Correspondence Driven Adaptation for Human Profile Recognition

Ming Yang[1], Shenghuo Zhu[1], Fengjun Lv[2], Kai Yu[1]
[1]NEC Laboratories America, Inc. Cupertino, CA 95014
{myang,zsh,kyu}@sv.nec-labs.com
[2]Huawei Technologies (USA) Santa Clara, CA 95050
felix.Lv@huawei.com

Abstract

Visual recognition systems for videos using statistical learning models often show degraded performance when being deployed to a real-world environment, primarily due to the fact that training data can hardly cover sufficient variations in reality. To alleviate this issue, we propose to utilize the object correspondences in successive frames as weak supervision to adapt visual recognition models, which is particularly suitable for human profile recognition. Specifically, we substantialize this new strategy on an advanced convolutional neural network (CNN) based system to estimate human gender, age, and race. We enforce the system to output consistent and stable results on face images from the same trajectories in videos by using incremental stochastic training. Our baseline system already achieves competitive performance on gender and age estimation as compared to the state-of-the-art algorithms on the FG-NET database. Further, on two new video datasets containing about 900 persons, the proposed supervision of correspondences improves the estimation accuracy by a large margin over the baseline.

1. Introduction

In recent years, intelligent video analysis systems, *e.g.* gender and age estimation for customer profiling, face verification and recognition, and wide-area surveillance, have been steadily improved by advances in computer vision and machine learning technologies. In general, certain statistical models are learned offline from a huge amount of training data in the development stage. When being deployed to real-world environments, however, such systems are often confronted by the model mismatch issue, that is, the performance degradation stemmed from the fact that training data can hardly cover the large variations because of different illumination conditions, image quality, and noise, *etc.* It is extremely hard, if not impossible, to collect sufficient training data in that the possible variations are unpredictable in diverse scenarios. Thus, it is desirable to adapt the statistical models in visual recognition systems to their specific deployment environments, in order to enhance the systems' generalization capability.

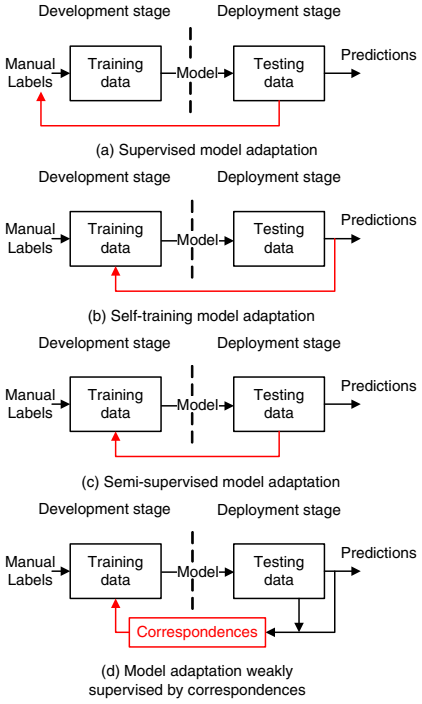

Figure 1. Different strategies to adapt a pre-trained statistical model after the deployment of a visual recognition system.

To address this model mismatch issue, people have developed various strategies. The most straightforward way is to obtain the ground truth labels of the testing data in the deployment scene and utilize them to perform supervised model adaptation, as shown in Fig. 1(a). Nevertheless, manual labels are costly and sometimes impractical to obtain after system deployment. Or we can trust the predictions and directly employ them to adapt the models in a self-training manner illustrated in Fig. 1(b). However, these direct positive feedbacks are very risky in practice which may result in model drift. An alternative way is to explore the structure and distances of unlabeled data using semi-supervised learning approaches [3] as in Fig. 1(c). Still,

whether the heuristic distance metric can capture the correct underlining structure of unlabeled data is in question.

In visual recognition systems, although true labels are generally unavailable after deployment, some weak supervision, such as the correspondence or co-occurrence of objects in successive frames, may be easier to obtain to mitigate the mismatch issue. For instance, in surveillance videos, by taking advantage of human tracking, the correspondences of persons in consecutive frames are usually available, which shall be very useful for tasks like human profile recognition because a person's profile barely changes in such a short span of time. It is reasonable to enforce the system to output consistent estimations for the same person. Therefore, we propose to leverage the correspondences as weak supervision to conduct online adaptation of pre-trained models. As shown in Fig. 1(d), we refer this new strategy by *correspondence driven adaption*.

We substantialize this idea in a fully automatic human profile recognition system on surveillance videos, which recognizes people's gender, age and race in real-time. Recognition of these biometric traits can greatly help applications like face verification and recognition [21, 14], digital signage [1], and retail customer analysis. Our baseline system uses advanced convolutional neural networks (CNNs) [17], which achieves competitive performance of gender and age estimation against the state-of-the-art algorithms on the FG-NET database [6]. The baseline system achieves accuracy 83.53% for gender recognition and mean absolute error (MAE) of 4.88 for age estimation. Upon these baseline models, we extract the correspondences of faces by tracking, and propose an online stochastic training approach for neural networks to impose the correspondence constraint, *i.e.* the outputs of the neural networks shall be consistent and stable for the same person. Evaluated on two video datasets containing about 900 persons, the proposed method achieves substantial improvement on gender, age, and race recognition accuracy over the baseline system.

The contributions of the paper are three-fold. First, the correspondence driven adaptation is a general method to enhance recognition systems for videos and adapt pre-trained models from development to deployment. Second, we derive a stochastic training approach for neural networks to realize this strategy. Furthermore, applying this method to a human profile recognition system, we demonstrate that the recognition accuracy can be improved automatically without human intervention. To our best knowledge, this correspondence driven adaptation for CNN models is novel. Moreover, most of existing gender/age/race recognition algorithms focus on static face images, which usually have much higher image resolution and quality than faces in surveillance videos. As far as we know, real-time gender/age/race estimation in videos are rarely investigated and reported in the literature before.

2. Related Work

Online model adaptation in video analysis has been explored in a variety of tasks such as video object classification [25], recognition [18], and visual tracking [23, 19, 2, 10], *etc*. Most of approaches adopt the self-training or semi-supervised strategies similar as Fig. 1(b) and (c) to conduct model adaptation. For example, [23, 19, 18] employ the inferred object in the current frame to incrementally update object appearance subspaces or manifolds; or [2, 10] collect new positive and negative samples according to their distances to the tracked object and learn boosting classifiers online. While, we do not directly use the recognition results of the pre-trained models but leverage the object correspondences in videos to online adapt the original models to yield stable recognition results.

We apply the correspondence driven adaptation to human profile recognition including estimation of gender, age and race from facial images, which are representative recognition tasks of binary classification, regression, and multiclass classification. They have attracted considerable research interests for decades [9, 20, 15], yet remain challenging problems, especially the age estimation, since the aging facial patterns are highly variable and influenced by many factors like gender, race, and living styles [11]. Thus, sophisticated representations and a huge amount of training data are required to tackle these problems in real-world applications. Age estimation has been formulated as a multi-class classification [16] or regression problem [26, 7, 11, 27, 24], using raw images [26], the anthropometric model [15], subspace [8] and manifold learning [7, 11], Gaussian mixture models of local image patches [27], biologically inspired features of Gabor filter banks [12], and active appearance models [5, 24]. All of these work focus on age estimation from still face images with relatively high resolution and low noise level. While, we learn gender, age, and race models using deep neural networks for surveillance quality videos, where the faces may be as small as 20 pixels.

The closest work to ours is the semi-supervised learning method in [4], which successfully collects training data with large variations from movies utilizing video tracks and *offline* re-trains SVMs of facial attributes. In contrast, our goal is to update a pre-trained recognition model in real-time after deployment to adapt it to new environments. We utilize the weak supervision of object correspondences and propose an *online* stochastic algorithm to adapt neural networks based human profile recognition models.

3. Correspondence Driven Adaptation

In this section, we start with a supervised adaptation framework for human profile recognition, and extend it to correspondence driven adaptation using stochastic training. Then, we discuss the relation and difference between corre-

spondence driven adaptation and the graph Laplacian based semi-supervised learning.

3.1. Supervised adaptation

Denote a recognition model as a score function $f(\mathbf{x}, \theta)$, where $\mathbf{x}$ is the input, *i.e.*, the raw image patch of a face, θ denotes the parameters of the recognition model (*e.g.*, a CNN model in the paper). To convert the score $f(\mathbf{x}, \theta)$ of a raw image $\mathbf{x}$ to the actual output label, we apply a mapping. For example, in the age estimation problem, the score would be the actual age, then we use an identity mapping, *i.e.* $y = f(\mathbf{x}, \theta)$; in gender recognition, the actual output labels are -1 for females and 1 for males, then we use $y = \text{sign}(f(\mathbf{x}, \theta))$; in race recognition, the output labels are multiple categories, the score function is a multivariate function, we use $y = \text{argmax}_i(f_i(\mathbf{x}, \theta))$.

In the supervised adaptation, we are given a set of additional training data $\mathcal{S}$, each element of which is $(\mathbf{x}, y)$, where $\mathbf{x}$ is the input and y is the supervised label. The adaptation can be formulated as an incremental learning problem, where the overall loss function is

$$J(\theta) = \frac{\lambda}{2}\|\theta - \theta_0\|^2 + \frac{1}{|\mathcal{S}|}\sum_{(\mathbf{x},y)\in\mathcal{S}} \ell(y, f(\mathbf{x}, \theta)), \quad (1)$$

where θ_0 is the parameter of the pre-trained recognition model, $\ell(y, f)$ is the individual loss function for each instance (image), which could be chosen according to the type of output labels. For example, a squared loss ($\|y - f\|^2/2$) for age estimation, a logistic loss for gender and race recognition. In the overall loss function, the first term keeps the parameter, θ, from deviating away from the original one, θ_0, with the regularization parameter λ. The second term reduces the loss on the additional training data.

Because of the characteristics of adaptation, we choose the stochastic gradient descent method to minimize $J(\theta)$. The update rule for an incoming sample $(\mathbf{x}_t, y_t)$ is

$$\theta \leftarrow \theta - \gamma_t \left[\lambda(\theta - \theta_0) + \frac{\partial}{\partial f}\ell(y_t, f(\mathbf{x}_t, \theta))\frac{\partial}{\partial \theta} f(\mathbf{x}_t, \theta)\right], \quad (2)$$

where γ_t is the step size and $\lambda(\theta - \theta_0)$ is the weight decay.

The stochastic gradient descent update rule is suitable for the CNN models in the paper. It also works well with other models, such as logistic regression and smoothed SVM.

3.2. Correspondence driven adaptation

Lack of supervised information, we aim to improve the recognition models by leveraging the correspondence information, *i.e.*, improving the consistency of the recognition scores of faces on the same trajectory. This is the core idea of the correspondence driven adaptation. For a given face $\mathbf{x} \in \mathcal{S}$ in the additional training image set, we have the set of all other faces on the same trajectory, denoted by the correspondence function $\mathcal{T}(\mathbf{x})$. Then, similar to Eq. (1), using the squared loss function, the overall loss function for the correspondence driven adaptation can be expressed as,

$$J(\theta) = \frac{\lambda}{2}\|\theta - \theta_0\|^2 + \frac{1}{4|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} \frac{1}{|\mathcal{T}(\mathbf{x})|} \times \sum_{\mathbf{z}\in\mathcal{T}(\mathbf{x})} (f(\mathbf{x}, \theta) - f(\mathbf{z}, \theta))^2. \quad (3)$$

The first term enforces a smoothness constraint on θ to the original parameter θ_0, where we use regularization parameter λ to control the deviation degree. The second term reduces the inconsistency of the profile recognition outputs of faces on the same trajectory. We normalize it by the size of $\mathcal{S}$ and the size of each trajectory, $|\mathcal{T}(\mathbf{X})|$.[1]

The derivative of the objective function of Eq. (3) can be written as,

$$\frac{\partial J}{\partial \theta} = \lambda(\theta - \theta_0) + \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} \frac{1}{|\mathcal{T}(\mathbf{x})|} \times \sum_{\mathbf{z}\in\mathcal{T}(\mathbf{x})} (f(\mathbf{x}, \theta) - f(\mathbf{z}, \theta))\frac{\partial}{\partial \theta} f(\mathbf{x}, \theta). \quad (4)$$

Note that each pair of images, $\mathbf{x}$ and $\mathbf{z}$, appears twice in the summation. Now we derive the stochastic update rule for each face image, $\mathbf{x}_t$,

$$\begin{aligned}\theta &\leftarrow \theta - \gamma_t[\lambda(\theta - \theta_0) + \frac{1}{|\mathcal{T}(\mathbf{x}_t)|} \times \sum_{\mathbf{z}\in\mathcal{T}(\mathbf{x}_t)} (f(\mathbf{x}_t, \theta) - f(\mathbf{z}, \theta))\frac{\partial}{\partial \theta} f(\mathbf{x}, \theta)] \quad (5)\\ &= \theta - \gamma_t\left[\lambda(\theta - \theta_0) + (f(\mathbf{x}_t, \theta) - \widetilde{y}_t)\frac{\partial}{\partial \theta} f(\mathbf{x}, \theta)\right],\end{aligned}$$

where $\widetilde{y}_t = \frac{1}{|\mathcal{T}(\mathbf{x}_t)|}\sum_{\mathbf{z}\in\mathcal{T}(\mathbf{x}_t)} f(\mathbf{z}, \theta)$. Note that Eq. (5) has the same form as Eq. (2), if we use squared loss in the supervised adaptation, $\frac{\partial}{\partial f}\ell(y_t, f) = f - y$. Thus, interestingly, this leads to that the average output of those images *other than* $\mathbf{x}_t$ on the trajectory is used as the pseudo labels, $\widetilde{y}_t$, for the supervised adaptation.

As the supervised adaption, the stochastic gradient descent update rule in Eq. (5) can be integrated into existing models without much effort. The updating process is implemented by the back propagation training for neural networks. Therefore, we can feed the pseudo labels into the existing training framework [17] to perform the correspondence driven adaptation. Because the validation set is not available, the stochastic training is only carried out in *one pass*, *i.e.*, each additional data $\mathbf{x}_t$ is used only once, which largely relieves the over-fitting risk. Note, since the face images are collected sequentially from videos, random data shuffle is critical in the stochastic training.

[1] The constants, 2 and 4 in the denominates, are put here to make the final update rule simpler. They can be absorbed by λ.

3.3. Relation to graph Laplacian based semi-supervised learning

Now we study the correspondence driven adaption from the view of semi-supervised learning. Unlike the online incremental fashion of our adaptation scheme, semi-supervised learning assumes that some image labels and correspondence information already exist during the offline training time.

Here, we use the normalized graph Laplacian regularization framework [29] for semi-supervised learning. The overall loss function of normalized graph Laplacian regularization has two parts, the loss part and the regularization part. The loss part is similar to the second term in Eq. (1). The regularization part is expressed as

$$\mathbf{f}^\top \mathbf{L} \mathbf{f},$$

where $\mathbf{f}$ is a vector, whose elements correspond to the scores of individual images; $\mathbf{L}$ is the normalized graph Laplacian matrix based on correspondence information, whose non-zero elements are

$$L_{ii} = 1; \quad L_{ij} = -\frac{1}{\sqrt{|\mathcal{T}(\mathbf{x}_i)||\mathcal{T}(\mathbf{x}_j)|}}, \text{ where } \mathbf{x}_j \in \mathcal{T}(\mathbf{x}_i).$$

As the vector $\mathbf{f}$ in the framework of [29] does not directly associate to the features of images, this framework is not inductive, *i.e.*, the scores can not be directly computed from the features of new images. Here, we constrain the i-th element of $\mathbf{f}$ to be $f(\mathbf{x}_i, \theta)$. Hence, once we have learned θ, we can compute the score for any image given its feature vector $\mathbf{x}$. Thus,

$$\mathbf{f}^\top \mathbf{L} \mathbf{f} = \frac{1}{2} \sum_i \frac{1}{|\mathcal{T}(\mathbf{x}_i)|} \sum_{j:\mathbf{x}_j \in \mathcal{T}(\mathbf{x}_i)} (f(\mathbf{x}_i, \theta) - f(\mathbf{x}_j, \theta))^2,$$

because $|\mathcal{T}(\mathbf{x}_i)| = |\mathcal{T}(\mathbf{x}_j)|$. This shows that the second term of Eq. (3) is exactly the regularization term of the normalized graph Laplacian, whose nodes are images, and edges are those pairs of images in the same trajectory.

Unlike the graph Laplacian based semi-supervised learning, our approach for correspondence driven adaptation does not require the unsupervised correspondence information to be known before training, which is the nature of adaptation, so that there is no gigantic Laplacian matrix to compute. Also, the adapted model can be applied to new images, which is not feasible in the above semi-supervised learning framework. We provide a stochastic training method by leveraging the existing training system (*e.g.* the CNN), which allows us to use the back propagation training algorithm to conduct the adaptation.

4. Human Profile Recognition System

We realize the correspondence driven adaptation in a fully automatic human profile recognition system in which convolutional neural networks [17] are learned for 3 important biological traits, *i.e.*, human gender, age, and race. They are chosen as representative recognition tasks of binary classification, regression, and multiclass classification. Next we briefly describe the architecture of CNNs and introduce the profile recognition system as follows.

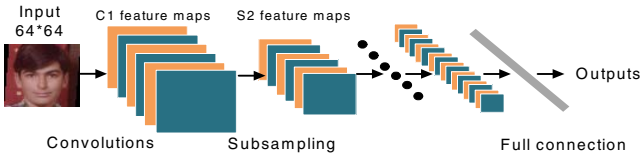

Figure 2. The architecture of the convolutional neural networks (CNNs) where each plane represents a feature map.

4.1. Convolutional Neural Networks

Conventional paradigm of pattern recognition usually consists of two steps: first compute hand-crafted features from raw inputs, then learn classifiers using the obtained features. The overall recognition performance is largely determined by the first step, which is, however, highly problem dependent and requires extensive feature engineering. Convolutional neural networks [17] are a class of deep learning approaches in which multiple stages of learned feature extractors are applied directly to the raw input images and the entire system can be trained end-to-end in a supervised manner. As shown in Fig. 2, in CNNs the convolution and subsampling operations are iteratively applied to the raw input images to generate multiple layers of feature maps. The adjustable parameters of a CNN model include all weights of the layers and their connections which are learned by the back-propagation algorithm [17].

We learn 5-layer CNN models for gender, female age, male age, and race, where the first two layers are shared across different tasks. The input layer includes 64×64 raw face image patches and the gradients of each RGB plane. CNN models are appealing since the design of hand-crafted features are avoided. In addition, they are efficient to evaluate at the testing stage, although in general the training of CNN models requires a large amount of data. In our C++ implementation, the evaluation of a CNN model runs at over 120 fps on a desktop with a Core 2 Duo 3.16GHz CPU.

4.2. System implementation

The correspondence driven adaptation can be readily incorporated in the human profile recognition system in Fig. 3. For every frame of the input video, we perform face detection and tracking, then the detected faces are aligned and normalized to 64×64 patches and fed to the CNN recognition engine to estimate the gender, age and race. The face detection and alignment modules are also based on certain CNN models. We employ multi-hypothesis tracking algorithms [13, 28] to obtain the correspondences of faces.

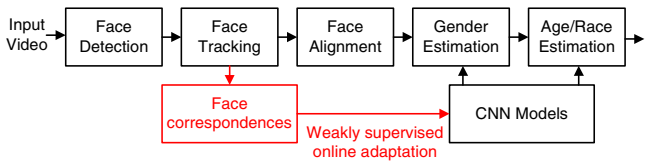

Figure 3. The diagram of the human profile recognition system.

(a) FG-Net dataset (b) FaceV dataset (c) Mall dataset

Figure 4. Example face image patches (64×64) output by the face alignment module on the 3 test datasets.

We tune the parameters to increase the precision of the tracker and tolerate the error of splitting frames of the same person to multiple tracks. After discarding too short tracks, the average length of trajectories is around 155 frames.

The recognition models trained using manually labeled faces are denoted by the *Baseline* CNN models. During processing a new video, the faces and their correspondences are stored as additional data which are used to update the Baseline CNN models online periodically. Then, the updated models are applied to new test videos. Note, we always update the *Baseline* models using the additional data, not the latest *updated* models, to avoid model drift.

5. Experiments

We apply the correspondence driven adaptation to gender, age and race recognition to validate the capability of the proposed method to improve binary classification, regression, and multiclass classification problems. For gender and age estimation, we first evaluate our Baseline CNN models on a widely-used public dataset called the FG-NET database [6] [2] to compare with the existing algorithms. Then we evaluate the proposed method on a dataset including 884 persons with 68759 faces extracted from 1.5 hours videos, referred as the *FaceV* dataset [3]. For race recognition, we test on a dataset with 908 persons and 19858 faces collected from video clips recorded in a shopping mall, thus, referred as the *Mall* dataset. Sample face images are shown in Fig. 4 where we can observe that the image quality and resolutions in the FG-NET are generally much higher than the faces extracted from videos.

The Baseline CNN models were trained using about 300K faces in which none of them is from the testing datasets, *i.e.*, the FG-NET, FaceV, and Mall datasets. The system shown in Fig. 3 is implemented in C++ and runs at 10 fps including face detection and tracking. The recognition engine alone runs at over 120 fps on a Core2Duo 3.16GHz desktop. A real-time live demo is available.

[2]The UIUC-Yamaha dataset was used in [26, 7, 11, 27, 12], however, it is not publicly available.

[3]The *FaceV* dataset is available upon signed license agreement.

Train	# persons	# img.	Test	# persons	# img.
Male	227	17788	Male	178	14181
Female	284	21185	Female	195	15605
Overall	511	38973	Overall	373	29786

Table 1. The split of training and testing data of the FaceV dataset in the *ExpI* setting.

5.1. Gender recognition

For the FG-NET database, using the Baseline CNN models as initialization, we adopt the Leave-One-Person-Out (LOPO) scheme, which is the common test scheme in the literature. The recognition accuracy of gender for females is 81.19% and 85.22% for males, the overall accuracy is 83.53%. The FG-NET database involves 37% of young kids (0-9 years old) whose genders are hard to recognize.

In the FaceV dataset, which was collected from videos resemble to surveillance videos in retail stores, there are 884 Asian persons with 68759 faces. The faces are extracted from nine 10-minute video clips denoted by $\{V_1, \cdots, V_9\}$. The number of persons in one video clip varies from 80 to 150. Each person appears only in one video clip. The labels are obtained using Amazon Mechanical Turk with 3 workers for the gender and 7 workers for the ages. These are the appearance gender and ages of interests in digital signage and customer analysis applications.

We conduct experiments on the FaceV dataset with two settings, denoted by *ExpI* and *ExpR*. ExpI employs a conventional setting of *i*ncremental adaptation. That is, on top of the Baseline models, we use 6 video clips V_1 to V_6 as the additional training set and the other 3 clips V_7 to V_9 as the testing set. ExpR uses a *r*ecursive adaptation setting that resembles data availability in real applications. We use one 10-minute clip as the additional training data and another one as the testing data, *e.g.*, using faces in V_1 for adaptation and V_2 for testing, next, using V_2 for adaptation and V_3 for testing, so on so forth (using V_9 to update the models to test on V_1). Afterwards, the test results are averaged. Note, this *ExpR* setting is different from a 9-fold validation in that much less data are used in each round of model adaptation. It is also worth noting that each time the Baseline models are updated without any accumulation to avoid drifting. The statistics of the FaceV dataset and the data split in *ExpI* are shown in Table 1.

We compare the proposed correspondence driven method, denoted by *CD-CNN*, against the Baseline CNN

ExpI	Male	Female	Overall
Baseline	83.60%	87.83%	85.82%
Self train	79.69%	90.15%	85.17%
CD-CNN	89.82%	86.93%	***88.31%***
ExpR	**Male**	**Female**	**Overall**
Baseline	85.28%	90.55%	88.10%
Self train	86.02%	90.39%	88.36%
CD-CNN	91.84%	87.84%	***89.70%***

Table 2. The accuracy of gender recognition in the ExpI (top) and the ExpR (bottom) settings.

	Male		Female		Overall	
Range	MAE	# img.	MAE	# img.	MAE	# img.
0- 9	2.56	240	4.44	131	3.22	371
10-19	3.12	187	4.36	152	3.67	339
20-29	4.30	83	5.49	61	4.80	144
30-39	8.97	33	9.21	46	9.11	79
40-49	11.30	27	16.49	19	13.45	46
50-59	8.70	7	24.64	8	17.20	15
60-69	13.91	5	29.88	3	19.90	8
Total	***3.93***	582	***6.20***	420	***4.88***	1002

Table 3. MAE (years) at different age groups on the FG-Net database.

model and the self-training approach in which the predictions of the offline pre-trained model are directly used as the labels to update the model. The accuracy of gender estimation is shown in Table 2 where we observe that the CD-CNN improves the Baseline models by 2.49% in the ExpI and 1.6% in the ExpR. While, the self-training approach shows very little or no improvement. The within-class variances of the appearances of female faces collected online may be even larger than those induced by deployment environments.

5.2. Age recognition

The performance of age estimation is measured by the mean absolute error (MAE) and the cumulative score (CS) [8] which is defined as $CS(l) = N_{e<l}/N$ where $N_{e<l}$ is the number of test images on which the absolute error of the age estimation is no higher than l (years) and N is the total number of test images.

For the FG-NET database, following the same evaluation criteria in [26, 27, 12], we show the breakdown of MAEs at different age groups in Table 3 and the cumulative scores up to 15 years in Fig. 5. The comparison with other methods is shown in Table 4. Our CNN method achieves MAE 4.88 which is very close to the state-of-the-art MAE 4.77 in [12], where the biological inspired "HMAX" model [22] bears some resemblance to the neural networks. However, the derivatives of some operations in HMAX units, *e.g.*, the max-pooling, are hard to calculate for adaptation.

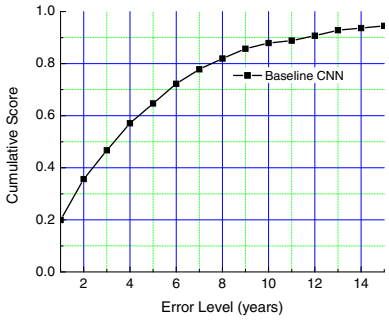

Figure 5. The cumulative scores of age estimation on the FG-NET.

Method	QM [16]	AGES [8]	RUN [26]	LAPR [11]
MAE	6.55	6.77	5.78	5.07
Method	**RPK [27]**	**BIF [12]**	**mkNN [24]**	**Ours**
MAE	4.95	4.77	4.93	*4.88*

Table 4. Comparison of the MAE (years) on the FG-NET database.

Method	Male	Female	Overall
Baseline	5.29	7.32	6.35
Self train	5.14	7.32	6.28
CD-CNN	*4.47*	*6.49*	***5.53***
Method	**Male**	**Female**	**Overall**
Baseline	5.56	7.34	6.51
Self train	5.53	7.33	6.49
CD-CNN	*5.24*	*6.86*	***6.11***

Table 5. The comparison of MAE (years) in the ExpI (top) and ExpR. (bottom)

For the FaceV dataset, we can observe from Table 5 that the CD-CNN achieves MAE=5.53 which is 0.82 years lower than the Baseline in the ExpI, and the MAE is reduced from 6.51 to 6.11 by 0.4 years in the ExpR. These substantial improvements clearly validate the advantage of the correspondence driven approach. We also test the performance of the supervised adaptation using true labels, where the MAE is 5.37 and 6.04 in the ExpI and ExpR respectively. These are the optimal results or the bounds we can achieve given the additional data, which reveals that the results of CD-CNN are very close. The cumulative scores in Fig. 6 show details of how close CD-CNN is to the supervised learning results. A higher performance gain (0.82 years) over the Baseline is obtained in the ExpI than that in the ExpR (0.40 years), which shows the benefit of more data for adaptation. The MAEs at different age groups are shown in Table 6 and 7. These tables indicate females' ages are harder to estimate than males' ages. The MAEs of kids and senior people are high probably due to insufficient training samples.

The strength of the correspondence driven adaptation originates from the capability to address the mismatch problem due to the factors such as lighting, view angels, image

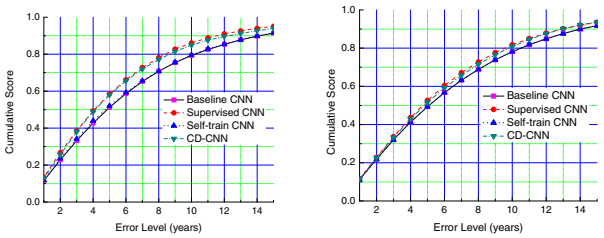

Figure 6. The comparison of the cumulative scores in the ExpI (left) and the ExpR (right). The CS curve of the Baseline CNN is largely occluded by that of the self-training CNN. The performance of the CD-CNN is slightly different from that of the CNN models obtained by supervised model adaptation.

	Male		Female		Overall	
Range	MAE	#img.	MAE	#img.	MAE	#img.
0- 9	5.57	1781	10.06	1361	7.52	3142
10-19	3.29	2525	8.67	2012	5.68	4537
20-29	4.60	3549	4.91	7338	4.81	10887
30-39	3.85	2051	6.54	2633	5.36	4684
40-49	4.59	2232	7.66	1238	5.68	3470
50-59	5.09	1594	7.17	937	5.86	2531
60-69	4.75	355	8.17	86	5.42	441
70-79	9.81	94	0.00	0	9.81	94
Total	*4.47*	14181	*6.49*	15605	***5.53***	29786

Table 6. The MAE (years) at different age groups of the CD-CNN in the ExpI.

	Male		Female		Overall	
Range	MAE	#img.	MAE	#img.	MAE	#img.
0- 9	6.01	2882	10.62	2959	8.34	5841
10-19	3.49	3182	9.01	4020	6.57	7202
20-29	4.85	9002	5.79	14132	5.42	23134
30-39	4.79	5685	5.95	7320	5.44	13005
40-49	4.99	5364	7.22	5337	6.10	10701
50-59	6.76	3669	6.07	2548	6.48	6217
60-69	5.62	1653	11.07	474	6.84	2127
70-79	14.06	532	0.00	0	14.06	532
Total	*5.24*	31969	*6.86*	36790	***6.11***	68759

Table 7. The MAE (years) at different age groups of the CD-CNN in the ExpR.

quality, or noise levels in the deployment environments, which may seldom be exactly the same as those in the training set. By forcing the models to produce consistent outputs for the same person, the models are adapted to be less sensitive to these factors. Therefore, the updated models outperform the original ones noticeably even if a small amount of additional data is added in the adaptation (*e.g.*, around 6K to 10K faces are added in the ExpR for each video clip).

Race	# persons	#img.	Baseline	Self train	CD-CNN
Caucasian	500	11086	88.95%	92.34%	93.38%
EastAsian	76	1487	54.82%	47.15%	55.52%
African	173	4068	78.86%	81.22%	85.96%
Hispanic	112	2231	55.49%	48.63%	53.88%
Indian	47	986	50.62%	44.60%	42.49%
Overall	908	19858	78.53%	79.32%	***81.65%***

Table 8. Race recognition performance on the Mall dataset.

5.3. Race recognition

For race recognition, we consider 5 races, *i.e.*, *Caucasian*, *EastAsian*, *AfricanAmerican*, *Hispanic*, and *Indian*. To our best knowledge, there is no public video dataset with race labels. So we collected several thousands of short clips of 908 people with 19858 tracked faces in a shopping mall, where the race labels were given by the subjects. We refer this collection as the Mall dataset. As shown in Fig. 4(c), this is a very challenging task due to the inherent ambiguity among different races even for human. Following the testing scheme of ExpI, we employ the correspondence driven adaptation to obtain the recognition accuracy 81.65% which improves by 3.12% against the Baseline accuracy 78.53% with no human intervention, as presented in Table 8.

5.4. Live demo system

The correspondence driven adaptation has been integrated in a live human profile recognition demo system. Example video sequences showing the gender and age recognition results in the FaceV dataset (all persons are Asians) are included in the supplemental materials, which also illustrate the face detection and tracking performance. Some example screen shots are shown in Fig. 7 [4], where the bounding boxes of faces are drawn in different colors to indicate the person id. Gender and age ranges centered at our estimation are also shown. For instance, "7:F10-15" means that the person #7 is a female with the age ranging from 10 to 15 years old.

6. Conclusions

Motivated by the mismatch issue of visual recognition models after deployment, we propose a novel correspondence driven adaptation strategy and apply it to an advanced CNN-based human profile recognition system. After collecting face correspondences by human tracking, we derive an online stochastic training method to enforce the updated models to output consistent gender, age and race estimations for the same person. The proposed algorithm enables a pre-trained CNN model adapt to the deployment environment and achieves significant performance improvement

[4]The privacy of the persons must be protected to the maximum extent.

Figure 7. Sample screen shots of the real-time human profile recognition system.

on two large video datasets containing about 900 persons. We demonstrate the correspondence driven approach can be readily integrated to a live visual recognition system.

Our future work includes investigation of online semi-supervised learning to impose the correspondence constraint and its applications to other video analysis problems.

References

[1] NEC digital signage solution. http://www.nec.com/global/solutions/digitalsignage/.

[2] S. Avidan. Ensemble tracking. In *CVPR'05*, volume 2, pages 494–501, San Diego, CA, June 20 - 25, 2005.

[3] O. Chapelle, B. Scholkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, Sept. 2006.

[4] N. Cherniavsky, I. Laptev, J. Sivic, and A. Zisserman. Semi-supervised learning of facial attributes in video. In *ECCV'10*, Crete, Greece, Sept.5-11, 2010.

[5] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active apperance models. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(8):681–685, 2001.

[6] FG-NET Aging Database. http://www.fgnet.rsunit.com.

[7] Y. Fu and T. S. Huang. Human age estimation with regression on discriminative aging manifold. *IEEE Trans. Multimedia*, 10(4):578–584, 2008.

[8] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Trans. Pattern Anal. Machine Intell.*, 29(12):2234–2240, 2007.

[9] B. A. Golomb, D. T. Lawrence, and T. J.Seinowski. SEXNET: A neural network identifies sex from human faces. In *NIPS'99*, pages 572–577, Denver, CO, Nov.29 - Dec.4, 1999.

[10] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. In *Proc. British Machine Vision Conference (BMVC'06)*, volume 1, pages 47–56, Edinburgh, UK, Sept. 4-7, 2006.

[11] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Trans. Image Processing*, 17(7):1178–1188, 2008.

[12] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *CVPR'09*, Miami, FL, June 21 - 23, 2009.

[13] M. Han, W. Xu, H. Tao, and Y. Gong. An algorithm for multiple object trajectory tracking. In *CVPR'04*, volume 1, pages 864 – 871, Washington, DC, Jun.27-Jul.2, 2004.

[14] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV'09*, Kyoto, Japan, Sept.29 - Oct.2, 2009.

[15] Y. H. Kwon and N. Lobo. Age classification from facial images. *Computer Vision and Image Understanding*, 74(1):1–21, Apr. 1999.

[16] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Trans. Syst., Man, Cybern. B*, 34(1):621–628, 2004.

[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based leraning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, Nov. 1998.

[18] K.-C. Lee and D. Kriegman. Online learning of probabilistic appearance manifolds for video-based recognition and tracking. In *CVPR'05*, volume 1, pages 852 – 859, San Diego, CA, June 20 - 25 2005.

[19] J. Lim, D. Ross, R.-S. Lin, and M.-H. Yang. Incremental learning for visual tracking. In *NIPS'04*, pages 801–808, Vancouver, Canada, Dec. 13-18, 2004.

[20] B. Moghaddam and M.-H. Yang. Gender classification with support vector machines. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 306–312, 2000.

[21] N. Ramanathan and R. Chellappa. Face verification across age progression. *IEEE Trans. Image Processing*, 15(11):3349–3361, 2006.

[22] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.

[23] D. Ross, J. Lim, and M.-H. Yang. Adaptive probabilistic visual tracking with incremental subspace update. In *ECCV'04*, volume 1, pages 215–227, Prague, Czech Republic, May 11-14, 2004.

[24] B. Xiao, X. Yang, Y. Xu, and H. Zha. Learning distance metric for gression by semidefinite programming with application to human age estimation. In *ACM Int'l Conf. on Multimedia (ACM MM'09)*, Beijing, Oct. 19 - 23, 2009.

[25] R. Yan, J. Zhang, J. Yang, and A. Hauptmann. A discriminative learning frameworkwith pairwise constraints for video object classication. In *CVPR'04*, volume 2, pages 284 – 291, Washington, DC, Jun.27-Jul.2, 2004.

[26] S. Yan, H. Wang, X. Tang, and T. S. Huang. Learning auto-structured regressor from uncertain nonnegative labels. In *ICCV'07*, Rio de Janeiro, Brazil, Oct.14-20, 2007.

[27] S. Yan, X. Zhou, M. Hasegawa-Johnson, and T. S. Huang. Regression from patch-kernel. In *CVPR'08*, 2008.

[28] M. Yang, F. Lv, W. Xu, and Y. Gong. Detection driven adaptive multi-cue integration for multiple human tracking. In *ICCV'09*, Kyoto, Japan, Sept.29 - Oct.2, 2009.

[29] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *NIPS'04*, pages 321–328, Vancouver, Canada, Dec. 13-18, 2004.