the top results that we achieved on the PIPA dataset.

3. CCAN

The workflow of CCAN is depicted in Figure 3 and explained in the caption. CCAN itself consists of three modules - a CNN model for face representation, another CNN for body context representation, and a 3-layer neural network for the prediction of clothing change and identity. Our face representation and body context representation models are based on ResNet [9] and is illustrated below. We adjust the face and body context representation models to increase cost-efficiency while preserving recognition performance.

3.1. Face Representation

We follow the CNN architecture introduced in [26] for face representation. At each block of the model, we add a residual shortcut to speed up model training. We also replace the Rectified Linear Units (ReLUs) [12] with a Max-Feature-Map (MFM) function [28] to shrink model size.

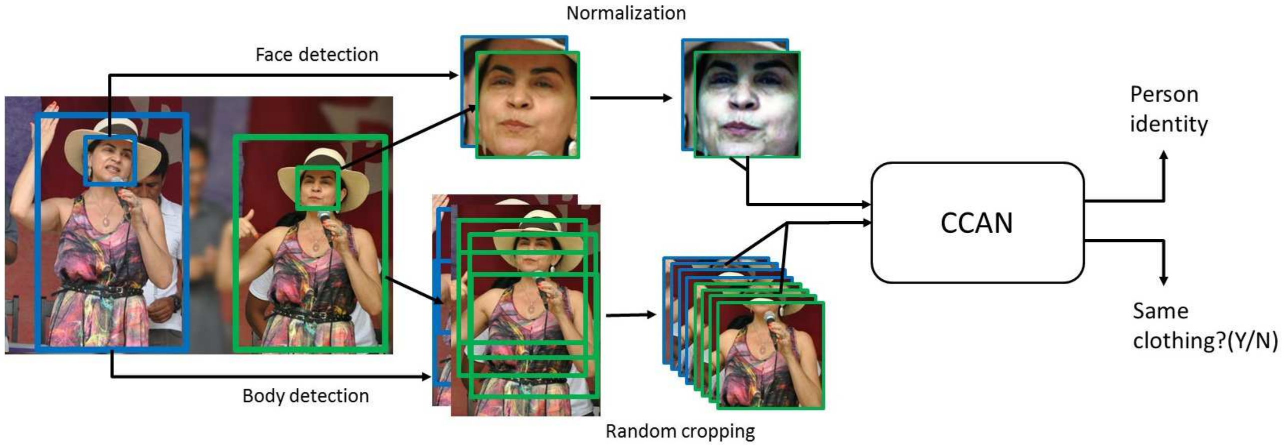

Figure 3: Workflow overview. CCAN jointly performs person identification and clothing change verification (checking if clothing has changed or not). Due to the existence of verification, CCAN requires an image pair of the same person as input. First, face and full body are detected and cropped on the image pair. Subsequently, on each of the cropped body image, 4 patches covering different body portions are further randomly cropped out. Finally, CCAN takes input of the cropped body patches (totally 4×2=8) and the cropped face pair, and predicts person identity and clothing change status simultaneously.

The loss function we employed is a combination of Softmax loss and Center loss [26].

Implementation: The network takes a 112x112 cropped face image as input. The face image is first aligned using the detection and alignment tools provided in Dlib [1]. When the alignment fails, we keep the unaligned face images for both training and testing. 100×100 patches are then randomly cropped from the 112×112 input and fed into the network. Each pixel (in [0, 255]) in the RGB images is normalized by subtracting 127.5. The dataset used for training is MegaFace [18] that includes 4.7 million images of 672 thousand identities. In practice we only select 20,000 identities that have the most instances, resulting in roughly 1.2 million training images. The model is trained for 1 million iterations using an initial learning rate of 0.01 and multiplied by 0.1 after every 200,000 iterations. On an NVIDIA Titan X card, the training takes 180 hours. The size of the trained model is 14.4M, only 1/40 of that of VGG-face [19] and similar to the Lightended models [28].

3.2. Body Context Representation

The body region has rich contextual information such as clothing style, gesture and hand poses. To capture these rich contexts, as shown in Figure 3, we perform full body detection and randomly crop 4 patches out of each detected body image. Unlike [31], random cropping saves the need of training various local attribute detectors, which reduces the number of used models. Meanwhile, with a high probability, random cropping covers both clothing and clothing-independent contexts, which is critical for the subsequent clothing change aware feature fusion.

A single model is used to represent each cropped body patch. For this purpose, we fine-tune a ResNet-50 [9] model that was pretrained on ImageNet [5]. First we add a fully-connected layer on top of the global average pooling layer of the ResNet-50 model. This helps us reduce the output feature dimension from 2048 to 512. We then add a classification layer on the top. Data employed for the fine tuning is the training split of the PIPA dataset [31].

Input to the fine-tuned network are 4 cropped patches of size 224×224. Given a detected body image we first resize its short side to 256 while keeping its Height/Width ratio. We then generate random crops by arbitrarily sampling a [0.8, 1.0] portion of the resized image and picking a random aspect ratio out of [3/4, 4/3]. We use a batch size of 64 to fine-tune the pre-trained ResNet-50 model. The learning rate is initialized at 0.01, and multiplied by 0.1 after every 80 epochs. The fine tuning takes 150 epochs in total. In the test phase, the feature representation used is the output of the added fully connected layer, which is 512-dimensional. Thus, the final length of body features is 512×4=2048.

3.3. Clothing Change Aware Identification

Once face and body contextual features are generated, they are fed into the subsequent Clothing Change Aware Network (CCAN), which performs identity recognition and clothing change detection simultaneously. The two tasks are coupled in such a way that CCAN learns shared features that are identity-friendly, especially when clothing change

[1] http://dlib.net/

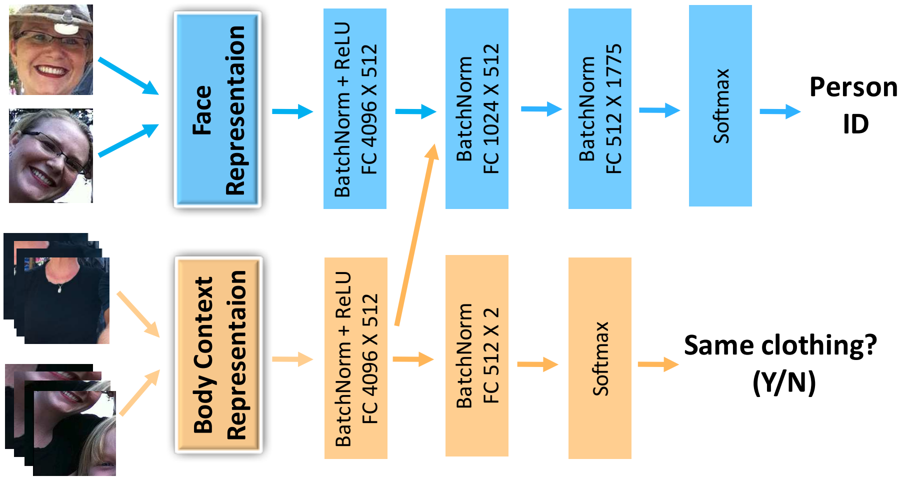

Figure 4: Detailed overview of CCAN. There are two parallel streams. The blue stream takes two face images of the same individual as input. The *FaceRepresentation* module converts the face images to feature vectors, as described in Section 3.1. The converted features then go through a 4-layer fully-connected network for getting a predicted identity. The yellow stream takes 4×2=8 body patches of the same individual as input. The *BodyContextRepresentation* module converts the patches to feature vectors, as described in Section 3.2. After that is a 3-layer fully-connected network, taking the converted body features as input and outputting if clothing is changed or not. There is a path linking the yellow *FC(4096×512)* layer to the blue *FC(1024x512)* layer, which associates the identity and labeled clothing information to together influence the learning of all layers. Best viewed in color.

is detected. Identity recognition is a multi-class classification problem, the output corresponding to the number of identities. Clothing change detection is a verification task, with the binary output being *changed* or *not changed*. To join the two tasks, we need patch pairs as input. Below are the details.

Architecture: Figure 4 details the architecture of CCAN. There are two parallel streams corresponding to the two tasks. The first stream is a 4-layer fully-connected network performing person identification. The output length corresponds to the number of identities in the training set. The second one is a 3-layer fully-connected network performing clothing change verification. The output length is 2, for YES and NO. We tried adding more layers but did not see any improvement. The two networks are associated by adding a path from the yellow FC(4096×512) layer to the blue FC(1024x512) layer. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ represent the extracted face and body features, respectively. The combination of face and body features is defined as

$$f(x) = \begin{bmatrix} F(\mathbf{x}_1, \mathbf{w}_1^1) \\ F(\mathbf{x}_2, \mathbf{w}_1^2) \end{bmatrix}^\intercal \begin{bmatrix} \mathbf{w}_2^1 \\ \mathbf{w}_2^2 \end{bmatrix}. \tag{1}$$

The function $F(\mathbf{x}_1, \mathbf{w}_1^1)$ and $F(\mathbf{x}_2, \mathbf{w}_1^2)$ represent the 1st fully connected layers on the two parallel streams, respectively - that is, the yellow and blue FC(4096×512) layers. Denote $F(\mathbf{x}_1, \mathbf{w}_1^1) = \sigma(\mathbf{w}_1^1 \mathbf{x}_1)$ and $F(\mathbf{x}_2, \mathbf{w}_1^2) = \sigma(\mathbf{w}_1^2 \mathbf{x}_2)$, where σ denotes ReLU [12] and the biases are omitted for simplifying notations. Since the outputs from $F(\mathbf{x}_2, \mathbf{w}_1^2)$ are also employed for clothing change verification, they will distribute at different statue if the input body features $\mathbf{x}_2$ are from different statue (same clothing vs different clothing). Consequently, the identity and clothing information are coupled to influence all the layers through gradient back-propagation. In this way, the model learns identity-friendly features, which may be independent of clothing if change is detected.

Training: CCAN accepts as an input a pair of two face-body-stacked features of the same person. If the paired features have differently labeled clothing, they are a positive instance; otherwise, they are negative. We use a batch size of 128, which is actually 128 feature pairs. To form these pairs, we first randomly select 128 features out of the training set. From the identity associated with each selected feature, we then randomly choose another feature of the same identity to form a training pair. If a person only has one feature, it is duplicated to form a pair.

Inference: In the test phase, there are gallery and probe sets, both of which are never used for training. Therefore, we cannot rely on the predicted identity of CCAN. Instead we only use intermediate output as feature representation. Specifically, we use the output of the blue FC(1024x512) layer as features. The gallery features will be pre-extracted. Each probe feature is then matched against the gallery set, and the best match will be picked. Similarly to training, test needs a pair of two input face-body-stacked features of the same individual from either probe or gallery. In the training phase we do not consider the order of the 4 input body features. Therefore, to enhance performance, we do a

Method	Face	Body Context	#Models	Accuracy
PIPER [31]	✓	✓	109	83.05%
naeil [11]	✓	✓	6	86.78%
Li et al. [15]	✓	✓	6	88.75%
PSM [13]	✓	✓	7	89.05%
COCO [16]	✓	✓	4	92.78%
Face	✓	-	1	85.05%
Body Context	-	✓	1	86.57%
Face + Body	✓	✓	2	90.86%
CCAN	✓	✓	3	92.59%

Table 1: Person recognition result on PIPA [31] test set. Note that we get comparable state-of-art result with fewer model.

comprehensive image pairing. Let $\{\mathbf{x}\}_{1:N}$ denote the face-body-stacked feature set belonging to an individual. Each time we first pick a feature $\mathbf{x}_i$ with i from 1 to N. Then we sequentially pair $\mathbf{x}_i$ with $\mathbf{x}_j$ with j from 1 to N, and feed $(\mathbf{x}_i, \mathbf{x}_j)$ into CCAN for feature extraction. Totally we get N such feature vectors for each i. Subsequently we average these N features and treat the averaged vector as the representation of $\mathbf{x}_i$. For individuals having only one face-body-stacked feature, we simply duplicate the feature for pairing. Finally we get N such averaged feature vectors corresponding to the N face-body-stacked inputs. In our experiments N differs for different identities. We used two different decision metrics for the probe-gallery set matching:

- **Averaging:** We average all features of each identity and perform an average-average vector matching;

- **Majority Voting:** Each feature of a probe identity is matched against each individual feature of each gallery identity. The gallery identity which is most often voted in the individual matchings, is the best match.

Implementation: Training data is the train split of the PIPA dataset [31]. We manually label the clothing information of all the 36,352 images of 1775 identities. The model is initialized with the Xavier distribution [7]. The initial learning rate is 0.01, and multiplied by 0.1 after every 12 epochs. The training totally takes 25 epochs. On a NVIDIA TITAN 1080TI card, the training takes less than 1 hour.

4. Experiments

We conducted experiments on two public datasets - the People in Photo Albums (PIPA) dataset [31] and the Celebrity In Place (CIP) dataset [30]. Only the setting of person identification was covered. In all experiments we only use the training split of the PIPA dataset for model training and validation. The rest are used for evaluation. All of our results were evaluated on the same machine.

4.1. Results on PIPA

PIPA contains 63,188 instances of 2,356 identities. Originally only face regions were labeled. To crop body regions, we apply person detection based on Faster R-CNN [21]. Then we compute the overlap area between each detected body and each labeled face region. If the overlap is larger than 75% of the face label, we say the face and body are from the same person. For each identity we randomly assign each of his/her photos as probe or gallery. No image appears in both gallery and probe. If a person only has one photo, the person will be added to gallery as a distractor.

Table 1 shows the results of comparing CCAN with five recent approaches and three baselines. ‘Face’ indicates using face feature only. Similarly, ‘Body Context’ indicates using body features only. ‘Face + Body’ refers to simply concatenating face and body features. CCAN shows top results, very close to the best result from COCO [16], which was reported very recently. For the result, CCAN uses three deep learning models while COCO relies on four.

Table 2 shows CCAN under two different matching metrics. CCAN consistently outperforms all the baselines. Especially, Body-only appears to drop more dramatically under the majority voting rule, which is also the leading reason of the significant drop of Face+Body. However, CCAN appears to be more robust to the different rules.

Figure 2 shows examples that CCAN correctly predicts while Face+Body fails at. Most of the cases have clothing changes, which Face+Body is not aware of. Figure 5 shows examples that CCAN fails to identify. In these cases, not only clothing but also face appearance vary a lot.

4.2. Results on CIP

The CIP dataset contains over 38k images with 4,611 celebrities involved in 16 places. The dataset was originally collected for celebrity retrieval but perfectly fits our research owing to the clothing changes at different places. We did the same steps as on PIPA to get full body labeling. For each identity we randomly assign each of his/her pictures to gallery or probe. The identities only having one

Method	Averaging	Majority Voting
Face	85.0%	82.4%
Body Context	86.6%	78.1%
Face + Body	90.9%	84.3%
CCAN	92.6%	92.4%

Table 2: Person recognition results on PIPA test set under two different decision metrics.

Method	Averaging	Majority Voting
Face + Body	30.4%	N/A
CCAN	35.6%	35.8%

Table 3: Person recognition results on CIP test set under two different decision metrics.

picture are simply added as gallery distractors. Totally we get 4,104 gallery identities and 3,176 probe identities.

On the dataset we only compared CCAN with the approach of concatenating face and body features (Face+Body). The results are shown in Table 3. CCAN outperforms Face+Body by 5%. On the dataset the averaging and majority voting metrics show similar results. Note that the overall accuracy is apparently lower than that of PIPA due to three reasons. First, CCAN is trained on the training split of PIPA, and not fined-tuned for CIP; Second, CIP images are captured at six very different places, therefore variations being large (examples shown in Figure 5); Third, many CIP identities have only one gallery or probe photo, which severely affects matching robustness. Note that the Majority Voting rule was not applied to Face+Body, mainly due to its long running time (over 10 days).

4.3. Results on PIPA+CIP

The last experiment we did is follow the PIPA protocol but incrementally adding CIP as gallery distractors. The purpose is to test the accuracy of CCAN under different gallery sizes. The decision metric used is *averaging*.

Figure 6 summarizes the results. When 100 distractors are added, CCAN beats Face+Body by 2%, similar to that of no distractors (see Table 1). As distractors grows to [300, 700], CCAN and Face+Body perform similarly, indicating that distractors have a key side effect. Alongside the further growth of distractors, the gain of CCAN goes up from 2% to 4%, showing that modeling clothing change in CCAN is more helpful for large-scale gallery set.

5. Concluding Remarks

This paper presents CCAN, a deep learning approach for person identification in the wild. CCAN is logically straightforward - when clothing is changed, we should suppress the role of clothing; otherwise, clothing should be equally used to boost the identification accuracy. CCAN handles the logic automatically, without needing to do any extra heuristic judgment. The advantage of CCAN was well verified on the public PIPA [31] and CIP [30] dataset. Future work will be to 1) fine-tune the body context model on larger datasets and 2) apply CCAN to a realistic scenario, e.g. recognizing patients in a hospital.

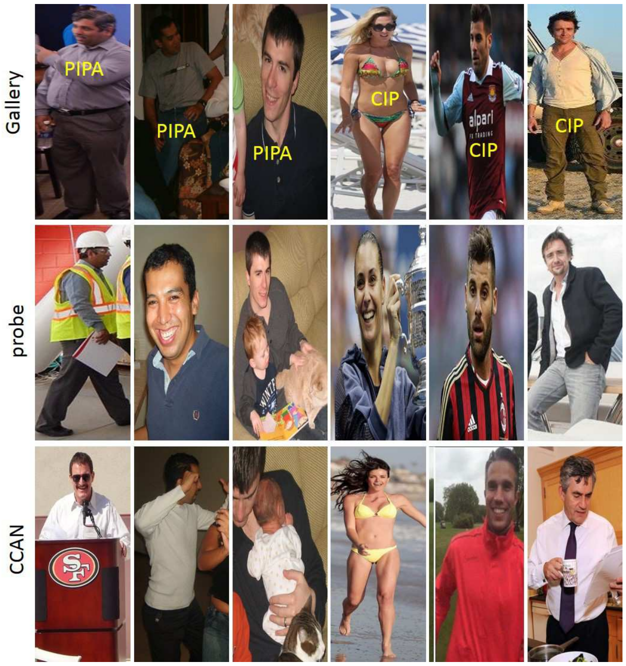

Figure 5: Examples that CCAN mis-recognized. CIP has more significant clothing changes. Since CCAN is trained on PIPA, it does not tackle well the changed CIP clothings.

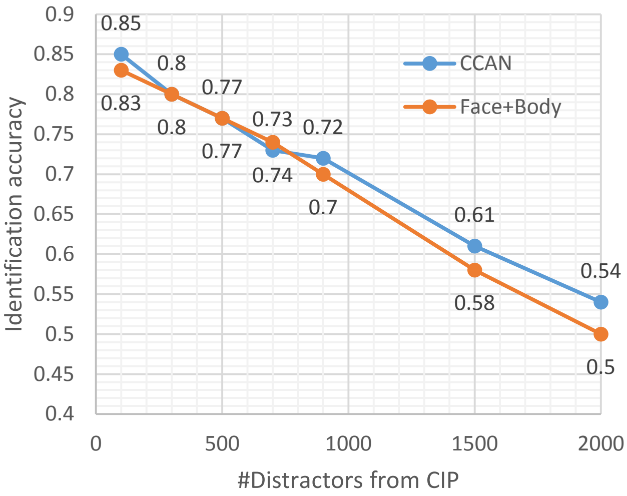

Figure 6: Results of adding CIP as gallery distractors to PIPA. Each time we randomly picked 100, 300, 500, 700, 900, 1,500 and 2,000 CIP distractors. Each experiment is repeated five times and the average accuracy is reported.

References

[1] D. Anguelov, K. c. Lee, S. B. Gokturk, and B. Sumengen. Contextual identity recognition in personal photo albums. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2007. 1, 2, 3

[2] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Transactions on Information Forensics and Security*, 9(12):2144–2157, 2014. 2

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision*, 2009. 3

[4] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):955–966, 1995. 1

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 4

[6] A. C. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 2

[7] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010. 6

[8] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *European Conference on Computer Vision*, 2016. 2

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 3, 4

[10] S. M. S. I. Kemelmacher-Shlizerman, S. Suwajanakorn. Illumination-aware age progression. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[11] S. Joon Oh, R. Benenson, M. Fritz, and B. Schiele. Person recognition in personal photo collections. In *International Conference on Computer Vision*, pages 3862–3870, 2015. 1, 2, 3, 6

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3, 5

[13] V. Kumar, A. Namboodiri, M. Paluri, and C. Jawahar. Pose-aware person recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3, 6

[14] E. Learned-miller, G. Huang, A. Roychowdhury, H. Li, G. Hua, E. Learned-miller, G. B. Huang, A. Roychowdhury, and H. Li. Labeled faces in the wild: A survey, 2016. 2

[15] H. Li, J. Brandt, Z. Lin, X. Shen, and G. Hua. A multi-level contextual model for person recognition in photo albums. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1305, 2016. 1, 2, 3, 6

[16] Y. Liu, H. Li, and X. Wang. Learning deep features via congenerous cosine loss for person recognition. *arXiv preprint arXiv:1702.06890*, 2017. 2, 3, 6

[17] D. J. McDuff, J. R. Estepp, A. M. Piasecki, and E. B. Blackford. A survey of remote optical photoplethysmographic imaging methods. In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6398–6404, 2015. 2

[18] A. Nech and I. Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 4

[19] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *British Machine Vision Conference*, volume 1, 2015. 2, 4

[20] G. W. Q. Patrick J. Grother and P. J. Phillips. The mbe 2010 report on the evaluation of 2d still-image face recognition algorithms. *NIST report on Multiple-Biometric Evaluation (MBE)*, 2010. 2

[21] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 6

[22] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 1, 2

[23] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2014. 1, 2

[24] H. Wang, K. van Zon, I. Kirenko, and M. Rocque. Monitoring patients in the wild. In *IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 997–997, 2017. 2

[25] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017. 2

[26] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515, 2016. 2, 3, 4

[27] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 529–534. IEEE, 2011. 1

[28] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *arXiv preprint arXiv:1511.02683*, 2015. 2, 3, 4

[29] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *International Conference on Computer Vision*, 2017. 2

[30] R. A. Yujie Zhong and A. Zisserman. Faces in places: compound query retrieval. *Proceedings of the British Machine Vision Conference (BMVC)*, pages 56.1–56.12, September 2016. 2, 6, 7

[31] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using

multiple cues. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4804–4813, 2015. 1, 2, 3, 4, 6, 7

[32] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. 2