博士論文

カメラ画像を用いたディープラーニングによる被写人物の年齢推定

AGE ESTIMATION OF PEOPLE USING DEEP LEARNING WITH CAMARA IMAGES

２０２２年4月２２日

April 22, 2022

東京都市大学　総合理工学研究科　情報専攻　視覚メディア研究室

Visual Media Laboratory, Department of informatics, Graduate School of Integrative Science and Engineering, Tokyo City University

張　北辰

Beichen Zhang

概　要

顔画像からの年齢推定は，コンピュータビジョンにおいて重要かつ挑戦的な課題であった。これまでに提案された年齢推定法では，人種や性別が違う場合にパフォーマンスが低下する，画像を撮る時のポーズが異なる場合にもパフォーマンスが低下するなど問題がある。

それらの問題を解決するため，本論文は Cross-dataset 学習法，頭部の向き推定を併用した学習法，Face Landmarks を用いたマルチタスク学習法の 3 つ手法を提案し，Morph，CACD，AFAD，UTK Face など代表的な顔画像のデータセットを用いて年齢推定の検証実験を行った。

既存の年齢推定用データセットは，含まれた画像数の不足による，画質の問題及び単一人種の問題があった。高品質な学習データ不足の課題解決を目的として，複数のデータセットを併用した Cross-dataset 学習法を提案した。提案手法は数が少なく質も低い CACD と AFAD データセットと数が多く質も高い Morph データセットを併用して，年齢推定の精度は従来技術より CACD で 0.7 歳（平均推定年齢），AFAD で 0.2 歳を向上できた。

写真や動画の中での年齢推定の精度が低下する問題について，人間のポーズが異なることによる精度の低下への影響を実験で明らかにした。ポーズによる精度低下の課題解決を目的として，頭部の向き推定を併用した学習法を提案した。提案手法は画像に対して，まず頭部の向き推定を行って，推定した 3 つの角度により，30 度以内の画像のみ年齢推定を行った。頭部の向き角度の制限を用いて，近年よく利用されるデータベース CACD と AFAD による年齢推定の精度を従来技術より 0.8 歳を向上できた。

近年よく利用されているマルチタスク学習法に着目し，年齢推定の精度向上のために報告されている，性別推定を用いたマルチタスク学習法による年齢推定法があった。この方法の課題であるさらなる精度向上を目的として，Face Landmarks を用いたマルチタスク学習法を提案した。提案手法は Face Landmarks を用いたマルチタスク学習法により，よく利用されるデータベース CACD と UTK Face で年齢推定の精度は従来技術より 0.5 歳を向上できた。

顔画像からの年齢推定における課題を分析するとともに三つの問題点を着目し，三つの解決案を提案した。さらに提案したアルゴリズムによって実験を行ってそれらの有効性を確認し、顔画像からの年齢推定の精度向上を果たした。

Abstract

Age estimation from facial images has been an important yet challenging task in computer vision. The age estimation methods proposed so far have problems such as poor performance because of the differences of race and gender as well as different poses in images.

In order to solve these problems, this paper proposes three methods: Cross-dataset learning method, learning method using head pose estimation, and multi-task learning method using Face landmarks. Some representative facial image databases such as Morph, CACD, AFAD and UTKFace were used to conduct a verification experiment of age estimation.

The existing age estimation dataset had image quality problems and single race problems due to the lack of images included. We proposed a Cross-dataset learning method that uses multiple datasets together for the purpose of solving the problem of lack of high-quality learning data. The proposed method uses a combination of CACD and AFAD datasets with small number and low-quality images and Morph dataset with large number and high-quality images, resulting in the accuracy improvement of 0.7 years for CACD (average estimated age) and 0.2 years for AFAD compared to the state-of-the-art method.

It was clarified by experiments that different human poses in images and videos cause decrease on accuracy of age estimation. For the purpose of solving the problem of accuracy deterioration due to poses, we proposed a learning method using head pose estimation. In the proposed method, the orientation of the head was first estimated from the image, and then the age was estimated only for the head pose within 30 degrees from the estimated three angles. By limiting the orientation angle of the head, we were able to improve the accuracy of age estimation on the databases CACD and AFAD, which are often used in recent years, by 0.8 years compared to the state-of-the-art method.

Focusing on the multi-task learning method that is often used in recent years, there was an age estimation method by the multi-task learning method using gender estimation, which has been reported to improve the accuracy of age estimation. We proposed a multi-task learning method using Face landmarks for the purpose of further improving the accuracy, which is the problem of this method. The proposed method improved the accuracy of age estimation by 0.5 years compared to the state-of-the-art method on CACD and UTKFace dataset.

We analyzed the problems in age estimation from facial images, focus on three problems, and proposed three methods to solve then. Experiments were conducted using the proposed algorithms and the effectiveness are confirmed with improvement of the accuracy for age estimation from facial images.

目次

第 1 章 序論

1.1 研究背景

人間の顔画像には，性別，年齢，人種，表情，健康状態など，多くの特徴が含まれている。これらの特徴のうち，年齢推定は，ヒューマンコンピュータインタラクション [1,2,3]，セキュリティ [4]， 医療[5]，カスタム広告[6]など様々な分野で利用されており，非常に挑戦的かつ重要な課題となっている。

図 1.1 のように、年齢を一つの情報として、顔認識システムに利用されている。図 1.2 を示したカスタム広告に対して、年齢情報が抽出できれば、特定年齢層の客に相応しい商品を推薦できる。将来的には、AI 診断（図 1.3）を大幅に応用されると想定できるが、年齢情報を使って、薬の用量や施術可否の判断も可能である。

カスタム広告などの応用に対して正確な年齢数字ではなく、年齢層をある程度判断できれば十分だが、セキュリティなどの場合は推定年齢に対して高い精度を求められている。例えば、お酒を売る場所で、AI により年齢抽出すれば、非常に高い精度（誤差 1 歳以内）が必要である。より広い領域で応用できるため、年齢推定の精度をより向上することを目指す。

顔画像を用いた年齢推定問題についての研究は，1994 年までさかのぼることができる [7]。当初は，皮膚のしわや顔の一部情報のみを用いて，年齢を一定範囲内に推定していた。2009 年に生物学的な特徴量（BIFs）[8] という広く利用されている表現方法が提案されて，さらに局所バイナリパターン（LBP）[9] やスケール不変特徴変換（SIFT）[10]，bor [11] などの方法は従来法よりもはるかに高い精度を達成することができた。近年，畳み込みニューラルネットワーク（CNN）は，物体検出 [12]，セグメンテーション [13]，顔認識 [14] など，コンピュータビジョンの様々な分野で素晴らしい性能を示している。また，CNN は ChaLearn looking at people challenge [15]で，年齢推定問題に対する主流のアプローチであり，目覚しい進歩を遂げていた。

図 1.1 監視カメラよる年齢・表情・性別を瞬時に推定顔認識システム

図 1.2　google カスタム広告

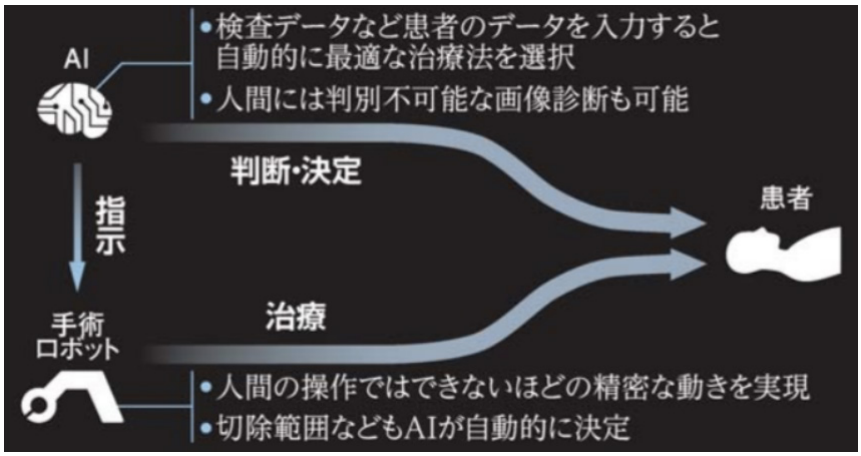

図 1.3　AI 診断

年齢推定問題に関する広範な研究にもかかわらず，既存の手法の性能は，精度や信頼性の点で実生活の要求を満たすにはまだ程遠い。この問題を難しくしている要因は，照明，ポーズ，表情などの外的要因（図 1.4)[16] と，人種，性別，健康状態などの内的要因(図 1.5) [8] の 2 つに分類されている。外的要因は画像処理や写真撮るときの調整などで簡単に補正できるが内的要因の解決は困難であるため，多くの先行研究が内的要因に着目している。年齢推定の困難さ，1）同じ年齢でも顔の見た目が大きく異なること（図 1.6），2）人間の顔は年齢によって変化すること（例えば，子供の頃に骨の成長が早いが，大人になってから，顔の変化は少ない）（図 1.7）により、様々な年齢層の顔画像を正確に推定できる年齢推定器を設計することは困難な課題である。

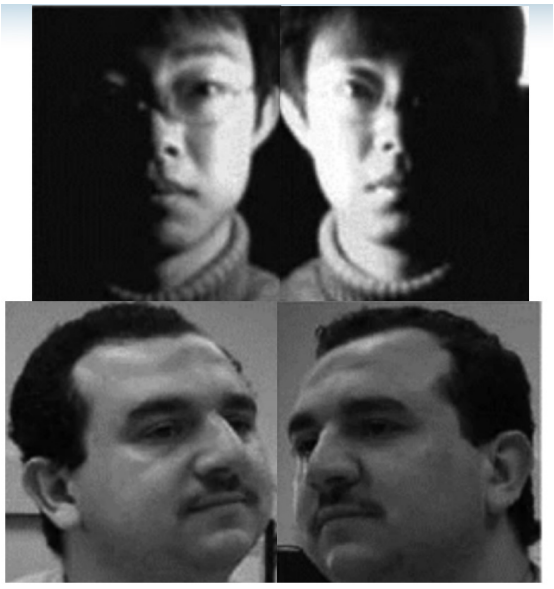

図 1.4　照明、ポーズにより見た目の違い

図 1.5　人種、性別により見た目の違い

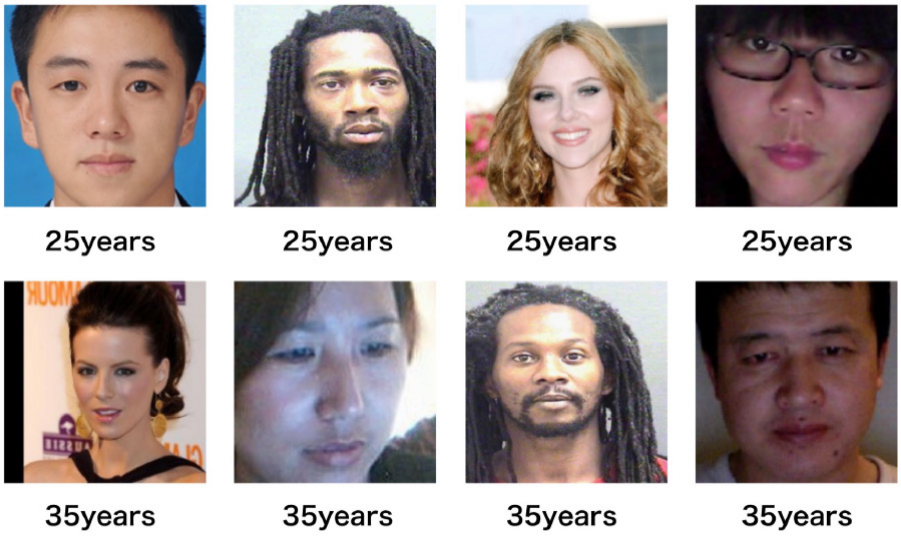

図 1.6 同年代の異なる人たちの外見の違い

図 1.7 小児期から成人期にかけての顔貌の変化

同じ年齢推定方法では，人種や性別が変わるとパフォーマンスが低下する。画像を撮る時のポーズが異なる場合にもパフォーマンスが低下するなど様々な問題がある。それらの問題を解決するため，本論文は Cross-dataset 学習法，頭部の向き推定を併用した学習法，Face Landmarks を用いたマルチタスク学習法の 3 つ手法を提案し，Morph，CACD，AFAD，UTK Face など代表的な顔画像のデータセットを用いて年齢推定の検証実験を行った。

1.2 本論文の構成及び概要

本論文の内容を章ごとに概説する。

第 1 章では，画像より年齢推定の研究背景と研究の目的，および本論文の全体構成について説明した。

第 2 章では，顔画像からの年齢推定に関して従来研究を調査・分析し，関連ディープラーニングの特性について解説した。そして，近年の技術についてまとめ，現存の問題点，改善すべきところなど課題を抽出した。

第 3 章では，既存の年齢推定用データセットの概要を説明し，含まれた画像数の不足による問題，画質の問題及び単一人種による問題を明らかにした。そして，高品質な学習データ不足の課題解決を目的として，複数のデータセットを併用した Cross-dataset 学習法を提案した。Cross-dataset 学習法のアルゴリズム，ネットワーク構造を説明した。実験については，数が少なく質も低い CACD と AFAD データセットと数が多く質も高い Morph データセットを用いて提案方法による年齢推定実験を行った。その結果，提案方法による年齢推定の精度は従来技術より CACD で 0.7 歳（平均推定年齢），AFAD で 0.2 歳を向上できたことを示した。

第 4 章では，写真や動画の中での年齢推定の精度が低下する問題についてさらに分析し，

人間のポーズが異なることによる精度の低下への影響を実験で明らかにした。そして，ポーズによる精度低下の課題解決を目的として，頭部の向き推定を併用した学習法を提案した。提案した年齢推定と頭部の向き推定法について，詳細なアルゴリズムやネットワーク構造を説明した。画像に対して，まず頭部の向き推定を行なって，推定した 3 つの角度により，30 度以内の画像のみ年齢推定を行なった。実験を行った結果，頭部の向き角度の制限を用いて，近年よく利用されるデータベース CACD と AFAD による年齢推定の精度を従来技術より 0.8 歳を向上できたことを示した。

第 5 章では，近年よく利用されているマルチタスク学習法に着目し，年齢推定の精度向上のために報告されている，性別推定を用いたマルチタスク学習法による年齢推定法について説明した。この方法の課題であるさらなる精度向上を目的として，Face Landmarks を用いたマルチタスク学習法を提案した。Face Landmarks の概要とマルチタスク学習について，アルゴリズムやネットワーク構造を説明し，年齢推定実験を設計・実行した。その結果，Face Landmarks を用いたマルチタスク学習法により，よく利用されるデータベース CACD と UTK Face で年齢推定の精度は従来技術より 0.5 歳を向上できたことを示した。

第 6 章では，以上の各章で得られた結論を総括し、将来的な展望について述べした。
本研究成果は，顔画像からの年齢推定における課題を分析するとともに従来技術の問題を明らかにし，複数の精度向上案を提案した。さらに実装のためのアルゴリズムやディープラーニングモデルを開発し，実験でそれらの有効性を確認した。今後はリアルタイム処理など改良を加え，実用化を目指したい。

1.3 本研究に関連する学術論文・研究発表

A. 学術論文

1. Beichen Zhang and Yue Bao. “Cross-Dataset Learning for Age Estimation.” *IEEE Access*, vol 10, pp 24048-24055, 2022

2. Beichen Zhang and Yue Bao. “Age Estimation of Faces in Videos Using Head Pose Estimation and Convolutional Neural Networks.” *Sensors*, 22(11), 4171, 2022

3. Beichen Zhang and Yue Bao. “Facial Age Estimation Using Deep Multitask Learning with Face Landmarks.” *IEEE Access*, 2022（Responding to peer-reviewed comments）

B. 国際学会における研究発表

Beichen Zhang and Yue Bao. “A Deep CNN Model for Age Estimation.” *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision,* 2022.5.18

参考文献

1. A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. IEEE Transaction on Systems, Man, and Cybernetics, Part B: Cybernetics, 34(1):621–628, 2004.
2. H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In Proc. ICB, pages 1–8, 2013.
3. X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. In Proceedings of the ACM International Conference on Multimedia, pages 307–316, 2006.
4. A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(4):442–455, 2002.
5. Z. Song, B. Ni, D. Guo, T. Sim, and S.Yan. Learning universal multi-view age estimator using video context. In Proceedings of the IEEE International Conference on Computer Vision, pages 241–248, Nov 2011. 1
6. C. Shan, F. Porikli, T. Xiang, and S. Gong, editors. Video Analytics for Business Intelligence. Studies in Computational Intelligence. Springer, 2012.
7. Y. Kwonand, and N. Lobo. Age classification from facial images. In IEEE CVPR, pages 762–767, 1994.
8. G. Guo, G. Mu, Y. Fu, and T. Huang. Human age estimation using bioinspired features. In IEEE CVPR, pages 112–119, 2009.
9. T. Ojala, M. Pietikinen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In ICPR, pages 582–585, 1994.
10. D. Lowe. Object recognition from local scale-invariant features. In IEEE ICCV, pages 1150–1157, 1999.
11. D. Gabor. Theory of communication. J. Inst. Electr. Eng., 93(26):429–457, Nov. 1946.
12. T. -Y. Lin, P. Dolla r, R. B. Girshick, K. He , B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In IEEE CVPR, volume 1, page 4, 2017.
13. K. He, G. Gkioxari, P. D. ar, and R. Girshick. Mask R-CNN. In IEEE ICCV, pages 2980–2988, 2017.

14. Zhang, K., Zhang, Z., Li, Z., and Qiao, Y.. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503. 2016.
15. S. Escalera, J. Fabian, P. Pardo, X. Baro, J. Gonzalez, H. J. Escalante, and I. Guyon. ChaLearn 2015 apparent age and cultural event recognition: Datasets and results. IEEE International Conference on Computer Vision, ChaLearn Looking at People workshop, pages 1–9, 2015.
16. H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic estimation from face images: Human vs. machine performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(6):1148–1161, 2015. 1

第 2 章 年齢推定に関する既往研究

本章では年齢推定について 3 つの観点から既往の研究についても概観する。2.1 節ではデータセットの人種単一性に注目して，その単一性を解決するため計画，2.2 節では写真や動画の中での年齢推定の精度が低下する問題を注目し，その不安定性を解決するため計画，2.3 節ではマルチタスク学習法に着目し，その精度を上がるに基づく計画について示す。

2.1 データセットに注目した研究

既存の年齢ラベル付き顔画像データセットのほぼ全ては，単一人種の顔画像を含んでいる。いくつかの先行研究は，各人種の年齢を別々に推定することを提案している[1]，[2]. しかし，個別モデルは，ほとんどのオープンデータセットが各人種のデータが不十分であり，一部のデータが誤ったラベルを持つという新たな問題に直面する。モデルに特化したデータセットを作成するという選択肢のコストが高いし，データを手作業で収集して，十分なラベルを付けることも困難である。この問題に対処するために，より多くのデータをラベリングする代わりに，ある人種の大規模な高品質データセットを使用して，別の人種の顔の画像を含む小規模または低品質のデータセットにてファインチューニングして、年齢推定の性能を向上させることができる[3]。また，この方法は，学習したモデルの汎化性能を向上させ，実際のアプリケーションにおいてより信頼性の高いモデルを実現することができる。このクロスデータセット年齢推定アプローチ（図 2.1）の問題点は，対象データセットが少数の高品質な学習データしか持っていないため，ファインチューニングしても、学習過程が不安定になり，また精度の低い学習結果になってしまうことである 。

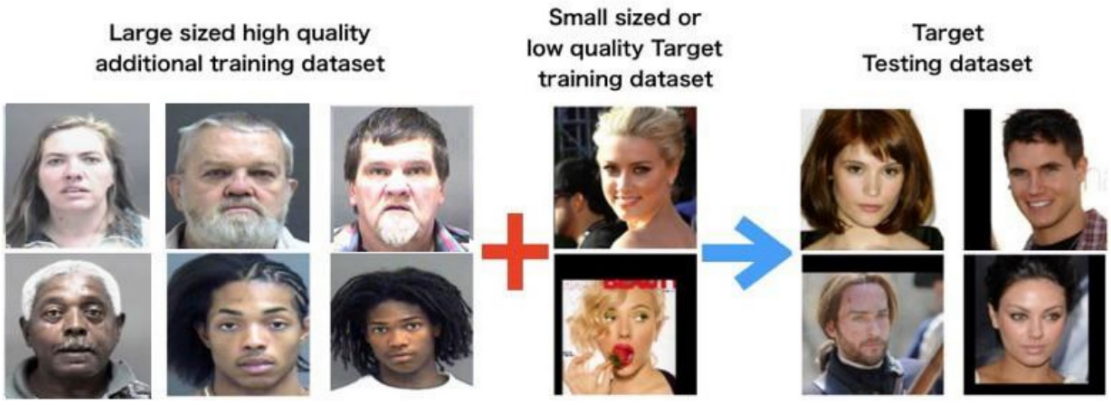

図 2.1 クロスデータセット年齢推定アプローチ

クロスデータセット学習を用いた先行研究は，顔画像を用いた年齢推定に適用されていない。その理由は，クロスデータセットは通常，異なるクラスのラベルを持つ複数のデータセットを用いるため，ラベルが異なるデータセットから複製される可能性があるからだ。

統合型顔分析ネットワーク[4]では，異なるデータセットを持つ様々なタスクを一緒に学習する。顔のランドマークや顔の感情などの複数の特徴を統合して汎用的な顔モデルを作成するため，データサンプルごとにすべての特徴にラベルを付ける必要はない。本手法は，特徴量間の相互作用をマルチタスク学習の範囲の要素としてモデル化することで，性能を大幅に向上させる。

もう一つ関連するアプローチは，Kuang ら[5]が ChaLearn Looking at People（LAP）challenge 2015 で提案したものである。彼らは，制約のない顔画像から年齢を推定するために，公開されているラベル付き顔年齢データセットを活用する方法を提案し，深い CNN を使用して複数の公開されている年齢データセットで年齢関連の識別表現を学習した。CNN の学習は，リッチバイナリコードによって監督され，したがってマルチラベルクラス分類問題としてモデル化された。このコードは，複数の粒度における異なる年齢グループのパーティションと，性別情報を表す。そして，LAP 主催者から提供された小規模な訓練データセットに対して，ランダムフォレストと二次回帰法を融合し，局所調整法を用いて，深層表現から年齢への回帰器を訓練した。[6]に示されるように，回帰問題に対する非定常カーネルの学習は，学習過程においてオーバーフィッティングを引き起こしやすいため，トレーニングが困難である。

2.2 異なるポーズによる不安定性に注目した研究

近年，年齢推定[7,8,9]をはじめとする様々なコンピュータビジョンタスクにおいて，深層学習が重要な成果を上げている．しかし，これらの研究はすべて正面顔画像のみを含むデータセットを用いており，実際のアプリケーションの状況を十分に反映することができない．また，動画像やウェブカメラでは，顔画像とは異なり，頭部の姿勢が大きく変化するため，年齢推定に耐えられない誤差が生じる可能性がある．

映像中の顔から年齢を推定する場合，最も密接に関連する研究として，Deep Age（図 2.2）推定モデル[10]では，Ji らが注目機構を持つ CNN を使用し，顔の特徴量を抽出し，注目ブロックから顔の特徴量ベクトルを集計し最終的な年齢推定を行う。モデル全体の学習はフレームごとの精度と安定性の両方を保証する新しい損失関数を利用して、年齢推定結果を全フレームに渡って表示する。ただし，複数のフレームを同時にトレーニングすると、コストが高くなる。また，複数のフレームを利用しても長時間のポーズ変化に対応できない場合もある。動画から年齢の表情、両方を同時に学習することができる研究は，Spatially-Indexed Attention Model（SIAM）（図 2.3） [11]と呼ばれている。このモデルでは，Ji Pei らは CNN を使用して潜在的な外観表現特徴を抽出し，それをリカレントネットワークに送り込んで

テンポラリダイナミクスをモデル化している。さらに，畳み込み層の間で空間的にインデックス化された特殊メカニズムを定義し，個々の画像で顕著な顔領域から抽出し，時間的な注意層を使用して各フレームに重みを割り当てる。この 2 つのアプローチは，情報フレームと顔領域に焦点を当てることでパフォーマンスを向上させるだけでなく，空間的な顔領域と時間フレームと年齢推定のタスクとの相関性も表れている。しかし，この方法は顔の表情もトレーニングするので、利用できるデータセットは制限されている，具体的には笑顔と嫌悪感のラベルが付いているデータベースのみが使用された。

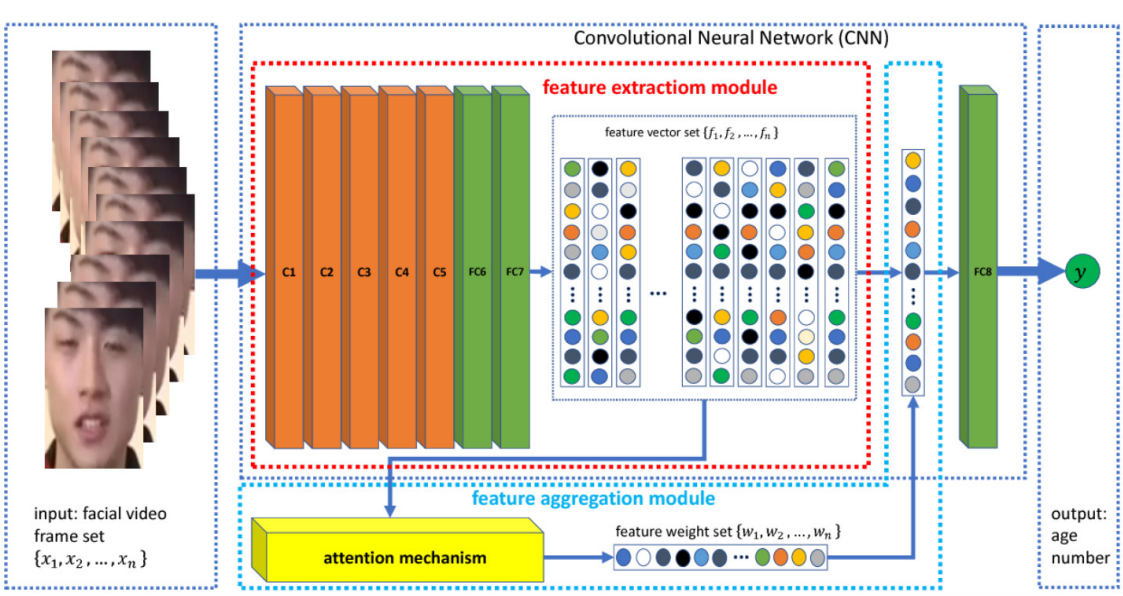

図 2.2 動画による年齢推定システム

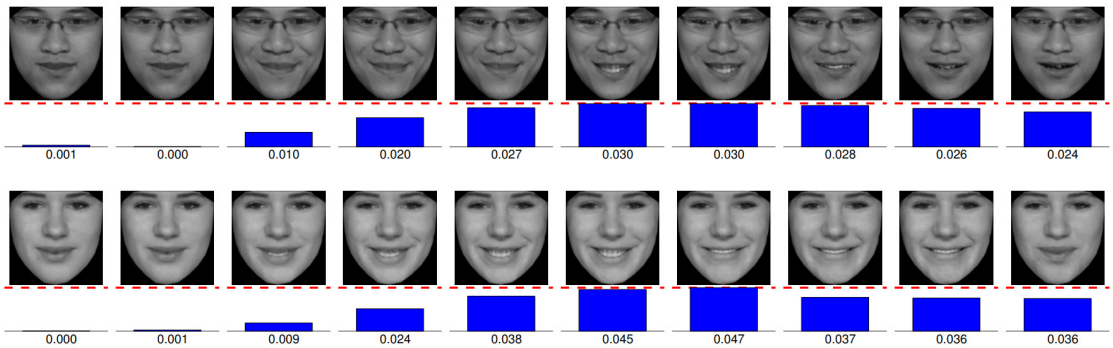

図 2.3 表情による年齢推定カリブレーション

2.3 マルチタスク学習法に注目した研究

マルチタスク学習(図 2.4)は，メインタスクが関連タスクの学習信号が持つドメイン固有の情報を利用する推論型移動学習法である。 関連タスクの学習信号が持つドメイン固有の情報を帰納的バイアスとして利用し，メインタスクの汎化性能を向上させる機械学習法である[12]。年齢推定問題に対してマルチタスク学習法も研究成果をいくつ出せた。

[13]と同様に，Yi ら[14]は年齢推定に CNN を用いて，顔の異なる領域から特徴を抽出し，測定基準として平均二乗損失を導入している。Niu.Z ら[9]は年齢に関する連続的な特徴に注目し，順序 CNN を学習させ，より良い最終結果を出した。また，[15]では，softmax の分類結果を直接利用するのではなく，各ニューロンの softmax 出力を年齢の重みとして利用し，加重平均値を算出するという別の方法を用いており，その結果，より良い性能を示すことが分かった。[16]，[17]では，年齢推定にマルチタスク学習法を用い，他の複数の顔特徴量を共同で学習し，各タスクの性能を向上させた。ディープリグレッションフォレスト(DRF) [8]はランダムフォレストと CNN を組み合わせて使用し，より良い性能を得た。 しかし，ほとんどの年齢特徴学習アプローチは， 単一のタイプの年齢特徴の学習に焦点を当て，年齢パターンに大きな影響を与える性別や人種などの他 の appearance 特徴を無視している[18]。Yo らの研究は，年齢と性別を認識のために新しい深層学習アーキテクチャを使用し[19]，ジェンダーの推論を伴う条件付き年齢推定を改善した。しかし，年齢と性別の特徴量の相関が弱いため，推定結果の性能を改善する余地がある(図 2.5)。この問題を解決するため，顔ランドマークを用いたマルチタスク学習を活用し，同じ入力画像から複数の特徴を学習し，これらの特徴を融合してより識別性が高く頑健な年齢推定の手法を構想する。

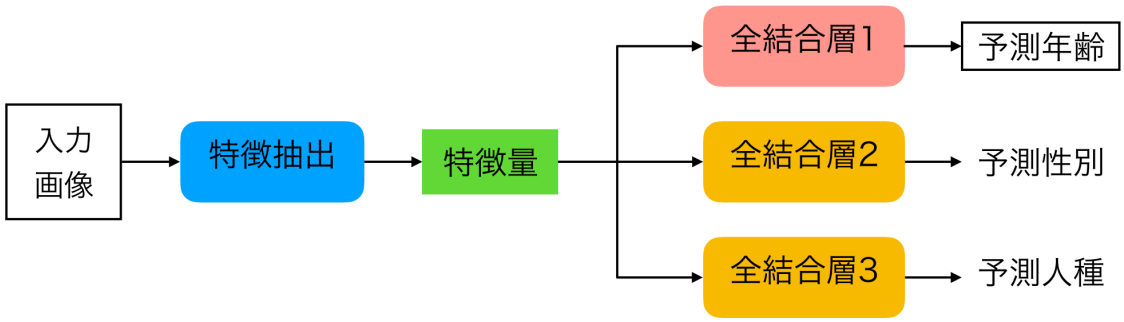

図 2.4 マルチタスク学習

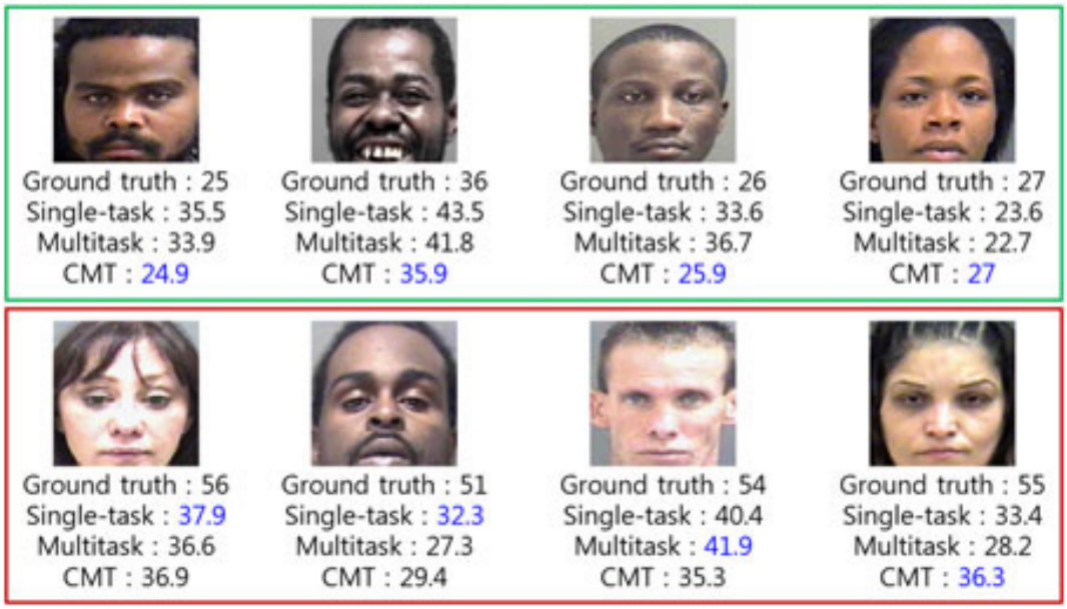

図 2.5 既往マルチタスク学習の精度不足

2.4 まとめ

本章では，顔画像からの年齢推定に関して従来研究を調査・分析し，関連ディープラーニングの特性について解説した。そして，近年の技術についてデータセット，ポーズ，マルチタスク学習の３つ角度から既往研究をまとめて，現存の問題点，改善すべきところなど課題を抽出した。年齢推定方法をより広い領域で応用できるため、従来研究より年齢推定の精度を向上することが必要である。

参考文献

1. G. Guo and G. Mu. Human age estimation: What is the influence across race and gender? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 71–78, 2010. 1

2. G. Guo, G. Mu, Y. Fu, C. Dyer, and T. Huang. A study on automatic age estimation using a large database. In Proceedings of the IEEE International Conference on Computer Vision, pages 1986–1991, 2009. 1

3. T. Perrett and D. Damen. Recurrent assistance: Cross- dataset training of lstms on kitchen tasks. In Computer Vi- sion Workshop (ICCVW), 2017 IEEE International Conference on, pages 1354–1362, IEEE, 2017.
4. J. Li, S. Xiao, F. Zhao, J. Zhao, J. Li, J. Feng, S. Yan, and T. Sim. Integrated face analytics networks through cross- dataset hybrid training. In Proceedings of the 2017 ACM on Multimedia Conference, pages 1531–1539, 2017.
5. Z. Kuang, C. Huang, and W. Zhang. Deeply Learned Rich Coding for Cross-Dataset Facial Age Estimation. In ICCVW, pages 338-343, 2015.
6. K. Chang, C. Chen, and Y. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In CVPR, pages 585–592, 2011.
7. S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao. Using ranking-CNN for age estimation. In IEEE ICCV, pages 5183–5192, 2017.
8. W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille. Deep Regression Forests for Age Estimation. In IEEE CVPR, pages 2304–2313, 2018.
9. Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4920–4928, 2016.
10. Z. Ji, C. Lang, K. Li and J. Xing. Deep Age Estimation Model Stabilization from Images to Videos. In International Conference on Pattern Recognition, 2018.
11. W. Pei, H. Dibeklio˘ glu, T. Baltrušaitis and D. Tax. Attended End-to-End Architecture for Age Estimation From Facial Expression Videos. In IEEE Transactions on Image Processing, volume 29, pages 1972-1984, 2019.
12. Caruana, R. Multitask Learning. Machine Learning 28, 41–75 (1997).
13. Y. Kwonand, and N. Lobo. Age classification from facial images. In IEEE CVPR, pages 762–767, 1994.
14. D. Yi, Z. Lei, and S. Z. Li. Age estimation by multi-scale convolutional network. In IEEE ICCV, pages 144–158, 2015.
15. R. Rothe, R. Timofte, and L. V. Gool. Deep expectation of real and apparent age from a single image without facial landmarks. Int. J. Comput. Vis., 126(2):1–14, Aug. 2016.
16. F. Wang, H. Han, S. Shan, and X. Chen. Multi-task learning for joint prediction of heterogeneous face attributes. In IEEE FG, pages 173–179, 2017.
17. H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. IEEE Trans. Pattern Anal. Mach. Intell., Aug. 2017.

18. M. Duan, K. Li and K.L. An ensemble CNN2ELM for age estimation. IEEE Trans. Inf. Forensics Secur, 13(3), 758–772, 2017.
19. B. Yoo, Y. Kwak, Y. Kim, C. Choi and J. Kim. Deep Facial Age Estimation Using Conditional Multitask Learning With Weak Label Expansion. In IEEE Signal Processing Letters, 25(6):808–812, 2020.

第 3 章 Cross-Dataset 学習法

3.1 緒論

既存の年齢ラベル付き顔画像データセットのほぼ全ては，単一人種の顔画像を含んでいる。いくつかの先行研究は，各人種の年齢を別々に推定することを提案している[1]，[2]。しかし，ほとんどのオープンデータセットが各人種のデータが不十分であり，一部のデータが誤ったラベルを持つという新たな問題に直面する。モデルに特化したデータセットを作成するという選択肢は高価であり，データを手作業で収集して，十分なラベル付けることも困難である。この問題に対処するために，より多くのデータをラベリングする代わりに，ある人種の大規模な高品質データセットを併用して，別の人種の顔の画像を含む小規模または低品質のデータセットに対する年齢推定の性能を向上させることができる[3]。また，この方法は，学習したモデルの汎化性を向上させ，実際のアプリケーションにおいてより信頼性の高いモデルを実現することができる。

このクロスデータセット年齢推定アプローチの問題点は，対象データセットが少数の高品質な学習データしか持っていないため，学習過程が不安定になり，精度の低い学習結果になってしまうことである。本章では，より高精度な年齢推定を実現するために，Cross-Dataset CNN（CDCNN）というエンド・ツー・エンドの学習モデルを提案する。

CDCNN モデルは，年齢推定のために顔画像から特徴を抽出する CNN を用いる。この学習方法で抽出された特徴は，アクティブアピアランスモデル（AAM）[4]や BIF モデル[5]で用いられるような手作りの年齢特徴と比べて，より識別性が高く，顔の見た目の変動に対してロバスト性であることが期待される。本章では，年齢推定を回帰問題ではなく，分類問題として扱い，Softmax 関数を用いて，出力をマッピングし，平均期待値を算出する。また，対象データセットでより高い性能を得るために，このモデルは複数のデータセットからラベル付けされたデータを追加で利用する 。実年齢推定の標準的なベンチマークとして，Asian Face Age Dataset (AFAD) [6], Cross-Age Celebrity Dataset (CACD) [7], Craniofacial Longitudinal Morphological Face Database (MORPH) [8] の 3 つデータベースを用いて実験が行われている。さらに，Cross-Dataset 学習は，結合されたデータセットに対して行われ，その性能結果は 1 つのデータセットに対する結果と比較される。実験結果より，提案モデルが ADAD と CACD において，複数の最先端手法を上回ることが示された。

3.2 提案手法

3.2.1 フェイスクロッピング

顔画像の背景が変わると性能が変化することがある。また，データセット内の顔の配置が異なる場合も，性能のばらつきにつながる。周囲の画素の影響を防ぐために，すべての顔画像をスクイーズして 256×256 画素をランダムに 224×224 画素に切り出し(図 3.1)，ニューラルネットワークに入力する。顔画像の切り出しは，すべての顔がランダムに異なる場所に配置されるようにする。位置は元のデータセットに関係なく，またこの方法によってできたモデルは実際のアプリケーションにおける様々な顔位置を合わせることができるほどロバスト性がある。 すべてのデータセットをピクセル単位の精度で一貫して整列させたわけではないが，この切り出し方法は実験にとって精度が十分である。

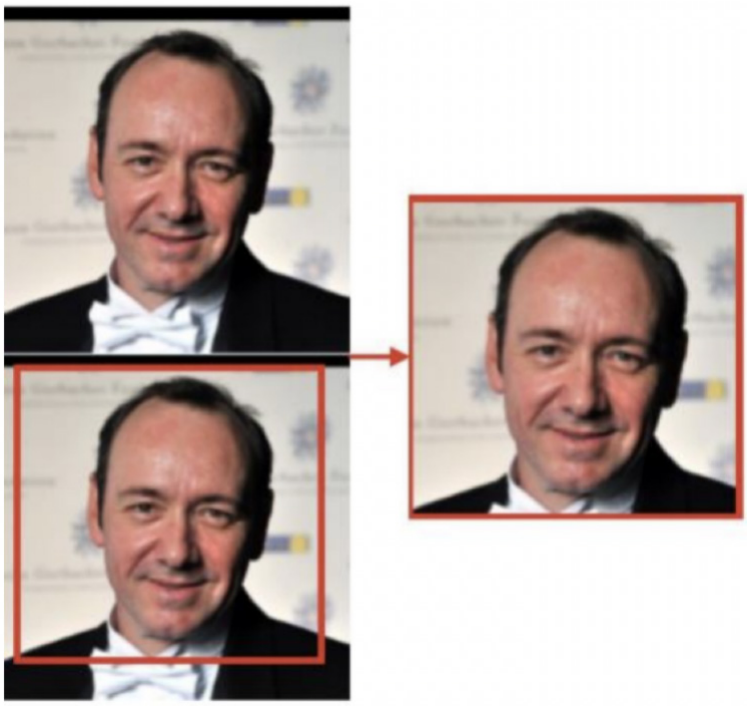

図 3.1 フェイスクロッピング

3.2.2　CNN 構造

顔画像から年齢を推定するために，CNN を導入する。このネットワークは顔画像を入力として用いて，年齢推定結果を出力する。CNN の学習には，年齢ラベルを持つ複数の顔画像データセットが用いられる。 CNN ネットワークとして VGG16 [9] (図 3.2)アーキテクチャが選ばれた理由は，(i) 深いが扱いやすいアーキテクチャである，(ii) VGG16 を用いた過去の研究が ImageNet チャレンジで素晴らしい結果を出している，(iii) VGG16 を用いて事前に学習したモデルがあり，学習が容易にできる。他にも複数より複雑なネットワークがあるが、VGG16 よりトレーニング時間がかかるし、収束も困難だし、予測結果の精度も同じぐらいので、私はトレードオフして、VGG16 を利用した。VGG16 は AlexNet[10]のような初期のネットよりも層数が多く，16 層であり，そのうち 13 層が畳み込み層，3 層が全結合層となっている。VGG-16 では，(3×3)ピクセルという小さなフィルタが採用されており，初期の CNN と比較して，よりシンプルな構成になっている。しかし，より深いネットワークによって，より複雑な関数関係を表すことができる。本研究のすべての実験において，モデルは Imdb-wiki データセット[26]で事前に年齢推定問題に対して学習したパラメータを用いている。その後，年齢推定に適応した各顔データセットの画像でトレーニングして、CNN をファインチューニングする。ファインチューニングにより，ネットワークはターゲットデータセットの特徴を得ることができるため，推定結果の性能を最適化することができる。

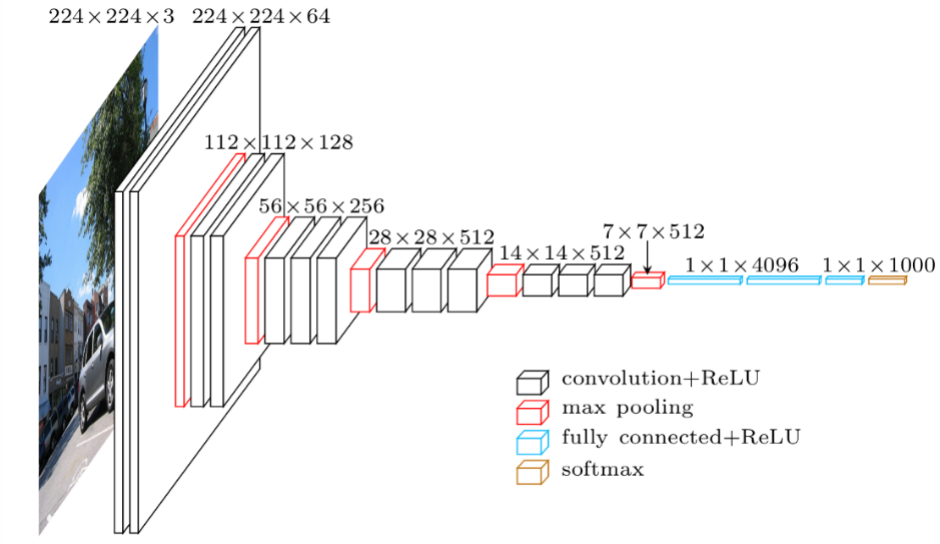

図 3.2 VGG-16 CNN 構造

3.2.3　出力層と期待値

VGG16 ネットワークは最初に ImageNet で学習して，様々な物体に対して 1,000 個の出力で分類を行った。出力層からの各ニューロンはオブジェクトクラスを表す。一方，年齢推定は年齢の値が連続的であるため，回帰問題でもあるし、各年齢クラスを分類することも可能である。回帰学習では，VGG16 の最終層に 1 ニューロンだけを出力し，ユークリッド損失関数が用いられる。しかし，このような回帰モデルを直接学習させると，外れ値によってかなりの誤差が生じるため，比較的に不安定である。そのため，ネットワーク全体が収束しにくい大きな勾配を作り出し，不安定な予測結果になる。R. Rothe ら[26]の研究結果により、年齢推定問題を分類問題と扱うと、回帰もでるより、ネットワークも収束しやすいし、予測結果の精度も高くなる。なので，私も年齢推定問題を分類問題として扱って，年齢の異なる値を X クラスに離散化して、各 $\boldsymbol{x_i}$ は 1 つの年齢値をカバーする。。

分類問題なので、トレーニングするときは、クロスエントロピー誤差を利用して、誤差逆伝播法により，CNN パラメータを学習し、$|\boldsymbol{X}|$ニューロンからのソフトマックス関数を用いた出力確率により予測値を計算する。ここで，P = 1, 2, 3, ..., X は最終層の出力，$\boldsymbol{p_i}$ はソフトマックス関数で正規化したクラス i の確率を表す。実験結果により，予測年齢は確率加重平均値を利用すれば、直接に一番確率高い分類結果をを利用するより場ロバスト性や精度を向上できることを示していた。学習の流れは図 3.3 に示した。

$$\boldsymbol{E(P)} = \sum_{\boldsymbol{i=1}}^{|\boldsymbol{X}|} \boldsymbol{x_i} \cdot \boldsymbol{p_i}$$

3.2.4　Cross-Dataset 学習

データセットの結合は，データセット間学習における重要な操作である。例えば，ある人物の人種と性別のように，同一人物に対して異なるラベルを持つ列を交互的に結合することが頻繁に要求される。従来の Cross-Dataset 学習では，異なるクラスに対して異なるラベルを持つ複数のデータセットを用いるため，ラベルが異なるデータセットで重複によって衝突の可能性がある。しかし，年齢ラベルは全て整数であるため，年齢推定に用いるデータセットが異なっても，衝突の可能性がない。提案した年齢推定に対して Cross-Dataset 学習法は複数のデータセットを混合し，一つ CNN を用いて一つの損失関数で学習できる。したがって，Cross-Dataset 学習法はシンプルな構造で，異なるデータセットに独立に埋め込まれた知識を把握することができるため，より高い精度が達成できる。ただオーバーフィッティングを防ぐため，異なるデータセットからのデータのバランスを注意しなければならない。本研究では最小のデータセットから得られる画像の数が最大のデータセットから得られる画像の数の半分以上であることを確保する。

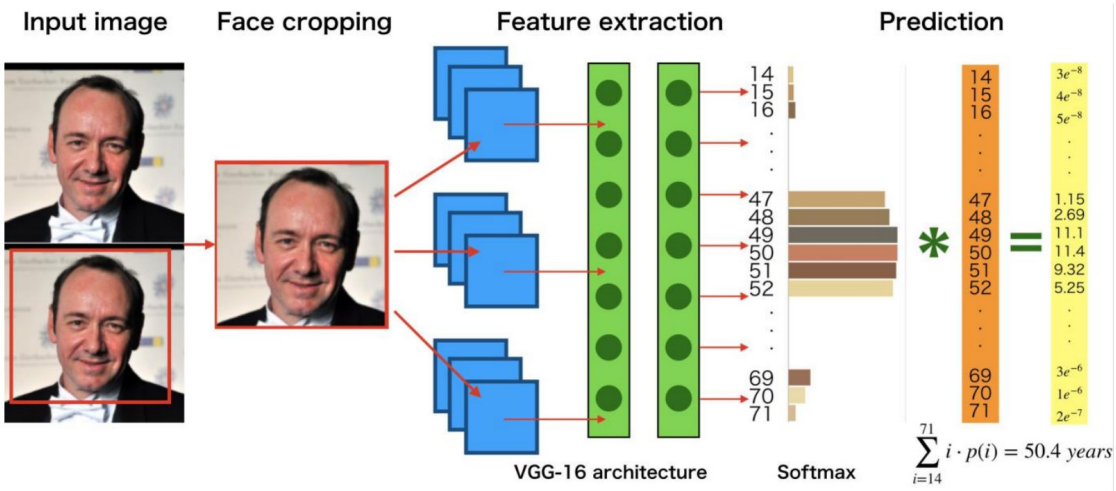

図 3.3 提案した Cross-Dataset 学習の流れ

3.3 年齢推定実験パラメータ

3.3.1 評価プロトコール

これらの実験では，結果を定量的に評価するため、2 つの基準を用いている。

一般的な評価指標である平均絶対誤差（MAE）は，異なる年齢推定アルゴリズムの性能を測定するために使用されている。平均絶対誤差にあるように，真実と予測年齢の間の絶対誤差を MAE として平均化したものであり，以下のように定義される。

$$\boldsymbol{MAE = \frac{1}{K}\sum_{i=1}^{K}|\tilde{y}_i - y_i|}$$

ここで，K はデータサンプルの総数，$\boldsymbol{y_i}$はグランドトゥルース年齢，$\boldsymbol{\tilde{y}_i}$は i 番目のサンプルの予測される年齢を表す。一般的に，より優れた年齢推定アプローチは，テスト結果における MAE 値が小さくなる。

もう一つの評価指標である累積スコア（CS）も，異なる年齢推定アプローチを評価するために使用されている。CS は，誤差 e（年齢推定における MAE）が与えられた数値 i（年齢推定における年）より小さくなるサンプルの割合であり，以下のように定義される。

$$\boldsymbol{CS(i) = K_{e \le i}/K}$$

関連する論文では，0 から 10 までの様々な CS の値を与えているか，単に CS の固定値を設定している。ここでは，[11]，[12]，[13]のアプローチと同様に ，常に i = 5 が使用されている。すべての論文が CS 値を報告しているわけではないので, CS 値は一部の比較対象についてしか提供することができない。

3.3.2 データセット

本章では，年齢推定の 3 つの基本データセットを用いて，それらの 3 つの異なる組み合わせをクロスデータセット学習に用いている。表 3.1 に各データセットのサイズ（学習用とテスト用の分割分布も含む）を示す。

データセット	画像枚数
CACD-MORPH-AFAD	136,473
CACD(used)	18,171(200 celebs)
MORPH	55,128
AFAD(used)	50,449
CACD-MORPH	86,024
AFAD-MORPH	105,577
CACD(total)	163,446(2000 celebs)
AFAD (total)	164,432

表 3.1 各データセットの分割分布を含む画像枚数

CACD [7]（図 3.5），2,000 人の有名人の 163,446 枚の人物画像が収録されている。これらの画像は，各著名人の年号と名前によって検索エンジンを使ってオンラインで収集された。写真の撮影日と有名人の生年月日情報を計算して得られる年齢ラベルは，そのため部分的なラベリングや画像エラーを含むノイズの多いデータである。性能をよりよく評価するために，有名人ごとに，学習用 1800 人の画像，検証用 80 人，テスト用 120 人に分けられる。このうち，検証用とテスト用 200 人の画像に対して，ノイズの多い画像を手動で除去して、クリーンなサブセットを作成した。本研究の実験では，18,171 枚の画像からなるクリーンなサブセットのみを使用した。

MORPH [8] （図 3.6）は 13,000 人以上から 55,000 枚のユニークな画像を含み，手動データ収集による一般公開の顔データベースとしては最大である。このデータセットに含まれる年齢は 16 歳から 77 歳で，中央値は 33 歳である。本研究では，年齢が 16 歳から 71 歳までのサブセットのみを実験に採用し，データセットのバランスをとるために約 55,000 枚

の写真を利用されている。

AFAD [6] （図 3.7）は 160K 枚以上の顔画像とそれに対応する年齢・性別ラベルを含む，年齢推定に用いられる新しいデータセットである。このデータセットは，アジア人の顔を対象としたデータセットであるため，すべての顔画像は，アジアの学生や大学院生に広く利用されている RenRen Social Network（RSN）から収集されている。データセットに含まれる年齢は，15 歳から 40 歳以上までと幅広い。データセットのバランスをとるため，18 歳から 39 歳までの 22 の連続した年齢別の 59,344 枚の画像を含む AFADLITE と名付けられたサブセットのみが実験に用いられた。

使用したすべてのデータセットの様々な年齢分布は図 3.4 に示されている。CACD は 20 歳から 60 歳の間でバランスの取れた年齢分布をしているが，この範囲外のサンプルはわずかしかない。MORPH は，2 つのデータソースから収集されているため，年齢分布曲線のピークは明らかに約 20 歳と 40 歳の 2 つである。AFAD は 18 歳から 40 歳までを対象としており，高校生や大学生の画像が多いため，20 歳にピークがある。FG-NET は AFAD よりもさらに年齢分布が若く，ほとんどのサンプルが 30 歳未満である。CACDMORPH は MORPH と同様の分布，CACD-MORPHAFAD と AFADMORPH は AFAD と同様の分布となった。

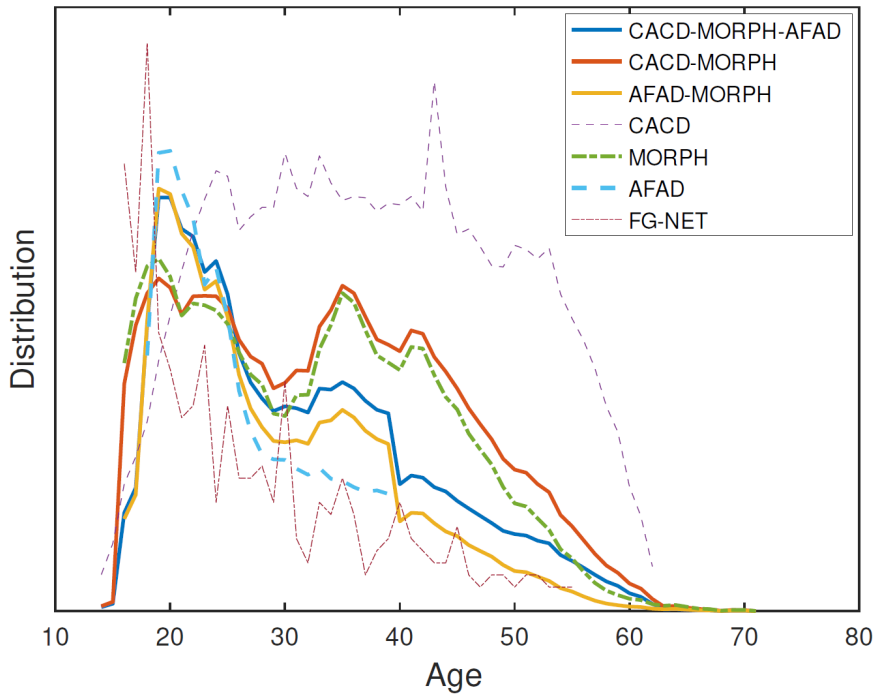

図 3.4　各データセットの年齢分布

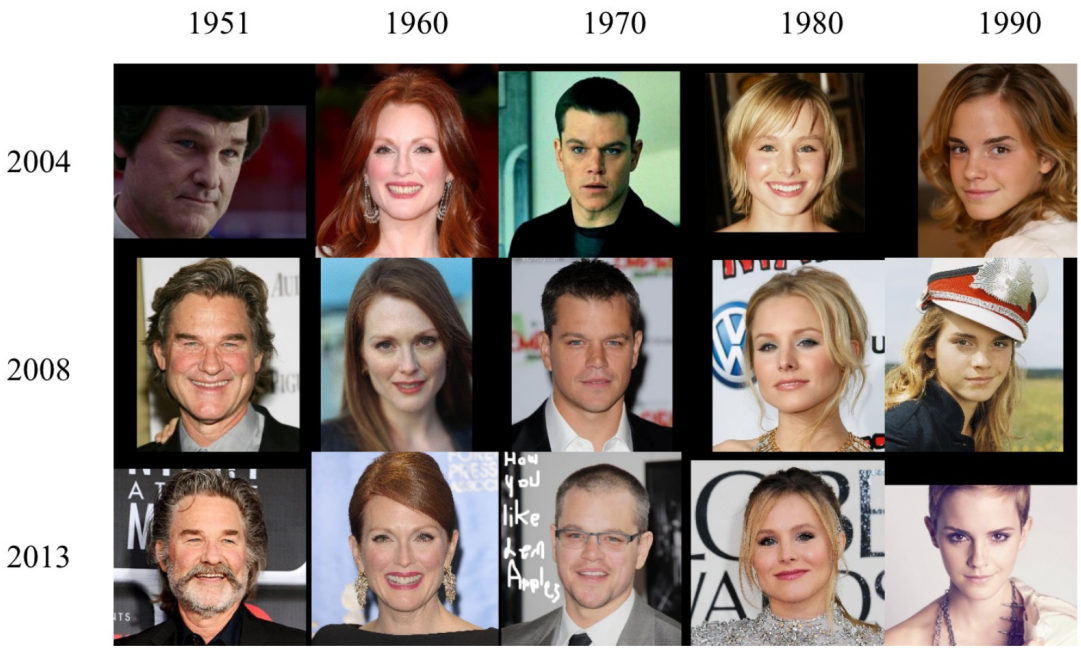

図 3.5　CACD データセット

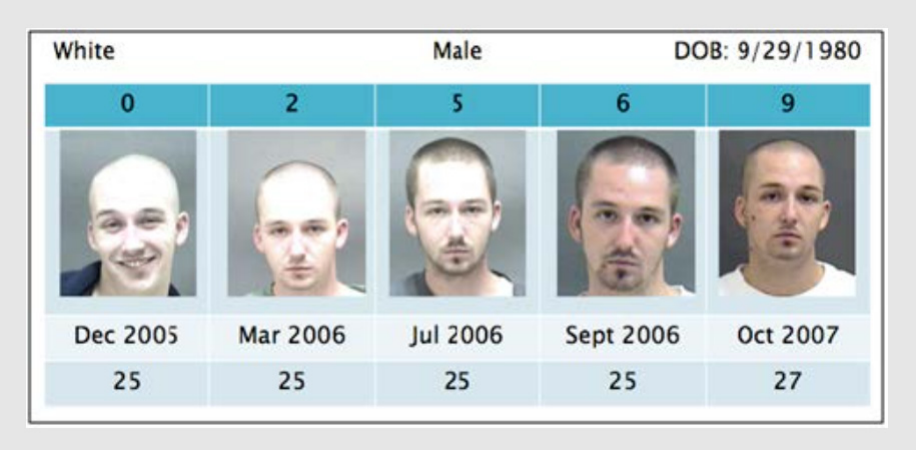

図 3.6　Morph データセット

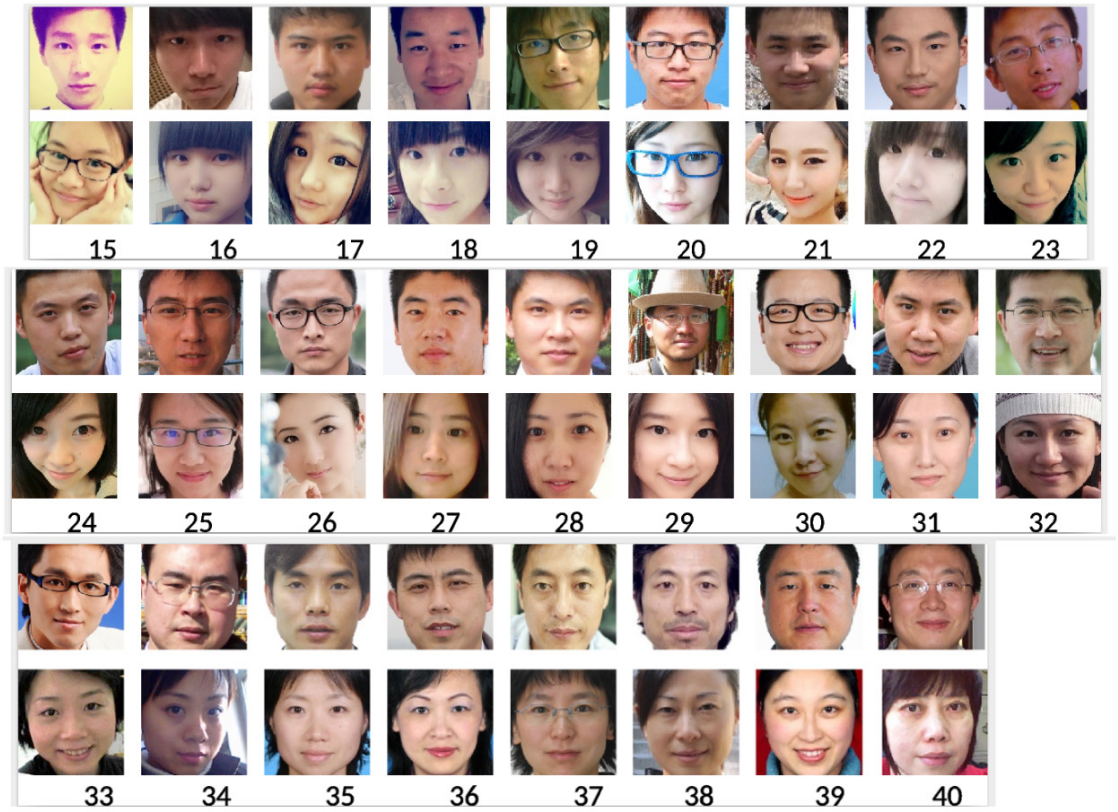

図 3.7　AFAD データセット

3.3.3 実装の詳細

前述の通り，本手法における CNN は分類モデルとして学習される。予測年齢はソフトマックス正規化した確率加重出力により算出される。

すべての実験において，Imdb-wiki 上で学習させた初期化された重みを使用する。このモデルは，分類対象のデータセットで出力ニューロンを微調整し，年齢クラスの数と同じ数のニューロンを使用する。学習前に，Imdb-wiki [26]で学習したのと同じ重みで全ての実験モデルを初期化した。学習段階では，全てのデータセットが分割され，65%のデータが重みの学習に，20%が検証に，そして最後の 15%がテストに使用される。検証セットで CNN がオーバーフィットした後，学習プロセスを終了する。学習バッチサイズは 64 であり，ドロップアウト率は 0. 5 である。確率的勾配降下法（SGD）を用いて，初期学習率 0.01 でネットワークを学習し，15k イテレーションごとに 1/10 に減少させる。異なるデ

ータセットに対して年齢クラスの数が異なるため，出力ニューロンの数も変更する。全てのネットワークは Nvidia GTX1080 GPU の上で Caffe フレームワーク[14]を用いて学習した。

3.4 実験結果

まず，3 つのベースラインベンチマークにおいて，提案手法のアーキテクチャの性能を他の研究の性能と比較する。次に，AFAD と CACD ベンチマークで学習させた提案するクロスデータセット手法の結果を，ベースライン結果や他の研究の結果と比較する。CACDMORPH と AFADMORPH のクロスデータセットは，MORPH，CACD，AFAD の各データセットのデータを組み合わせている。次に，公平な比較を行うために，同じネットワークと学習プロセスを使用する。また，MAE と CS5 の値これら全ての設定について評価する。

3.4.1 ベースライン

実験には 3 つのベースラインベンチマークを用いた。本節では，提案手法の年齢推定に関する性能を示す.

MORPH について提案した CNN 構造は，MORPH 訓練データセット上で CNN を微調整した場合，MAE 値 2.76 を達成する。MORPH は[15], [16], [17], [18]と同様 85%/15%の割合でランダムに訓練/試験セットに分割されている。実験誤差を減らすため，異なるランダム分割で 5 回実験を繰り返した。5 回の実験の平均値を最終的な結果として，ロバストな性能を示した。定量的な結果は表 3.2 にまとめられている。MORPH は年齢推定を研究する多くの研究者にとって最も一般的なデータセットであるため，この 0.6 年のマージンは最先端モデルに匹敵するものである。提案した CDCNN 手法は ，画像数が不十分なデータセットや画質の悪い画像に対して，他のデータセットから特徴を把握して性能を向上させるために用いられるため，本研究では MORPH をデータ横断学習にのみ用い，主に高品質なラベル付き顔画像を持たないデータセットに着目している。

手法	MAE	CS
SVR [19]	3.48	78.8%*
OR-CNN [6]	3.27	73.0%*
Ranking-CNN [20]	2.96	85.0%*
Our method	**2.76**	**84.9%**
MA-SFV2 [21]	2.68	90.0%*
CORAL [22]	2.64	N/A

DRFs [23]	2.17	91.3%

表 3.2 .MORPH での性能比較（*: 値は論文中の CS 曲線からの推定値）

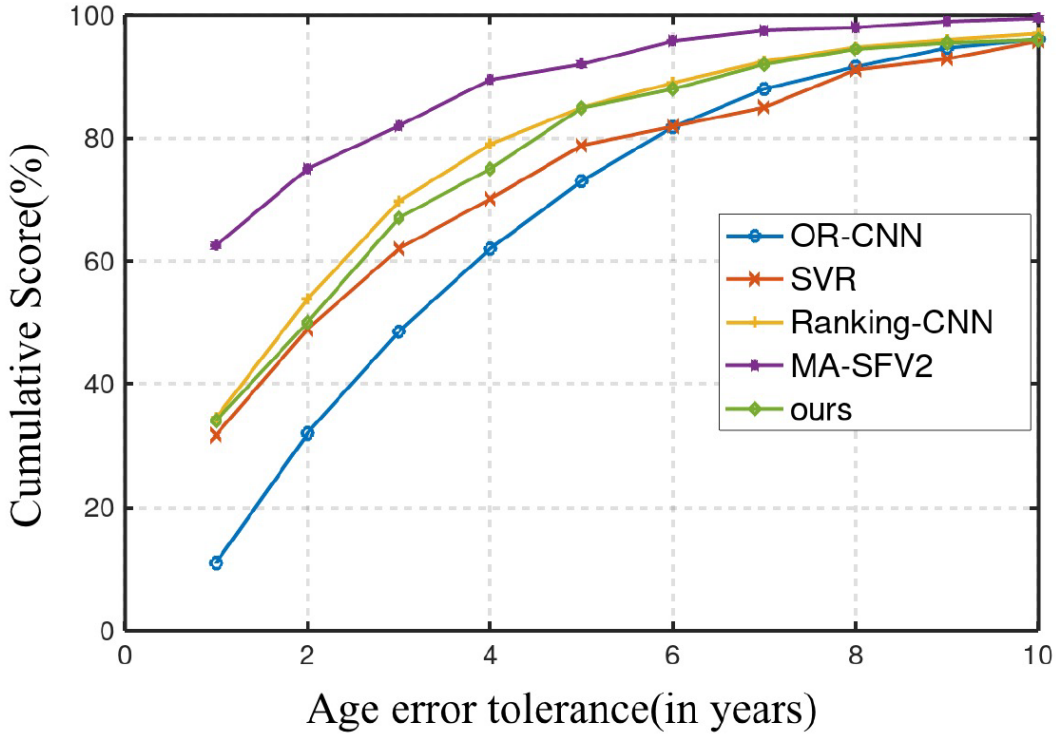

図 3.8　MORPH での CS カーブ

AFAD において提案した CNN 構造は，AFAD の訓練データセットでネットワークを訓練したとき，MAE 値 3.30 を達成した。AFAD はランダムに 85%/15%の学習/テストセットに分割され， 5 回の実験が繰り返されて，ロバストな性能を示した。 定量的な結果は表 3.3 にまとめられている。明らかなマージンはないが，最先端の性能が達成された。AFAD において年齢推定の研究が少ないかつ古いため、そのまま VGG-16 ネットワークでより良い結果ができたが、精度を改善する余地がかなりある。

手法	MAE
BIFS + OHRank [13]	3.84
CORAL [22]	3.48
OR-CNN [6]	3.34
Our method	**3.30**

表 3.3　AFAD における性能(MAE)の比較

CACD では，CACD の訓練用データセットに対して CNN を微調整したところ，MAE 値 4.58 を達成した。CACD はランダムに 85%/15%の学習/テストセットに分割され，5 回実験が繰り返されて，5 回の実験の平均値を最終結果とした。定量的な結果は表 3.4 にまとめられている。CACD は MORPH や AFAD と比較して，データ量は多いが，部分的なラベリングや画像エラーでノイズが多い。そこで，CACD の 10 万枚以上の写真の代わりに，手動でクリーニングした 18,171 枚の写真からなるサブセットのみを使用した。 明らかなマージンはないが，最先端の性能が達成された。CACD において年齢推定の研究が少ないかつ古いため、そのまま VGG-16 ネットワークでより良い結果ができたが、精度を改善する余地がかなりある。

手法	MAE
CORAL [22]	5.35
dLDLF [16]	4.73
DRFs [23]	4.60
Our method	**4.58**

表 3.4　CACD における性能(MAE)の比較

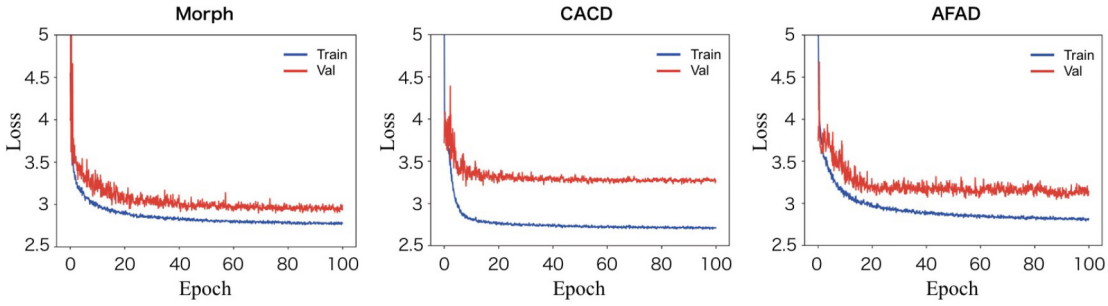

図 3.9　MORPH，CACD，AFAD の学習曲線

3.4.2 Cross-Dataset トレーニング

特に CACD や AFAD のような高品質なラベル付き顔画像が少ないデータセットでは，回帰法は勾配が大きく，ネットワーク全体が収束しにくいため，分類法よりも安定性に欠ける。したがって，表 3.5，表 3.6 に示したデータセットでは，より複雑な回帰モデルと比較しても，提案した単純な分類モデルの方が良い性能を得ることができる。

AFAD や CACD のようなデータセットでは，高品質なラベル付き顔画像が少ないため，様々な複雑な CNN 構造を用いて性能を向上させることは困難である。この問題を解決し，学習モデルを改善する代わりに，年齢推定用データセットの品質を向上させる CDCNN 手法を提案する。AFAD-MORPH と CACD-MORPH のデータセットを用いて学習を行った結果，大きな改善が見られた。

提案した CNN 構造は，AFAD-MORPH トレーニングデータセットでネットワークをトレーニングし，AFAD テストセットでテストした場合，MAE 値 3.11 を達成する。AFAD-MORPH データセットでは，テストデータが学習されないように，AFAD の分割学習部分のデータのみを使用した。定量的な結果は表 3.5 にまとめられた。この Cross-Dataset 学習方法により，以下のように結果が改善された。[6]で報告された従来の最先端手法の結果より 0.2 年良い。

手法	MAE
BIFS + OHRank [13]	3.84
CORAL [22]	3.48
OR-CNN [6]	3.34
Our method (CDCNN 使わず)	**3.30**
Our method (CDCNN)	**3.11**

表 3.5　AFAD における性能(MAE)の比較

提案した CNN 構造は，CACD-MORPH トレーニングデータセットでネットワークをトレーニングし，CACD テストセットでテストした場 合，MAE 値 3.96 を達成する。CACDMORPH データセットは，テストデータが学習されないように，CACD の分割学習部分のデータのみを使用していた。定量的な結果を表 3.6 にまとめられた。この Cross-Dataset 学習方法により，以下のように結果が改善された。[23]で報告された従来の最先端手法の結果より 0.7 年良い。

手法	MAE
CORAL [22]	5.35

dLDLF [16]	4.73
DRFs [23]	4.60
Our method (CDCNN 使わず)	**4.58**
Our method (CDCNN)	**3.96**

表 3.6　CACD における性能(MAE)の比較

3.5 結論

本章では年齢推定のための顔データセットにおける画像数の不足と画質の悪さの問題を解決するために，異なる特徴量でラベル付けされた複数のデータセットを年齢推定用に共同学習する CDCNN 法を提案し，CACD と AFAD で最先端の結果を示した。

本章の主な貢献は以下の通りである。(1)クロスデータセット学習によって年齢推定のための複数のデータセットを共同で学習させたのは初めてである．(2) 年齢推定問題を従来の回帰タスクではなく，分類タスクとして扱うことで，画像数が十分でない顔データセットや画質の悪い画像において，より良い性能を発揮する。(3) 提案する CDCNN モデルは，CACD において MAE 値 3.96，AFAD において MAE 値 3.11 と，最新の年齢推定精度を達成することができた。

しかし，ベースラインネットワークの精度はまだ改善可能であり，より多くの顔データをトレーニングに使用することができる。将来的には，年齢ラベルを持つより多くの顔画像と，残差ネットワーク[24]のようなより深いネットワークを使用することで，性能を向上させることができる。また，138M のパラメータを持つ VGG16 とは異なり，39.7K のパラメータしか持たない C3AE [25]のような小規模で効率の良いモデルを作ることも可能である。最終的には，提案した CDCNN 法は性別，人種，表情，健康状態の推定など，他の顔特徴推定タスクにも利用する可能性がある

参考文献

1. G. Guo and G. Mu. Human age estimation: What is the influence across race and gender? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 71–78, 2010. 1

2. G. Guo, G. Mu, Y. Fu, C. Dyer, and T. Huang. A study on automatic age estimation using a large database. In Proceedings of the IEEE International Conference on Computer Vision, pages 1986–1991, 2009. 1

3. T. Perrett and D. Damen. Recurrent assistance: Cross- dataset training of lstms on kitchen tasks. In Computer Vi- sion Workshop (ICCVW), 2017 IEEE International Conference on, pages 1354–1362, IEEE, 2017.

4. Cootes TF, Edwards GJ, and Taylor CJ. Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 23(6):681–685, 2001.
5. G. Guo, G. Mu, Y. Fu, and T. Huang. Human age estimation using bioinspired features. In IEEE CVPR, pages 112–119, 2009.
6. Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4920–4928, 2016.
7. B. Chen, C. Chen, and W. H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. IEEE Trans. Multimedia, 17(6):804–815, 2015.
8. K. Ricanek and T. Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In Proc. FG, pages 341–345, 2006.
9. Simonyan. K, Zisserman. A. Very deep convolutional networks for largescale image recognition. International Conference on Learning Representations, 2015.
10. Krizhevsky. A, Sutskever. I, Hinton. GE. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (NIPS), pages 1097–1105, 2012.
11. I. Huerta, C. Fernandez, and A. Prati. Facial age estimation through the fusion of texture and local appearance descriptors. In Porc. ECCV Workshops, pages 667–681, 2014.
12. K. Chen, S. Gong, T. Xiang, and C. L. Chen. Cumulative attribute space for age and crowd density estimation. In Proc. CVPR, pages 2467–2474, 2013.
13. K. Chang, C. Chen, and Y. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In CVPR, pages 585–592, 2011.
14. Jia. Y, Shelhamer. E, Donahue. J, Karayev. S, Long. J, Girshick. R, Guadarrama. S, Darrell. T. Caffe: Convolutional architecture for fast feature embedding. In International Conference on Multimedia, pages 675–678, 2014.
15. B. B. Gao, C. Xing, C.W. Xie, J.Wu, and X. Geng. Deep label distribution learning with label ambiguity. IEEE Transactions on Image Processing, PP(99):1–1, 2016.
16. W. Shen, K. Zhao, Y. Guo, and A. Yuille. Label distribution learning forests. In Proc. NIPS, page 834-843, 2017.
17. X. Geng. Label distribution learning. IEEE Transactions on Knowledge and Data Engineering, 28(7):1734–1748, 2016
18. X. Geng, K. Smith-Miles, and Z. Zhou. Facial age estimation by learning from label distributions. In Proc. AAAI, pages 451–456, 2010.
19. H. Liao, Y. Yan, W. Dai, and P. Fan. Age Estimation of Face Images Based on CNN and Divide-and-Rule Strategy. In Mathematical Problems in Engineering, 2018.

20. S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao. Using ranking-CNN for age estimation. In IEEE ICCV, pages 5183–5192, 2017.
21. X. Liu, Y. Zou, H. Kuang, and X. Ma. Face Image Age Estimation Based on Data Augmentation and Lightweight Convolutional Neural Network. In symmetry, 12(1):146, 2020.
22. W. Cao, V. Mirjalili, and S. Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. In Pattern Recognition Letters, volume 140, page 325-331, 2020.
23. W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille. Deep Regression Forests for Age Estimation. In IEEE CVPR, pages 2304–2313, 2018.
24. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.
25. C. Zhang, S. Liu, X. Xu, and C. Zhu. C3AE:Exploring the Limits of Compact Model for Age Estimation. In IEEE CVPR, pages 12587-12596, 2019.
26. R. Rothe, R. Timofte, and L. V. Gool. Deep expectation of real and apparent age from a single image without facial landmarks. Int. J. Comput. Vis., 126(2):1–14, Aug. 2016.

第 4 頭部姿勢推定と併用した学習法

4.1 緒論

近年，年齢推定[1,2,3]をはじめとする様々なコンピュータビジョンタスクにおいて，深層学習が重要な成果を上げている。しかし，これらの研究はすべて正面顔画像のみを含むデータセットを用いており，実際のアプリケーションの状況を十分に反映することができない。また，動画像やウェブカメラでは，顔画像とは異なり，頭部の姿勢が大きく変化するため，年齢推定に耐えられない誤差が生じる可能性がある。

本研究では，動画像やウェブカメラストリーム中の顔から年齢を推定する問題を解決するために，年齢推定と頭部姿勢推定を組み合わせたシステムを提案する。まず，顔画像の年齢推定にディープリグレッションフォレスト（DRF）[4]を用いて，正面顔画像に対して高精度な年齢推定を行う。一方，マルチロス畳み込みニューラルネットワーク(CNN)も頭部姿勢の推定に利用される[5]。そして，学習させたシステムを用いて，複数の動画からフレームごとに年齢と頭部姿勢を推定することができる。学習した年齢と頭部姿勢の対応付けを用いる場合，頭部姿勢に度数の閾値を設定し，頭部姿勢がこの閾値内にあるフレームのみ年齢推定を行い，映像から推定した年齢の値を絞り込むことができるようにする。

実験は2段階に分けて行われた。まず，300W-LP [6] (300W across Large Poses)データセットでmulti-loss CNNを学習させて，トレーニングしたモデルで頭部姿勢推定を行う。また，Cross-Age Celebrity Dataset (CACD) [7]とAsian Face Age Dataset (AFAD) [8] を，推定した頭部姿勢角度に基づいて分割し，正面と非正面の画像のサブセットに対して別々にDRFを学習させた。その結果，正面顔画像からの年齢推定は，異なる角度の顔画像からの推定よりも精度が高いことが示された。次に，様々な閾値を設定し，異なる角度によって正面画像のサブセットをトレーニングして，最適な閾値を決める。そして，学習したモデルを用いて，同一人物の顔の角度を変えて年齢を推定する実験を複数の動画フレームに対して行った。実験の結果，頭部ポーズ角度制約を用いた提案システムは，動画に対する推定誤差の標準偏差が，年齢推定を単独で行った場合よりも小さくなることが示された。この結果，提案方式は従来の方式と比較して，動画中の顔に対する年齢推定の精度と信頼性を向上させることができる。

4.2 提案手法

4.2.1 フェイスアライメント

顔の周辺環境は変化するため，検出性能も変化する。また，顔の位置合わせの方法が異なると，さらに性能の変化が生じる可能性がある。理想的な顔画像は，サイズが似ていて，正面から見たときに顔の中心が定まるもので，かつ，顔の並び方が固定ロケートで正規化され，背景がきれいになって姿勢推定の精度も高くなる。そこで，図 4.1 のように，画像から顔を求めるために MTCNN[9] 顔検出器を選択した。また，周辺画素の影響を最小限にするため，すべての画像を 256×256 にリサイズし，224×224 にランダムにクロップしていた。この切り出し処理により，元データによらず，画像の異なる位置にランダムに顔が配置された。この手法により，顔の位置が異なる様々なシーンにも特徴を把握できるロバスト性があるモデルを学習した。また，224×224 画素の画像は，VGG-16 ネットワークの入力サイズに適合している。

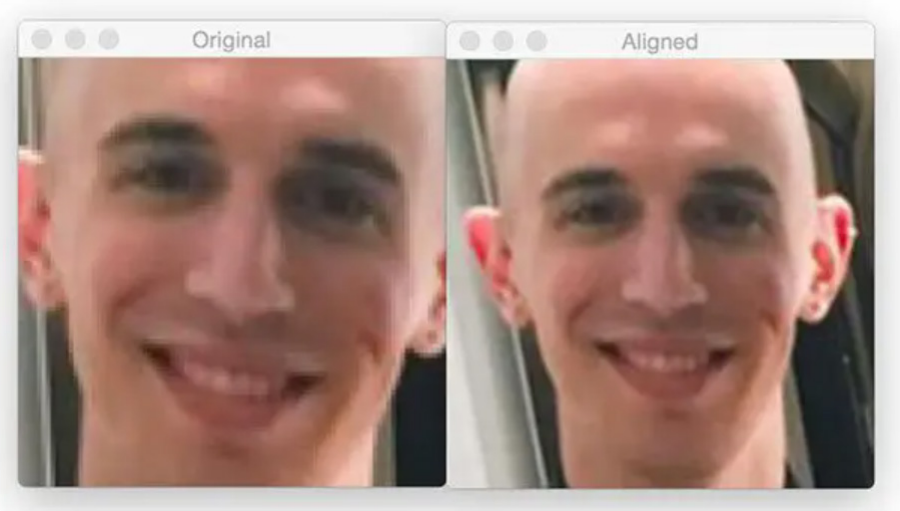

図 4.1 フェイスアライメント

4.2.2　DRF の構造

本研究では，顔画像から実年齢を推定するために，CNN とディープリグレッションフォレストを組み合わせたモデルを導入している。このモデルは，既知の年齢をラベルとして持つ顔画像データセットで学習される。本論文の学習プロセスは，[23]で用いたモデルと同様に，Imdb-wiki データセットから事前に学習した重みを用いて開始される。次に，年齢推定に用いる 2 つのターゲットデータセットに対して CNN の微調整を行う。この微調整により，CNN は各データセットの特徴，分布，バイアスを取得し，性能を最適化する。

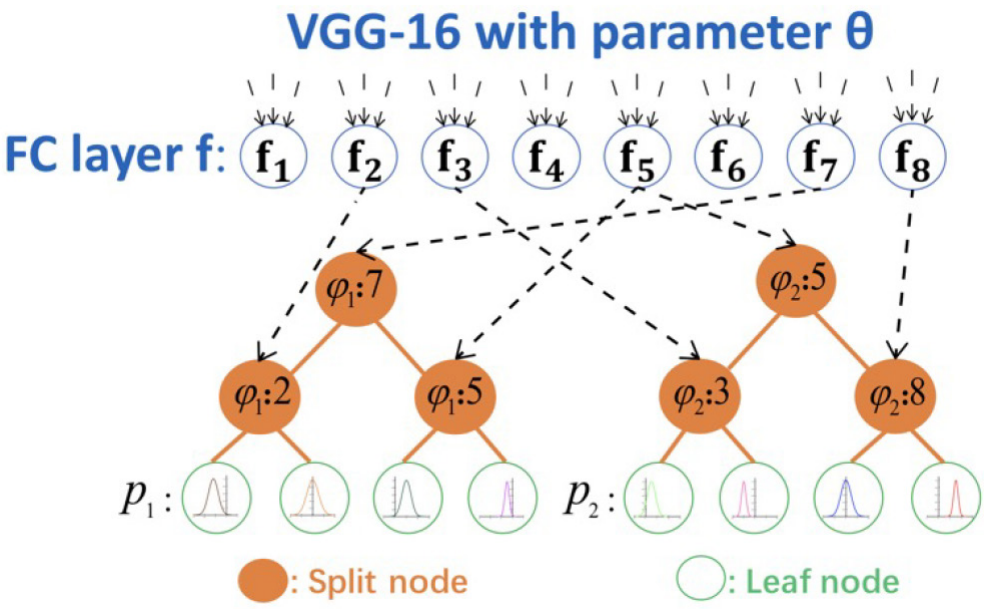

図 4.2 DRF 構造

上の青の円は，関数 **f** で定義された CNN の出力ニューロンを表し，その出力ニューロンはパラメータ Θ。これらのニューロンは全て VGG-16 の最後の全結合（FC）層から来たものである。真ん中のオレンジの円は分割ノード，下の緑の円はディープリグレッションフォレストの葉ノードを表している。$\phi 1$ と $\phi 2$ は各ツリーのインデックス関数を表している。黒破線の矢印は，各ツリーのスプリットノードから VGG-16 FC 層のニューロンへのマッピング関係を示す。各ニューロンは，異なるツリーのスプリットノードにマッピングする可能性がある。各ツリーはその葉ノードに対して独自の分布 π を持つ（葉ノード上の分布曲線で表される）。フォレスト全体の最終的な出力は，個々のツリーの予測値の混合

として計算することができる。パラメータ $f(;\ \Theta\ )$ と π はエンド・ツー・エンドで同時に学習される。

VGG-16 CNN アーキテクチャが選ばれたのは以下の理由からである。VGG-16 のアーキテクチャを選択した理由は，第一に，VGG-16 のアーキテクチャは深くて，高性能でありながら，扱いやすくて，拡張性があること。第二に，Russakovsky らは ImageNet 課題において VGG-16 モデルで素晴らしい成果を上げていること[11]。第三に VGG-16 の分類用の事前学習モデルが公開されていて，学習プロセスを加速することが可能であることが挙げられる。VGG-16 ネットワークは，AlexNet[12]などの従来のアーキテクチャよりもはるかに深く具体的には 13 畳み込み層と 3 FC 層で構成されている。または，コンボリューションカーネルとフィルタのストライドを 1 に設定した 3 つのフィルタによって特徴を付けられる。そのため，VGG-6 の各畳み込みフィルタは，よりシンプルな形状になっていますが，深さが増すことで，より複雑な関数を学習できる。

4.2.3 DRF のアルゴリズム

ディープリグレッションツリー

DRF は複数の深層回帰木を組み合わせたものである。各ツリーに対して，3 つの入出力ペア$\{x_i, y_i\}_{n=1}^{N}$があって，このうち$x_i \in R^{D_x}\ \ y_i \in R$。ディープリグレッションツリーモデルは，CNN の入力から出力へのマッピング関係をレッションツリーで接続することにより記述する。ディープリグレッションツリーT は多数のスプリットノード N とリーフノード L を持つ。具体的には，入力x_iを左右のどちらのサブツリーに至るかは，各スプリットノードで決定される。一方，葉ノード f はガウス分布 $p\,(yi\,)$で表され，μl が平均，σ_l^2が平均を表す。

スプリットノード

スプリットノードはスプリット関数$S_n\left(x_i;\Theta\right):x_i \to [0,1]$で表示される。スプリット関数は CNN のパラメータ Q でパラメータ化される，通常次のように定義される　$s_n\left(x_i;\Theta\right) = \sigma\left(f_{\varphi(n)}\left(x_i;\Theta\right)\right)$。ここで，$\sigma(\,)$はシグモイド関数，$\varphi(\,)$はスプリットノード n と一致する$f\left(x_i;\Theta\right)$の$\varphi$(n)個の要素を指し示すインデックス関数，$f\left(x_i;\Theta\right)$は学習した深層特徴量である。与えられた$x_i$に対して，葉ノード l に到達する確率は以下のように計算できる。

$$\boldsymbol{\omega_{\mathrm{l}}(x_i|\Theta) = [\prod_{n\in\mathcal{N}} s_n(x_i;\Theta)^{[\mathrm{l}\in\mathcal{L}_{n_\ell}]}\left(1 - s_n(x_i;\Theta)\right)^{[\mathrm{l}\in\mathcal{L}_{n_r}]}]}$$

ここで，$\boldsymbol{\mathcal{L}_{n_\ell}}$と$\boldsymbol{\mathcal{L}_{n_r}}$は，サブツリー$\boldsymbol{T_{n_\ell}}$と$\boldsymbol{T_{n_r}}$に属するリーフノードの集合である。サブツリー$\boldsymbol{T_{n_\ell}}$はツリーの根がノード n の左ツリー$\boldsymbol{n_\ell}$であることを意味し，$\boldsymbol{T_{n_\ell}}$は木の根がノード n の右ツリー $\boldsymbol{n_r}$であることを意味する．

スプリットノード

ツリー$\mathcal{T}$を考えると，各入力x_iに対して，$\mathrm{l} \in \mathcal{L}$ 個の葉ノードは y_i の予測分布を表し，$p_{\mathrm{l}}(y_i)$で示される。具体的には，そこでは$p_{\mathrm{l}}(y_i)$はガウス分布$\mathcal{N}(y_i|\mu_l, \sigma_l^2)$に従順であると仮定する。したがって，$x_i$に対する$y_i$の条件付き確率を持つ最終的な分布は，各リーフノードへの経路の確率を平均化することで次のように計算することができる。

$$p_{\mathcal{T}}\left(y_i\middle|x_i;\Theta,\pi\right) = [\sum_{\mathrm{l}\in\mathcal{L}} \omega_{\mathrm{l}}\left(x_i\middle|\Theta\right) p_{\mathrm{l}}(y_i)]$$

ここで，Θは CNN のパラメータ，πは分布のパラメータ$\{\mu_l, \sigma_l^2\}$ である。この分布は混合分布とみなすことができ，$\omega_{\mathrm{l}}\left(x_i\middle|\Theta\right)$は混合係数，$p_{\mathrm{l}}(y_i)$は$\mathrm{l}^t h$葉ノードにおけるガウス分布である$\pi$は木によって異なるため，以下では$\pi_k$をインデックスとして使用する。

ディープリグレッションフォレスト

ディープリグレッションフォレストは，複数のディープリグレッションツリー $\mathcal{F} = \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$ の組み合わせであり，入力x_iによる予測の最終出力分布は，全てのツリーの平均として計算することが可能である。

$$p_{\mathcal{F}}\left(y_i\middle|x_i,\Theta,\Pi\right) = \frac{1}{N}[\sum_{n=1}^{N} p_{\mathcal{T}_n}\left(y_i\middle|x_i,\Theta,\pi_n\right)]$$

ここで，Nはツリーの総数，$Pi = \{pi_1, \ldots, \pi_N\}$，$p_{\mathcal{F}}\left(y_i\middle|x_i,\Theta,\Pi\right)$ は i^{th} 入力からy_iを出力するときの可能性を表す。

最適化

トレーニングセット S が与えられると、ディープリグレッションツリーを学習するのは以下の損失関数を最小化する：

$$R(\boldsymbol{\pi}, \boldsymbol{\Theta}; \mathcal{S}) = -\frac{1}{N}\sum_{i=1}^{N}\log(p(\mathbf{y}_i|\mathbf{x}_i, \mathcal{T}))$$
$$= -\frac{1}{N}\sum_{i=1}^{N}\log\Big(\sum_{\ell\in\mathcal{L}} P(\ell|\mathbf{x}_i;\boldsymbol{\Theta})\pi_\ell(\mathbf{y}_i)\Big),$$

ここで、スプリットノードパラメータΘ（VGG16 のパラメータも含む）と葉ノードパラメータπをそれぞれ学習する必要がある。トレーニングとき、まずは、πを固定し、Θを最適化する、そしてΘを固定し、πを最適化する。それぞれの最適化を交代して、最終的にすべてのパラメータが収束する。

πを固定すると、損失関数の勾配は：

$$\frac{\partial R(\boldsymbol{\pi}, \boldsymbol{\Theta}; \mathcal{S})}{\partial \boldsymbol{\Theta}} = \sum_{i=1}^{N}\sum_{n\in\mathcal{N}} \frac{\partial R(\boldsymbol{\pi}, \boldsymbol{\Theta}; \mathcal{S})}{\partial f_{\varphi(n)}(\mathbf{x}_i;\boldsymbol{\Theta})} \frac{\partial f_{\varphi(n)}(\mathbf{x}_i;\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}$$

その中で：

$$\frac{\partial R(\boldsymbol{\pi}, \boldsymbol{\Theta}; \mathcal{S})}{\partial f_{\varphi(n)}(\mathbf{x}_i;\boldsymbol{\Theta})} = \frac{1}{N}\Big(s_n(\mathbf{x}_i;\boldsymbol{\Theta})\Gamma^i_{n_r} - \big(1-s_n(\mathbf{x}_i;\boldsymbol{\Theta})\big)\Gamma^i_{n_l}\Big)$$

一般的なスプリットノード n に対して：

$$\Gamma^i_n = \frac{p(\mathbf{y}_i|\mathbf{x}_i;\mathcal{T}_n)}{p(\mathbf{y}_i|\mathbf{x}_i;\mathcal{T})} = \frac{\sum_{\ell\in\mathcal{L}_n} P(\ell|\mathbf{x}_i;\boldsymbol{\Theta})\pi_\ell(\mathbf{y}_i)}{p(\mathbf{y}_i|\mathbf{x}_i;\mathcal{T})}$$

すべてのスプリットノードに対してΓ^i_nが計算できで、そして$\Gamma^i_n = \Gamma^i_{n_l} + \Gamma^i_{n_r}$，葉ノードから根ノードまですべてのノードがボトムアップ方式で計算できる。なので、パラメータΘが誤差逆伝播法で学習できる。

Θを固定すると、損失関数は制約付き最適化問題になる：

$$\min_{\boldsymbol{\pi}} R(\boldsymbol{\pi}, \boldsymbol{\Theta}; \mathcal{S}), \mathbf{s.t.}, \forall\ell, \int \pi_\ell(\mathbf{y})d\mathbf{y} = 1$$

簡単に計算できるため、葉ノードにおける確率分布をガウス分布で仮定する：

$$\pi_\ell(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^k \det(\boldsymbol{\Sigma}_\ell)}} \exp(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_\ell)^{\mathrm{T}}\Sigma_\ell^{-1}(\mathbf{y}-\boldsymbol{\mu}_\ell))$$

ここでμ_lはガウス分布の平均値、Σ_lはガウス分布の共分散行列。

この制約付き最適化問題に対して、Variational Bounding[22,23]方法を利用して、損失関数は以下と表示できる：

$$
\begin{aligned}
R(\boldsymbol{\pi}, \boldsymbol{\Theta}; \mathcal{S}) &= -\frac{1}{N}\sum_{i=1}^{N} \log\Big(\sum_{\ell\in\mathcal{L}} P(\ell|\mathbf{x}_i;\boldsymbol{\Theta})\pi_\ell(\mathbf{y}_i)\Big) \\
&= -\frac{1}{N}\sum_{i=1}^{N} \log\Big(\sum_{\ell\in\mathcal{L}} \zeta_\ell(\bar{\boldsymbol{\pi}};\mathbf{x}_i,\mathbf{y}_i)\frac{P(\ell|\mathbf{x}_i;\boldsymbol{\Theta})\pi_\ell(\mathbf{y}_i)}{\zeta_\ell(\bar{\boldsymbol{\pi}};\mathbf{x}_i,\mathbf{y}_i)}\Big) \\
&\le -\frac{1}{N}\sum_{i=1}^{N}\sum_{\ell\in\mathcal{L}} \zeta_\ell(\bar{\boldsymbol{\pi}};\mathbf{x}_i,\mathbf{y}_i)\log\Big(\frac{P(\ell|\mathbf{x}_i;\boldsymbol{\Theta})\pi_\ell(\mathbf{y}_i)}{\zeta_\ell(\bar{\boldsymbol{\pi}};\mathbf{x}_i,\mathbf{y}_i)}\Big) \\
&= R(\bar{\boldsymbol{\pi}}, \boldsymbol{\Theta}; \mathcal{S}) - \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell\in\mathcal{L}} \zeta_\ell(\bar{\boldsymbol{\pi}};\mathbf{x}_i,\mathbf{y}_i)\log\Big(\frac{\pi_\ell(\mathbf{y}_i)}{\bar{\pi}_\ell(\mathbf{y}_i)}\Big),
\end{aligned}
$$

ここで上界$\phi(\pi,\bar{\pi})$が得られる：

$$
\phi(\boldsymbol{\pi}, \bar{\boldsymbol{\pi}}) = R(\bar{\boldsymbol{\pi}}, \boldsymbol{\Theta}; \mathcal{S}) - \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell\in\mathcal{L}} \zeta_\ell(\bar{\boldsymbol{\pi}};\mathbf{x}_i,\mathbf{y}_i)\log\Big(\frac{\pi_\ell(\mathbf{y}_i)}{\bar{\pi}_\ell(\mathbf{y}_i)}\Big)
$$

最終的にμとΣは以下の式より更新できる：

$$
\boldsymbol{\mu}_\ell^{(t+1)} = \frac{\sum_{i=1}^{N} \zeta_\ell(\boldsymbol{\pi}^{(t)};\mathbf{x}_i,\mathbf{y}_i)\mathbf{y}_i}{\sum_{i=1}^{N} \zeta_\ell(\boldsymbol{\pi}^{(t)};\mathbf{x}_i,\mathbf{y}_i)}
$$

$$
\boldsymbol{\Sigma}_\ell^{(t+1)} = \frac{\sum_{i=1}^{N} \zeta_\ell(\boldsymbol{\pi}^{(t)};\mathbf{x}_i,\mathbf{y}_i)(\mathbf{y}_i-\boldsymbol{\mu}_\ell^{(t+1)})(\mathbf{y}_i-\boldsymbol{\mu}_\ell^{(t+1)})^{\mathrm{T}}}{\sum_{i=1}^{N} \zeta_\ell(\boldsymbol{\pi}^{(t)};\mathbf{x}_i,\mathbf{y}_i)}
$$

4.2.4 頭部姿勢推定

コンボリューショナルネットワークを用いた頭部姿勢の予測に関する多くの研究において，最も簡単な方法は，平均二乗誤差損失を用い，頭部姿勢の出力角度を直接回帰する方法である。しかし，この方法では，年齢推定に利用したいデータセットで十分な性能を満たすことができない。

そこで，Ruiz の方法[5]を採用し，深層マルチロス CNN を学習させ，満足のいく精度で頭部姿勢推定を行うこととした。また，ResNet50 net- works [13]を頭部姿勢推定用に導入し，3 つの角度に対して別々に 3 つの損失を用いる。各損失には，直接回帰された平均二乗誤差と，姿勢の分類によるクロスエントロピーの損失の 2 つがある。FC 層は 3 つの角度で使用され，これまでのネットワークと共通である。また，分類によるクロスエントロピロスを追加で採用することで，バックプロパゲートする信号を 3 つ構築し，学習過程を改善した。3 つの出力角度の予測値を計算し，最終的な頭部姿勢の結果とした。図 4.3 はその詳細である。

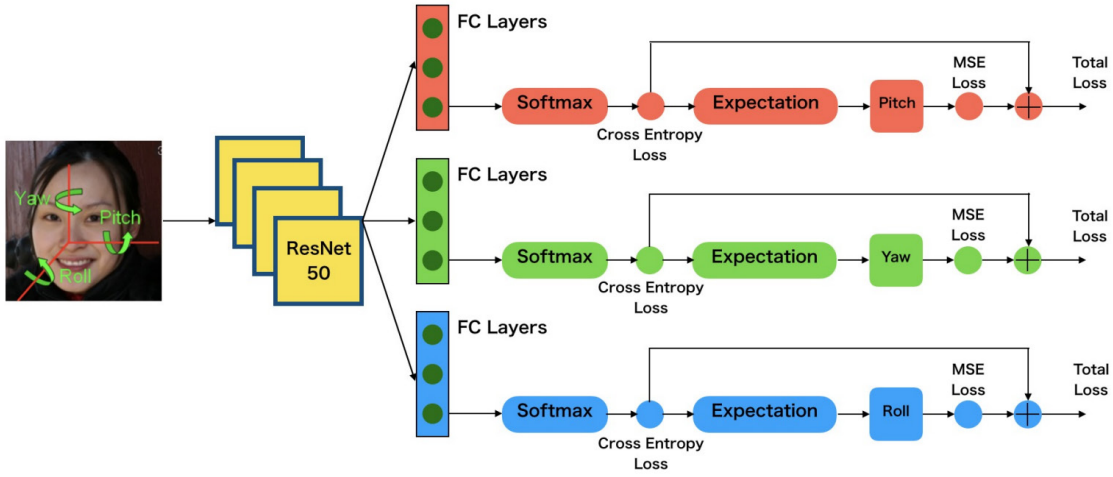

図 4.3 平均二乗誤差とクロスエントロピーの損失を合わせた頭部姿勢推定モデル

4.3 年齢推定実験パラメータ

4.3.1 データセット

本章では，年齢推定学習用に異なる人種からなる第三章も紹介された 2 つのデータセット CACD [7] と AFAD[8]を使用した。他には頭部姿勢推定用にのデータセットも一つ利用した。図 4.4 に各データセットの年齢推定用レプリカ画像を示す。

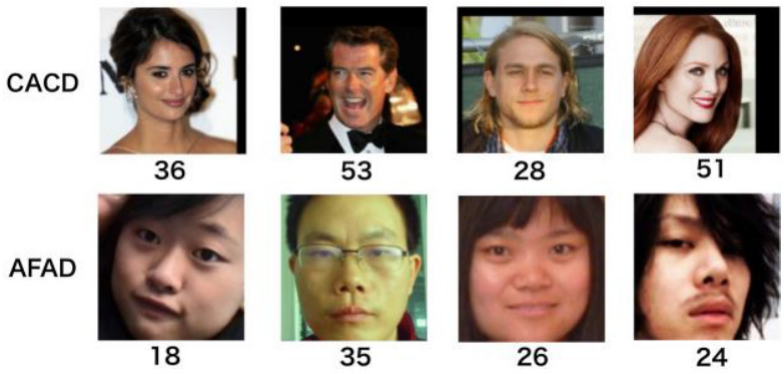

図 4.4 CACD と AFAD の顔画像の例，各画像の下にある数字は被写人物の実年齢

頭部姿勢推定は300W-LP [6]データセットを利用した。300W-LPは，AFW[14]，LFPW[15]，HELEN[16]，IBUG[17]，XM2VTS[18]を含む複数のアライメントデータベースを標準化するために68ランドマークを使用したものである。300Wに顔プロファイリング法を適用し，大きなポーズで 61,225 個のサンプル（IBUG 1,786 個，AFW 5,207 個，LFPW 16,556 個，HE- LEN 37,676 個，XM2VTS使用されてない）をフリップして122,450サンプルに拡張した。このようにして得られたデータベースを300W across Large Poses (300W-LP)と呼ぶ。

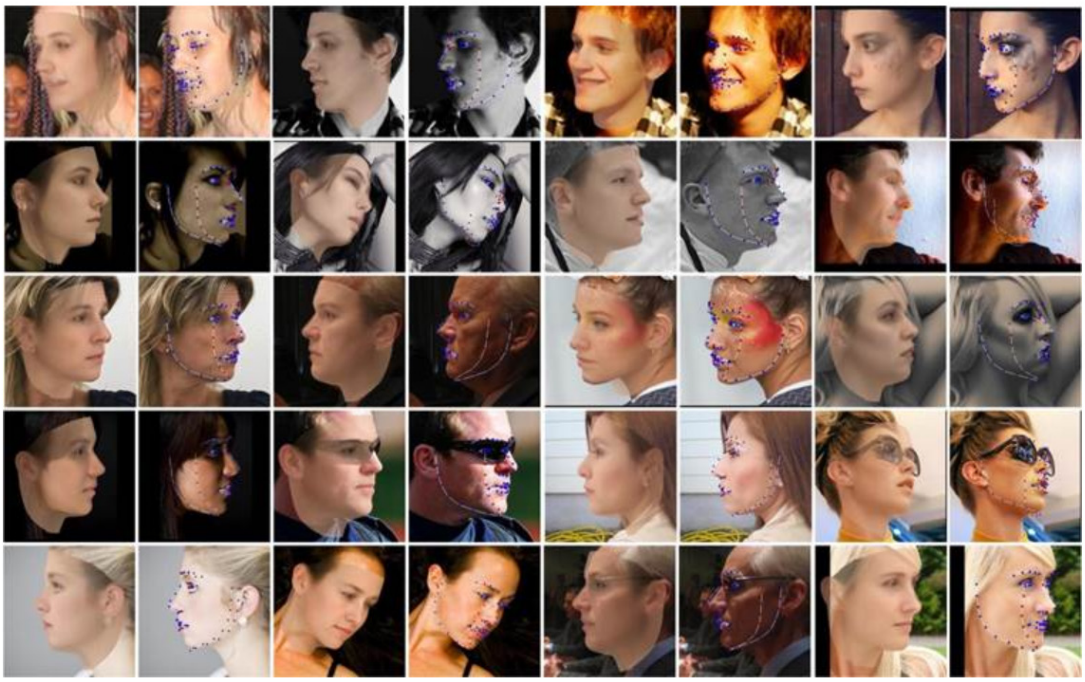

図 4.5　300W-LP データセット

4.3.2 実装の詳細

各実験では，ImageNet から VGG16 の既存の重みを初期値として使用した。ニューラルネットワークの学習パラメータは，学習データのバッチサイズを 64，ドロップアウト層の比率を 0.5，勾配降下法として SGD（stochastic gradient descent）を用い，学習率を初期値として 0.2，5k 反復で半分に減衰させる，となっている。リグレッションフォレストの学習パラメータは，ツリーの本数を 4，各ツリーの深さを 5，出力ユニットの数を 64，

葉ノードの値を 10 反復ごとに更新，葉ノードからの予測結果を 30 反復ごとに更新とした。このモデルを CACD と AFAD で微調整し，年齢推定を行う。ResNet50 は，300W-LP データセットを用いて，頭部姿勢推定を行う。

ResNet50 は，Adam 最適化を勾配降下法として用い，学習率を10^{-5}とし，$\beta 1 = 0.9$, $\beta 2 = 0.999$, E =10^{-8} とした。学習段階では，学習データは以下のように分割されます。80%が学習用，20%が検証用である。検証セットでモデルがオーバーフィットした場合，学習プロセスは早期に中断される。モデルは Nvidia GTX 1080 GPU で学習された。

4.4 実験結果

まず，頭部姿勢推定のために，300W-LP データセットで multi-loss CNN を学習させた。その後，推定された頭部姿勢角に基づいて画像を AFAD と CACD に分割し，正面と非正面の画像のサブセットに対して別々に DRF を学習させた。次は、様々な閾値を設定し，異なる角度によって正面画像のサブセットをトレーニングして，最適な閾値を決める。そして，2 つの顔映像データセットでモデルをテストし，同一人物の顔の角度を変えて年齢を推定し，その結果を従来の方法と比較した。公平な比較を行うために，同じネットワーク構造と学習ストラテジーを使用した。

4.4.1 頭部姿勢推定のテスト

頭部姿勢推定を行うため，採用したマルチロス CNN を 300W-LP のデータセットでトレーニングした。また，300W-LP のサブセットである AFLW2000 データセットでは，顔領域が小さく切り取られた画像で構成されている。姿勢推定アルゴリズムは主に AFLW2000 で性能をテストする。AFLW2000 のデータセットには，グラウンドトゥルースランドマークが記録されているため，本手法と，一般的に用いられている FAN[19]や Dlib[20]顔検測手法との比較を行った。定量的な結果は表 4.1 の通りである。本手法は従来の顔検出アルゴリズよりも優れており，提案したシステムには適している。

手法	Yaw	Pitch	Roll	平均
Dlib[20]	23.153	13.633	10.545	15.777
Fan[19]	6.358	12.277	8.714	9.116
Multiloss CNN	**6.470**	**6.559**	**5.436**	**6.155**
Ground truth landmarks	5.924	11.756	8.271	8.651

表 4.1 各手法における平均誤差（オイラー角）

4.4.2 AFAD と CACD のテスト

本章では，前頭部と非前頭部の顔画像に基づく年齢推定に対する DRF の性能を示す。この実験では，頻繁に使用される AFAD と CACD のデータセットを使用した。AFAD データセットほとんどはアジア人の画像を含めて，CACD は大分ヨーロッパ人の画像を含めている。両データセットの頭部ポーズを推定するために，学習した multi-loss CNN を使用した。各顔画像に対して，各軸に 1 つずつ，計 3 つの回転角度を推定した。このとき，3 つの角度の和の閾値を 30 度とし，頭部姿勢角度の推定値の和が 30 度以上の画像を非正面画像と定義した。図 4.6 に，各データセットの非正面顔画像の模範画像を示す。

推定された角度に基づき，AFAD は 53,983 枚と 5,361 枚の画像からなる正面サブセットと非正面サブセットに分割された。両サブセットともを学習/テスト(85%/15%)セットに分割し，学習プロセスを異なるランダムな分離で 5 回繰り返し，最終結果は 5 回の出力の平均とした。定量的な結果は表 4.2 にまとめられている。この結果から，正面顔画像からの年齢推定精度は，非正面顔画像からの推定精度よりも有意に優れていることがわかる。

推定角度に基づき，CACD は 15,145 枚の画像からなる正面サブセットと 3,026 枚の画像からなる非正面サブセットに分割された。両サブセットを訓練セットとテストセット（85%/15%）に分割し，異なるランダムな分離で 5 回訓練プロセスを繰り返し，最終結果は 5 回の出力の平均である。定量的な結果は表 4.3 にまとめられている。この結果から，正面顔画像からの年齢推定精度は，非正面顔画像からの推定精度よりも有意に優れていることがわかる。

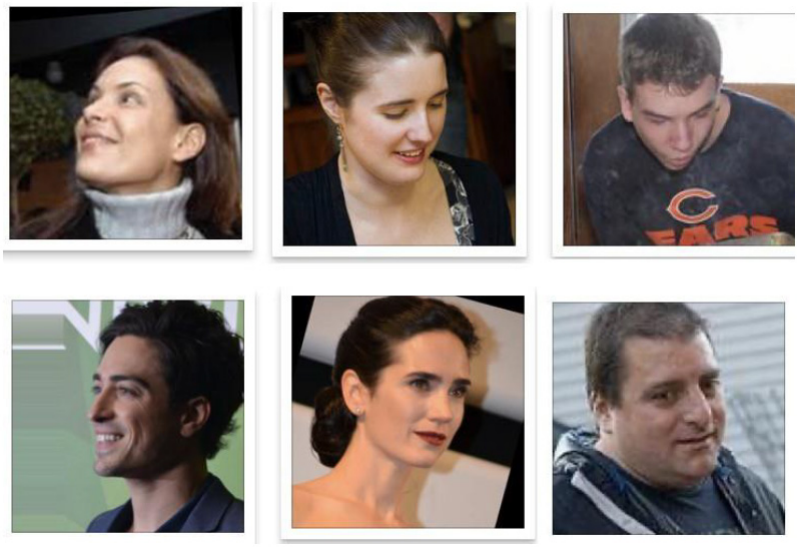

図 4.6　非正面の顔画像

サブセット	MAE
正面	3.73
非正面	4.97

表 4.2　AFAD の正面と非正面のサブセットでの性能（MAE）比較

サブセット	MAE
正面	4.59
非正面	5.65

表 4.3　CACD の正面と非正面のサブセットでの性能（MAE）比較

4.4.3 閾値確定

本章では, CACD データセットと AFAD データセットを, 頭部姿勢角度の閾値を変えて, いくつかのサブセットに分割する。また, 閾値は 50 度から 10 度まで 10 度刻みで設定し, 閾値が厳しくなるにつれて, 頭部姿勢角度が閾値内にあるサンプル数が少なくなった。閾値とサンプル数の対応、および年齢推定の性能は表 4.4 のようにまとめられた。閾値が 30 度より小さい場合、サンプル数は急激に減少するが、年齢推定の性能はほとんど変化しない。したがって、サンプル数と性能のトレードオフを考慮し、30 度を閾値として選択した。

閾値（度）	AFAD		CACD	
	MAE	画像枚数	MAE	画像枚数
50	3.97	59173	4.87	18023
40	3.84	57232	4.70	16842
30	3.73	53983	4.59	15145
20	3.72	36748	4.58	10398
10	3.71	18753	4.58	7569

表 4.4　CACD と AFAD において、異なる閾値による正面画像サブセットでの性能（MAE）と画像枚数の比較

4.4.4 顔動画データセットでのテスト

年齢推定性能の観点から提案したモデルを評価するために，2 つの新しい顔映像データセットを構築した。アジア人とヨーロッパ人の 12 分間の顔を含めて映像から，それぞれ 18,282 フレームと 18,944 フレームを収集した。ただし，これらのデータセットは同一人物から収集したものであり，年齢推定モデルを評価するためにのみ使用した。まず，アジア人およびヨーロッパ人を表す AFAD と CACD を用いて DRF を学習させた。次に，この 2 つの学習済みモデルを，頭部ポーズを同時に計測できる顔画像データセットでテストした。テスト画像の例を図 4.7 に示す．頭部姿勢が 30 度以内の顔のみ年齢推定を行い，頭部姿勢の制限のない全画像の結果と比較した。また，AFAD と CACD で学習した他のモデルについても，顔画像データセットでテストし，より全面的に比較した。

アジアの顔映像データセットで学習させた。また，DRF を AFAD 上で動作させ，頭部ポーズ制限のあるアジア映像データセットでテストした。本手法と他の年齢推定モデルの結果を比較し，定量的な結果を表 4.5 にまとめた．すべてのモデルは，公平に比較できるように，同じ学習ストラテジーを用いて AFAD で学習させた。顔画像から年齢を推定するタスクにおいて，本手法は MAE 5.12 を達成し，分散は既存の最良手法と比較して 0.62 減少している．

手法	MAE	分散
AlexNet [12]	6.19	6.92
DEX [21]	6.72	8.65
DRF [4]	5.96	4.12
Our method	**5.12**	**3.50**

表 4.5　アジアデータセットにおける性能(MAE，分散)の比較

欧州の顔映像データセットについて CACD で DRF を学習させ，頭部姿勢制限のある欧州の映像データセットでモデルをテストした。他の年齢推定モデルと比較し，その定量的な結果を表 4.6 にまとめた。すべてのモデルは，公平な比較を可能にするために，同じ学習ストラテジーを用いて CACD 上で学習された。顔画像から年齢を推定するタスクにおいて，本手法は MAE 5.56 を達成し，分散は既存の最適な手法と比較して 1.53 減少していることが確認された。

手法	MAE	分散
AlexNet [12]	6.93	7.15

DEX [21]	7.17	8.22
DRF [4]	6.39	5.84
Our method	**5.56**	**4.31**

表 4.6　欧州データセットにおける性能(MAE，分散)の比較

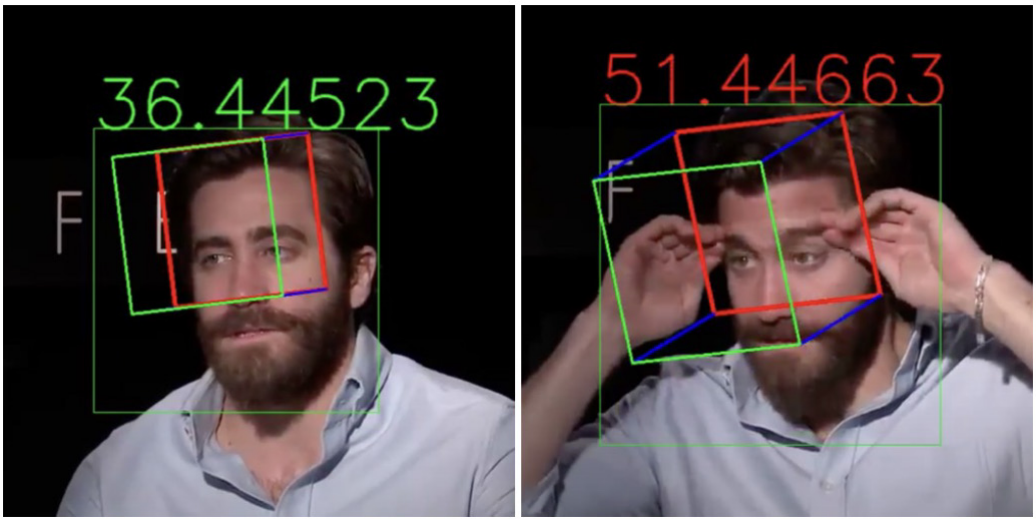

図 4.7　年齢と頭部ポーズを推定した顔動画データセットの例。数字は予測された年齢を表す。緑色は頭部姿勢の回転角度の総和が 30 度未満，赤色は 30 度以上であることを示す。

4.5 結論

本章では，ビデオやウェブカム中の顔画像から年齢を推定する際に，頭部ポーズが異なると推定年齢に耐え難い誤差が生じるという問題を解決するために，年齢推定と頭部ポーズ推定を組み合わせたシステムを提案する。本手法では，頭部ポーズを制限し，頭部ポーズが指定された閾値内にある顔画像に対してのみ年齢推定を行うことにより，動画像からの年齢推定において精度と安定性の大幅な向上を実現した。

本論文の主な貢献は以下の通りである。 (1)本研究は初めて、ビデオにおける年齢推定のために，年齢推定と頭部ポーズ推定を組み合わせた手法を提案した。(2) 本手法は，顔動画データセットにおける年齢推定において，他の最先端手法と比較して，精度・分散の両面で有意に高い性能を示す。

参考文献

1. B. Chen, C. Chen, and W. H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. IEEE Trans. Multimedia, 17(6):804–815, 2015.
2. K. Ricanek and T. Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In Proc. FG, pages 341–345, 2006.
3. Simonyan. K, Zisserman. A. Very deep convolutional networks for largescale image recognition. International Conference on Learning Representations, 2015.
4. W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille. Deep Regression Forests for Age Estimation. In IEEE CVPR, pages 2304–2313, 2018.
5. Ruiz. N, Chong. E, Rehg. J.M. Fine-grained head pose estimation without key-points. In CVPR workshops, pp. 2074–2083, 2018.
6. X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 146–155, 2016.
7. B. Chen, C. Chen, and W. H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. IEEE Trans. Multimedia, 17(6):804–815, 2015.
8. Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4920–4928, 2016.

9. Zhang, K., Zhang, Z., Li, Z., and Qiao, Y.. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503. 2016.
10. Simonyan. K, Zisserman. A. Very deep convolutional networks for large-scale image recognition. 444 CoRR abs/1409.1556, 2014.
11. Russakovsky. O, Deng. J, Su. H, Krause. J, Satheesh. S, Ma. S, Huang. Z, Karpathy. A, Khosla. A, Bernstein. M, Berg. AC, Fei-Fei. L. Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV), 115(3):211–252, 2015.
12. Krizhevsky. A, Sutskever. I, Hinton. GE. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (NIPS), 2012.
13. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.
14. X. Zhu, and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In IEEE CVPR, pages 2879–2886, 2012.
15. P.N.Belhumeur, D.W.Jacobs, D.Kriegman, and N.Kumar. Localizing parts of faces using a consensus of exemplars. In IEEE CVPR, pages 545–552, 2011.
16. F. Wang, H. Han, S. Shan, and X. Chen. Multi-task learning for joint prediction of heterogeneous face attributes. In IEEE FG, pages 173–179, 2017.
17. C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In IEEE ICCVW, pages 397– 403, 2013.
18. K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In Second international conference on audio and video-based biometric person authentication, volume 964, pages 965–966, 1999.
19. A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In International Conference on Computer Vision, pages 1021-1030, 2017
20. T. -Y. Lin, P. Dolla r, R. B. Girshick, K. He , B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In IEEE CVPR, volume 1, page 4, 2017.
21. R. Rothe, R. Timofte, and L. V. Gool. Deep expectation of real and apparent age from a single image without facial landmarks. Int. J. Comput. Vis., 126(2):1–14, Aug. 2016.
22. M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. Machine Learning, 37(2):183–233, 1999
23. G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In Proc. CVPR Workshops, pages 34–42, 2015.

第 5 章 Face Landmark を用いたマルチタスク学習法

5.1 緒論

近年，深層学習は年齢推定[1,2,3]を含む様々なコンピュータビジョンのタスクで目覚ましい成果を上げているが，人間の顔からの実年齢判定は依然として非常に難しい問題である。

Yoo ら[4]の研究は，これらの限界を克服するために行われたもので，年齢と性別推論を伴う条件付き年齢推定を改良することが行われた。しかし，年齢と性別の関係は特別なものではないため，精度を十分に向上させることができない。そこで，より年齢との関連性が高いものとして，顔面ランドマークが考えられます。顔面ランドマークとは，鼻の先や目の中心など，顔面上で特定される重要な点のことです。従来の手法では，顔画像から年齢と顔ランドマークを別々に推定する方法が研究されてきた。

本研究では，年齢推定と顔面ランドマーク推定を組み合わせたシステムを提案する。本システムでは，マルチタスク学習を用いて年齢推定と顔ランドマーク推定を行うことで，顔画像に対してより高い精度を実現する。具体的には，2 つの部分から構成される。まず，VGG16[5]ネットワークを用いてモデルを作成し，同時に年齢と Face Landmark を学習する。そして，マルチタスク学習を実現するために，CNN 出力ニューロンからパラメータをそろぞれのディープリグレッションフォレスト（DRFs）[2]とマッチングして年齢と Face Landmark を各自学習して、最終的に予測する。Cross-Age Celebrity Dataset (CACD) [6] と UTKFace データセット[7]での実験結果は，提案するシステムが従来の方法よりも年齢推定の精度を大幅に向上させ，CACD と UTKFace データセットにおいて最先端の性能を達成することを実証していた。

5.2 提案手法

5.2.1 Face Landmark

目頭や目尻，輪郭，唇の両端など各パーツの目印となるポイントが「Face Landmark」として認識され，それぞれのポイントに数字と座標を付けられた。これで，顔が数値データとなり，これらの点の距離や位置がその人の顔のデータとなる。

顔のランドマークの位置決め，または顔の位置合わせは，人間の顔上の事前定義されたランドマークのセットを検出することを意味する。顔認証・認識[8]，表情認識[9]，顔属性分析[10]など，顔に関連する多くのアプリケーションの基本ステップである。

本章はこの Face Landmark を利用して、大量の顔特徴量を抽出し、メインタスクであり年齢推定も共用して、より高い精度を達成する。

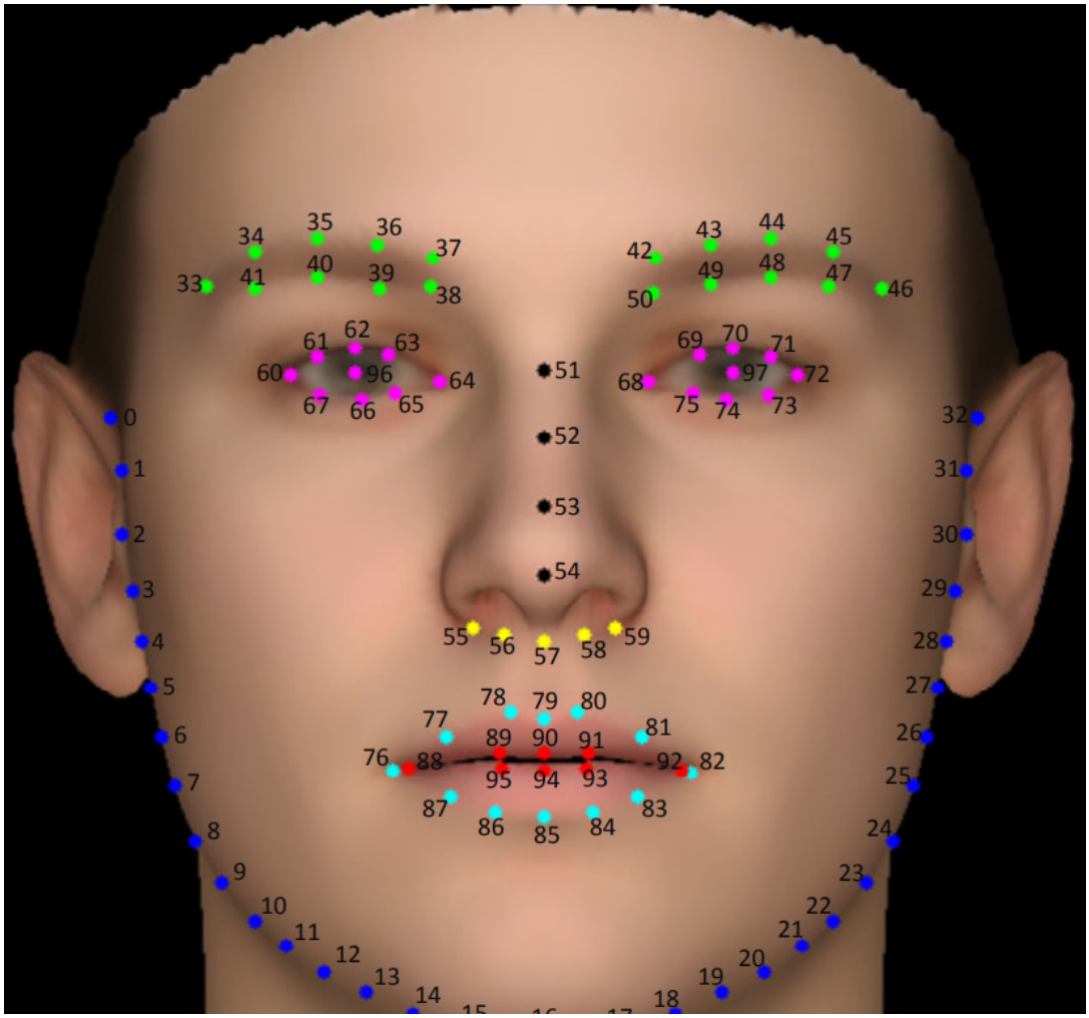

図 5.1 Face Landmark

5.2.2　Face Landmark 検出

Face landmark を推定するために，アンカーCNN[11]を導入する。このネットワークは分割と集約の 2 つのステップから構成される。図 5.2 にはパイプラインの概要を示す。

アンカーベースの設計に基づき，分割統治法を採用する。アンカーテンプレートを用いて顔特徴空間を分割し，各アンカーが回帰としてそれぞれの特徴量を抽出する。これにより，各アンカーの特徴空間からそれぞれの予測結果と予測信頼度によって重み付けして最終結果集計する。本研究は利用されたデータセットに対して，トレーニングしたアンカーCNN を利用して、Face landmark を推定する。推定された Face landmark はラベルとして、マルチタスク学習にトレーニングするときグラウンドトゥルースとして利用される。

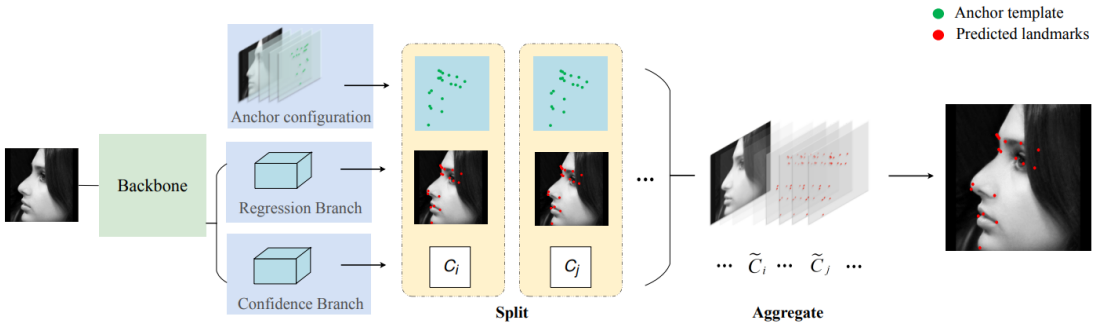

図 5.2　Face Landmark 検出手法

5.2.3　マルチタスク学習

マルチタスク学習の定義：共有表現に基づく機械学習アプローチで，関連する複数のタスクを一緒に学習する。

マルチタスク学習は，メインタスクが関連タスクの学習信号が持つドメイン固有の情報を利用する推論型移動学習法である。 関連タスクの学習信号が持つドメイン固有の情報を帰納的バイアスとして利用し，メインタスクの汎化性能を向上させる機械学習法である。 マルチタスク学習では，関連する複数のタスクが並行して学習し，同時に勾配をバックプロパゲートし，複数のタスクが基礎となる共有表現を通じて互いに学習し合うこと

で，汎化性能を向上させる。簡単に言えば，マルチタスクは複数の関連するタスクを一緒に学習するものであり（関連するタスクでなければならないことに注意，関連タスクの定義と共有する情報は後述），浅いレベルでの共有表現によって学習プロセスが促進されるということである。お互いの領域情報を共有・補完し，お互いの学習を促進し，汎化の効果を高めるために，表面的なレベルでの表現を共有することで学習プロセスを促進する。

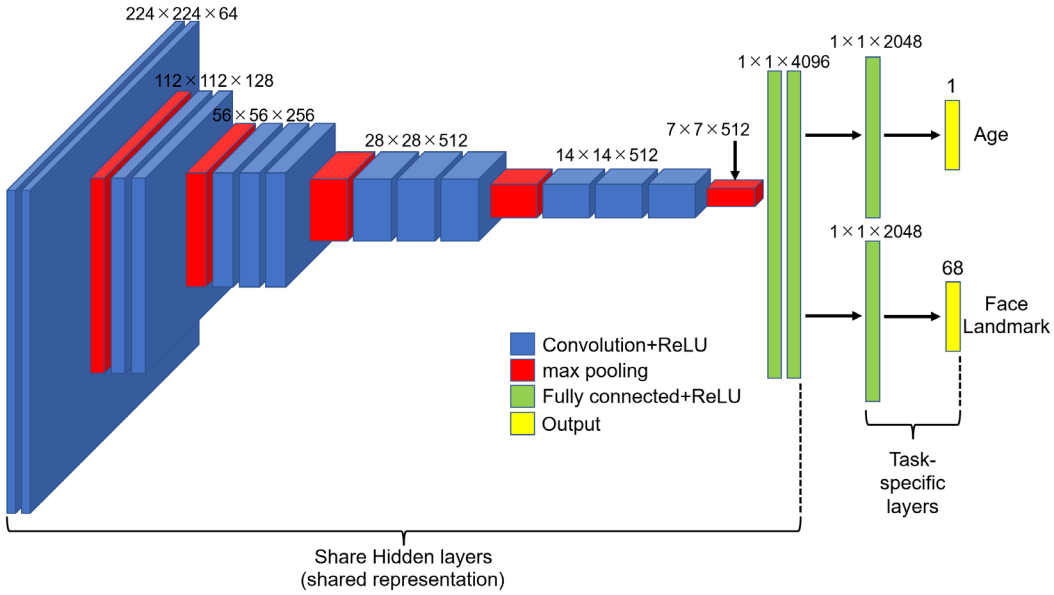

図 5.3　マルチタスク学習

5.2.4　Face Landmark を用いたマルチタスク学習

本システムでは，マルチタスク学習を用いて年齢推定と顔ランドマーク推定を行うことで，顔画像に対してより高い精度を実現する。具体的には，2 つの部分から構成される（図 5.3）。まず，VGG16[5]ネットワークを用いてモデルを作成し，同時に年齢と Face Landmark を学習する。そして，マルチタスク学習を実現するために，CNN 出力ニューロンからパラメータをそろぞれのディープリグレッションフォレスト（DRFs）[2]とマッチングして年齢と Face Landmark を各自学習して、最終的に予測する。DRF の構造やトレーニング方法は第 4 章に説明された。

マルチタスク学習に関して流れは図 5.4 に示した。損失関数は二つのタスクの総和となる：

$$L_{age} = \frac{1}{N}\sum_{i=1}^{N}(\tilde{a}_{ij} - a_{ij})^2$$

$$L_{landmark} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{68}[(\tilde{x}_{ij} - x_{ij})^2 + (\tilde{y}_{ij} - y_{ij})^2]$$

$$L_{multi} = \omega \cdot L_{age} + (1 - \omega) \cdot L_{landmark}$$

ここで、平均二乗誤差（MSE）を利用された。

年齢と Face Landmark 両方も収束できるため、二つの手法を利用された。まずは両方の勾配をバランスするため、重み係数$\omega = 0.9$を設定する。そして、学習において、補助タスクである Face Landmark 推定が学習セットにオーバーフィッティングし始め、主タスクに害を及ぼす前に、早期停止を行った。これは、学習の初期には、ネットワークが悪いローカルミニマムに陥るのを避けるために、すべてのタスクによって同時に学習し、結果によるパラメータを更新するが，学習が進むにつれて、補助タスクはピーク性能に達した後、もはや主タスクにとって有益ではなくなるので、その学習過程を停止させるべきである。

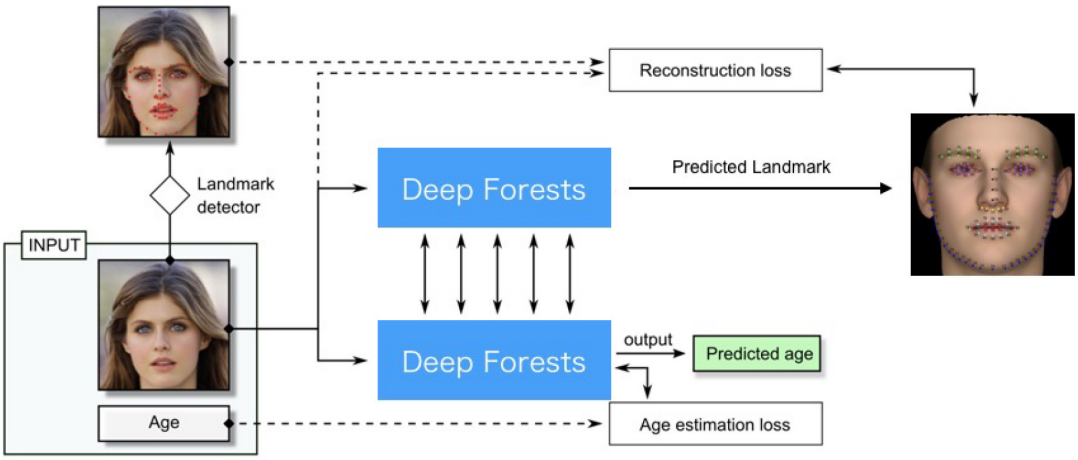

図 5.4 提案したマルチタスク学習の流れ

5.3 年齢推定実験パラメータ

5.3.1 データセット

本章では，年齢と Face Landmark ラベルを持つ 2 つのデータセットを使用した。

CACD [6]は第 3 章に紹介された。

UTKFaceデータセット（図5.5）は，年齢範囲が広く（0歳から116歳），年齢，性別，人種のアノテーションを持つ2万以上の顔画像を含む大規模な顔データセットである。このデータセットの画像は，既に位置合わせやトリミングが行われた状態で利用されている。また，このデータセットには68点の顔ランドマーク情報が含まれているため，本研究に容易に利用できます。評価用データセットは，85%をトレーニング用，15%をテスト用に分割された。

図 5.5　UTKFace データセット

5.3.2 実装の詳細

各実験では，Imdb-wiki からトレーニングした VGG16 の既存の重みを初期値として使用した。ニューラルネットワークの学習パラメータは，学習データのバッチサイズを 64，ドロップアウト層の比率を 0.5，勾配降下法として SGD（stochastic gradient descent）を用い，学習率を初期値として 0.2，5k イテレーションごとにで半分に減衰させる。リグレッションフォレストの学習パラメータは，木の本数を 4，各木の深さを 5，出力ユニットの数を 64，葉ノードの値を 10 イテレーションごとに更新，葉ノードからの予測結果を 30

イテレーションごとに更新とした。このモデルを CACD と UTKFace で微調整し，年齢推定を行う。。

学習段階では，学習データは以下のように分割されます。80%が学習用，20%が検証用である。検証セットでモデルがオーバーフィットした場合，学習プロセスは早期に中断される。モデルは Nvidia GTX 1080 GPU で学習された。

5.4 実験結果

まず，CACD と UTKFace のデータセットにおいて，年齢を直接推定する DRF 構造の性能を他の研究者と比較した。さらに，提案するマルチタスク学習モデルの性能を CACD と UTKFace のデータセットにおいて，DRF や他の研究成果と比較した。異なるデータセット間の比較が公平に行われるように，同じモデル構造とパラメータを用いた。年齢推定の性能の基準として MAE を用いた。

5.4.1 DRF でのテスト

ここでは，Face Landmark なしで直接年齢推定を行う DRFs 手法のみを用い，2 つのデータセットで実験結果を示す。

UTKFace 上 DRFs 法を用いて MAE が 3.96 を達成した。UTKFace をランダムに 2 分割し，85%を学習用，15%をテスト用とした。学習はランダムに分割して 5 回繰り返し，最終結果は 5 回の出力の平均である。定量的な結果は表 5. 1 にまとめられている。DRFs 法は UTKFace の年齢推定に最も良い性能を示した。UTKFace において年齢推定の研究が少ないかつ古いため、そのまま VGG-16 ネットワークでより良い結果ができたが、精度を改善する余地がかなりある。

手法	MAE
CORAL [12]	5.39
BIFS + OHRank [13]	4.55
GAP + VGG [14]	4.87
DRFs	**3.96**

表 5.1　UTKFace における性能(MAE)の比較

CACD上DRFs法を用いてMAEが4.67を達成した。CACDをランダムに2分割し，85%を学習用，15%をテスト用とした。学習はランダムに分割して5回繰り返し，最終結果は5回の出力の平均である。定量的な結果は表5.2にまとめられている。

手法	MAE
DEX [15]	4.79
dLDLF [16]	4.73
RNDF [17]	4.60
DRFs	**4.67**
CR-MTk [18]	4.58

表 5.2　CACD における性能(MAE)の比較

5.4.2 Face Landmark を用いたマルチタスク学習

提案モデルは，マルチタスクランドマーク学習のために Face Landmark ラベルも提供する，使用頻度の高い UTK- Face と CACD データセットで学習した。その結果，年齢推定性能が大幅に向上することが確認された.

UTKFace の学習データを用いて，提案したマルチタスク学習法は，MAE3.41 を達成した。マルチタスク学習とシングルタスク学習の比較が公平に行われるように，同じ DRF の構造とパラメータを用いた。定量的な結果は表 5.3 にまとめられている。提案するマルチタスク学習法は，以下のように結果を向上させる。同じく最先端手法である単一タスク学習と比較して精度が 0.55 年上がった。

手法	MAE
CORAL [12]	5.39
BIFS + OHRank [13]	4.55
GAP + VGG [14]	4.87
DRFs	**3.96**
DRFs(マルチタスク学習)	**3.41**

表 5.3　 UTKFace における性能(MAE)の比較

提案したマルチタスク学習法は，CACDの学習データでMAE4.23を達成した。マルチタスク学習とシングルタスク学習の比較が公平に行われるように，同じDRFの構造とパラメータを用いた。定量的な結果は表5.4にまとめられている。提案するマルチタスク学習法は，以下のように結果を向上させる。 単一タスク学習に対して0.44年，最先端手法に対して0.35年精度が上がった。

手法	MAE
DEX [15]	4.79
dLDLF [16]	4.73
RNDF [17]	4.60
DRFs	**4.67**
CR-MTk [18]	4.58
DRFs(マルチタスク学習)	**4.23**

表 5.4　CACD における性能(MAE)の比較

5.5 結論

従来の年齢特徴量学習では，1 種類の年齢特徴量を学習し，人種など他の外見的特徴を無視して年齢推定を行う手法が主流であった。マルチタスク学習の中には，性別など他の特徴を用いて年齢を推定する学習性能を向上させた手法もある。しかし，年齢と性別などの他の特徴量との相関が弱いため，その精度はまだ十分ではありません。この問題を解決するため，本章は Face Landmark を補助タスクとして用いる新しい年齢推定用マルチタスク学習法を提案し，頻繁に用いられる CACD および UTKFace データセットにおいて最新の結果を実証した。

本章の主な貢献は以下の通りである。(1) マルチタスク学習を用いた年齢推定と Face Landmark の組合せを世界で初めて提案した．(2) 本論文で提案する年齢推定のためのマルチタスク学習モデルは，CACD データセットにおいて MAE 4.23，UTKFace において MAE 3.41 という最高精度を達成した。

参考文献

1. S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao. Using ranking-CNN for age estimation. In IEEE ICCV, pages 5183–5192, 2017.
2. W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille. Deep Regression Forests for Age Estimation. In IEEE CVPR, pages 2304–2313, 2018.
3. Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4920–4928, 2016.
4. B. Yoo, Y. Kwak, Y. Kim, C. Choi and J. Kim. Deep Facial Age Estimation Using Conditional Multitask Learning With Weak Label Expansion. In IEEE Signal Processing Letters, 25(6):808–812, 2020.
5. Simonyan. K, Zisserman. A. Very deep convolutional networks for largescale image recognition. International Conference on Learning Representations, 2015.
6. B. Chen, C. Chen, and W. H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. IEEE Trans. Multimedia, 17(6):804–815, 2015.
7. Z. Zhang, Y. Song and H. Qi. Age Progression/Regression by Conditional Adversarial Autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4352–4360, 2017.
8. WANG, Lezi, et al. A coupled encoder–decoder network for joint face detection and landmark localization. Image and Vision Computing, 87: 37-46, 2019.

9. CHANG, Feng-Ju, et al. Expnet: Landmark-free, deep, 3d facial expressions. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition IEEE (FG 2018), pp. 122-129, 2018.
10. Yaniv, J., Newman, Y., & Shamir, A. The face of art: landmark detection and geometric style in portraits. ACM Transactions on graphics (TOG), 38(4), 1-15, 2019.
11. Z. Xu, B. Li, Y. Yuan, and M. Geng, "Anchorface: An anchor-based facial landmark detector across large poses," in Thirty-Fifth AAAI Conference on Artificial Intelligence, pp. 3092–3100, 2021.
12. W. Cao, V. Mirjalili, and S. Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. In Pattern Recognition Letters, volume 140, page 325-331, 2020.
13. K. Chang, C. Chen, and Y. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In CVPR, pages 585–592, 2011.
14. A. Al-Shannaq and L. Elrefaei. Age estimation using specific domain transfer learning. Jordanian Journal of Computer and Information Technology, 6(2):122–139, 2020.
15. R. Rothe, R. Timofte, and L. V. Gool. Deep expectation of real and apparent age from a single image without facial landmarks. Int. J. Comput. Vis., 126(2):1–14, Aug. 2016.
16. W. Shen, K. Zhao, Y. Guo, and A. Yuille. Label distribution learning forests. In Proc. NIPS, 2017.
17. Li, Shichao and Cheng, Kwang-Ting. Visualizing the Decision-making Process in Deep Neural Decision Forest. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.
18. N. Liu, F. Zhang, F. Duan. Facial Age Estimation Using a Multi-Task Network Combining Classification and Regression. IEEE Access, 8: 92441–92451, 2020.

第６章 本研究の総括と今後の展望

6.1 まとめ

顔画像からの年齢推定は，コンピュータビジョンにおいて重要かつ挑戦的な課題であった。これまでに提案された年齢推定法では，人種や性別が違う場合にパフォーマンスが低下する，画像を撮る時のポーズが異なる場合にもパフォーマンスが低下するなど問題がある。それらの問題を解決するため，本論文は Cross-dataset 学習法，頭部の向き推定を併用した学習法，Face Landmarks を用いたマルチタスク学習法の３つ手法を提案し，Morph，CACD，AFAD，UTK Face など代表的な顔画像のデータセットを用いて年齢推定の検証実験を行った。

第 1 章では，画像より年齢推定の研究背景と研究の目的，および本論文の全体構成について説明した。

第 2 章では，顔画像からの年齢推定に関して従来研究を調査・分析し，関連ディープラーニングの特性について解説した。そして，近年の技術についてまとめ，現存の問題点，改善すべきところなど課題を抽出した。

第 3 章では，既存の年齢推定用データセットの概要を説明し，含まれた画像数の不足による問題，画質の問題及び単一人種による問題を明らかにした。そして，高品質な学習データ不足の課題解決を目的として，複数のデータセットを併用した Cross-dataset 学習法を提案した。Cross-dataset 学習法のアルゴリズム，ネットワーク構造を説明した。実験については，数が少なく質も低い CACD と AFAD データセットと数が多く質も高い Morph データセットを用いて提案方法による年齢推定実験を行った。その結果，提案方法による年齢推定の精度は従来技術より CACD で 0.7 歳（平均推定年齢），AFAD で 0.2 歳を向上できたことを示した。

第 4 章では，写真や動画の中での年齢推定の精度が低下する問題についてさらに分析し，人間のポーズが異なることによる精度の低下への影響を実験で明らかにした。そして，ポーズによる精度低下の課題解決を目的として，頭部の向き推定を併用した学習法を提案した。提案した年齢推定と頭部の向き推定法について，詳細なアルゴリズムやネットワーク構造を説明した。画像に対して，まず頭部の向き推定を行なって，推定した３つの角度により，30 度以内の画像のみ年齢推定を行なった。実験を行った結果，頭部の向き角度の制限を用いて，近年よく利用されるデータベース CACD と AFAD による年齢推定の精度を従来技術より 0.8 歳を向上できたことを示した。

第 5 章では，近年よく利用されているマルチタスク学習法に着目し，年齢推定の精度向上のために報告されている，性別推定を用いたマルチタスク学習法による年齢推定法について説明した。この方法の課題であるさらなる精度向上を目的として，Face Landmarks を用いたマルチタスク学習法を提案した。Face Landmarks の概要とマルチタスク学習につい

て，アルゴリズムやネットワーク構造を説明し，年齢推定実験を設計・実行した。その結果，Face Landmarks を用いたマルチタスク学習法により，よく利用されるデータベース CACD と UTK Face で年齢推定の精度は従来技術より 0.5 歳を向上できたことを示した。

第 6 章では，以上の各章で得られた結論を総括している。

本研究成果は，顔画像からの年齢推定における課題を分析するとともに従来技術の問題を明らかにし，複数の精度向上案を提案した。さらに実装のためのアルゴリズムやディープラーニングモデルを開発し，実験でそれらの有効性を確認した。今後はリアルタイム処理など改良を加え，実用化を目指したい。

6.2 今後の課題

今後は，提案手法に対してベースラインネットワークの精度はまだ改善可能だと考えて，ネットワークを調整検討しつつ、より高い精度を目指す。頭部姿勢推定に関して、今は閾値外がの画像を除いたが、将来的には，非正面画像も利用し、カリブレーションなど方法を用いて、より幅広い動画でも利用できるように目指す。今提案したマルチタスク学習法は Face Landmarks のみ利用されたが、将来的には性別，人種，表情など，他の顔特徴タスクも利用して、より精度が高める可能性が予想できる。

参考文献

1. A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. IEEE Transaction on Systems, Man, and Cybernetics, Part B: Cybernetics, 34(1):621–628, 2004.

2. H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In Proc. ICB, pages 1–8, 2013.

3. X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. In Proceedings of the ACM International Conference on Multimedia, pages 307–316, 2006.

4. A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(4):442–455, 2002.

5. Z. Song, B. Ni, D. Guo, T. Sim, and S.Yan. Learning universal multi-view age estimator using video context. In Proceedings of the IEEE International Conference on Computer Vision, pages 241–248, Nov 2011. 1

6. C. Shan, F. Porikli, T. Xiang, and S. Gong, editors. Video Analytics for Business Intelligence. Studies in Computational Intelligence. Springer, 2012.

7. Y. Kwonand, and N. Lobo. Age classification from facial images. In IEEE CVPR, pages 762–767, 1994.

8. G. Guo, G. Mu, Y. Fu, and T. Huang. Human age estimation using bioinspired features. In IEEE CVPR, pages 112–119, 2009.

9. T. Ojala, M. Pietikinen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In ICPR, pages 582–585, 1994.

10. D. Lowe. Object recognition from local scale-invariant features. In IEEE ICCV, pages 1150–1157, 1999.

11. D. Gabor. Theory of communication. J. Inst. Electr. Eng., 93(26):429–457, Nov. 1946.

12. T. -Y. Lin, P. Dolla r, R. B. Girshick, K. He , B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In IEEE CVPR, volume 1, page 4, 2017.

13. K. He, G. Gkioxari, P. D. ar, and R. Girshick. Mask R-CNN. In IEEE ICCV, pages 2980–2988, 2017.

14. Zhang, K., Zhang, Z., Li, Z., and Qiao, Y.. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503. 2016.

15. S. Escalera, J. Fabian, P. Pardo, X. Baro, J. Gonzalez, H. J. Escalante, and I. Guyon. ChaLearn 2015 apparent age and cultural event recognition: Datasets and results. IEEE International Conference on Computer Vision, ChaLearn Looking at People workshop, pages 1–9, 2015.

16. H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic estimation from face images: Human vs. machine performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(6):1148–1161, 2015. 1
17. G. Guo and G. Mu. Human age estimation: What is the influence across race and gender? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 71–78, 2010. 1
18. G. Guo, G. Mu, Y. Fu, C. Dyer, and T. Huang. A study on automatic age estimation using a large database. In Proceedings of the IEEE International Conference on Computer Vision, pages 1986–1991, 2009. 1
19. T. Perrett and D. Damen. Recurrent assistance: Cross- dataset training of lstms on kitchen tasks. In Computer Vi- sion Workshop (ICCVW), 2017 IEEE International Conference on, pages 1354–1362, IEEE, 2017.
20. Cootes TF, Edwards GJ, and Taylor CJ. Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 23(6):681–685, 2001.
21. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4920–4928, 2016.
22. B. Chen, C. Chen, and W. H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. IEEE Trans. Multimedia, 17(6):804–815, 2015.
23. K. Ricanek and T. Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In Proc. FG, pages 341–345, 2006.
24. Simonyan. K, Zisserman. A. Very deep convolutional networks for largescale image recognition. International Conference on Learning Representations, 2015.
25. Krizhevsky. A, Sutskever. I, Hinton. GE. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (NIPS), pages 1097–1105, 2012.
26. I. Huerta, C. Fernandez, and A. Prati. Facial age estimation through the fusion of texture and local appearance descriptors. In Porc. ECCV Workshops, pages 667–681, 2014.
27. K. Chen, S. Gong, T. Xiang, and C. L. Chen. Cumulative attribute space for age and crowd density estimation. In Proc. CVPR, pages 2467–2474, 2013.
28. K. Chang, C. Chen, and Y. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In CVPR, pages 585–592, 2011.
29. Jia. Y, Shelhamer. E, Donahue. J, Karayev. S, Long. J, Girshick. R, Guadarrama. S, Darrell. T. Caffe: Convolutional architecture for fast feature embedding. In International Conference on Multimedia, pages 675–678, 2014.

30. B. B. Gao, C. Xing, C.W. Xie, J.Wu, and X. Geng. Deep label distribution learning with label ambiguity. IEEE Transactions on Image Processing, PP(99):1–1, 2016.
31. W. Shen, K. Zhao, Y. Guo, and A. Yuille. Label distribution learning forests. In Proc. NIPS, page 834-843, 2017.
32. X. Geng. Label distribution learning. IEEE Transactions on Knowledge and Data Engineering, 28(7):1734–1748, 2016
33. X. Geng, K. Smith-Miles, and Z. Zhou. Facial age estimation by learning from label distributions. In Proc. AAAI, pages 451–456, 2010.
34. H. Liao, Y. Yan, W. Dai, and P. Fan. Age Estimation of Face Images Based on CNN and Divide-and-Rule Strategy. In Mathematical Problems in Engineering, 2018.
35. S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao. Using ranking-CNN for age estimation. In IEEE ICCV, pages 5183–5192, 2017.
36. W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille. Deep Regression Forests for Age Estimation. In IEEE CVPR, pages 2304–2313, 2018.
37. Ruiz. N, Chong. E, Rehg. J.M. Fine-grained head pose estimation without key-points. In CVPR workshops, pp. 2074–2083, 2018.
38. X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 146–155, 2016.
39. Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4920–4928, 2016.
40. Simonyan. K, Zisserman. A. Very deep convolutional networks for large-scale image recognition. 444 CoRR abs/1409.1556, 2014.
41. Russakovsky. O, Deng. J, Su. H, Krause. J, Satheesh. S, Ma. S, Huang. Z, Karpathy. A, Khosla. A, Bernstein. M, Berg. AC, Fei-Fei. L. Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV), 115(3):211–252, 2015.
42. Krizhevsky. A, Sutskever. I, Hinton. GE. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (NIPS), 2012.
43. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.
44. X. Zhu, and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In IEEE CVPR, pages 2879–2886, 2012.
45. P.N.Belhumeur, D.W.Jacobs, D.Kriegman, and N.Kumar. Localizing parts of faces using a consensus of exemplars. In IEEE CVPR, pages 545–552, 2011.

46. F. Wang, H. Han, S. Shan, and X. Chen. Multi-task learning for joint prediction of heterogeneous face attributes. In IEEE FG, pages 173–179, 2017.
47. C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In IEEE ICCVW, pages 397– 403, 2013.
48. K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In Second international conference on audio and video-based biometric person authentication, volume 964, pages 965–966, 1999.
49. A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In International Conference on Computer Vision, pages 1021-1030, 2017
50. R. Rothe, R. Timofte, and L. V. Gool. Deep expectation of real and apparent age from a single image without facial landmarks. Int. J. Comput. Vis., 126(2):1–14, Aug. 2016..
51. B. Yoo, Y. Kwak, Y. Kim, C. Choi and J. Kim. Deep Facial Age Estimation Using Conditional Multitask Learning With Weak Label Expansion. In IEEE Signal Processing Letters, 25(6):808–812, 2020.
52. Z. Zhang, Y. Song and H. Qi. Age Progression/Regression by Conditional Adversarial Autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4352–4360, 2017.
53. WANG, Lezi, et al. A coupled encoder–decoder network for joint face detection and landmark localization. Image and Vision Computing, 87: 37-46, 2019.
54. CHANG, Feng-Ju, et al. Expnet: Landmark-free, deep, 3d facial expressions. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition IEEE (FG 2018), pp. 122-129, 2018.
55. Yaniv, J., Newman, Y., & Shamir, A. The face of art: landmark detection and geometric style in portraits. ACM Transactions on graphics (TOG), 38(4), 1-15, 2019.
56. Z. Xu, B. Li, Y. Yuan, and M. Geng, “Anchorface: An anchor-based facial landmark detector across large poses,” in Thirty-Fifth AAAI Conference on Artificial Intelligence, pp. 3092–3100, 2021.
57. W. Cao, V. Mirjalili, and S. Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. In Pattern Recognition Letters, volume 140, page 325-331, 2020.
58. A. Al-Shannaq and L. Elrefaei. Age estimation using specific domain transfer learning. Jordanian Journal of Computer and Information Technology, 6(2):122–139, 2020.

59. Li, Shichao and Cheng, Kwang-Ting. Visualizing the Decision-making Process in Deep Neural Decision Forest. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.
60. N. Liu, F. Zhang, F. Duan. Facial Age Estimation Using a Multi-Task Network Combining Classification and Regression. IEEE Access, 8: 92441–92451, 2020.
61. G. Guo and G. Mu. Human age estimation: What is the influence across race and gender? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 71–78, 2010. 1
62. G. Guo, G. Mu, Y. Fu, C. Dyer, and T. Huang. A study on automatic age estimation using a large database. In Proceedings of the IEEE International Conference on Computer Vision, pages 1986–1991, 2009. 1
63. T. Perrett and D. Damen. Recurrent assistance: Cross- dataset training of lstms on kitchen tasks. In Computer Vi- sion Workshop (ICCVW), 2017 IEEE International Conference on, pages 1354–1362, IEEE, 2017.
64. J. Li, S. Xiao, F. Zhao, J. Zhao, J. Li, J. Feng, S. Yan, and T. Sim. Integrated face analytics networks through cross-dataset hybrid training. In Proceedings of the 2017 ACM on Multimedia Conference, pages 1531–1539, 2017.
65. Z. Kuang, C. Huang, and W. Zhang. Deeply Learned Rich Coding for Cross-Dataset Facial Age Estimation. In ICCVW, pages 338-343, 2015.
66. K. Chang, C. Chen, and Y. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In CVPR, pages 585–592, 2011.
67. S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao. Using ranking-CNN for age estimation. In IEEE ICCV, pages 5183–5192, 2017.
68. Z. Ji, C. Lang, K. Li and J. Xing. Deep Age Estimation Model Stabilization from Images to Videos. In International Conference on Pattern Recognition, 2018.
69. W. Pei, H. Dibeklio˘ glu, T. Baltrušaitis and D. Tax. Attended End-to-End Architecture for Age Estimation From Facial Expression Videos. In IEEE Transactions on Image Processing, volume 29, pages 1972-1984, 2019.
70. Caruana, R. Multitask Learning. Machine Learning 28, 41–75 (1997).
71. Y. Kwonand, and N. Lobo. Age classification from facial images. In IEEE CVPR, pages 762–767, 1994.
72. D. Yi, Z. Lei, and S. Z. Li. Age estimation by multi-scale convolutional network. In IEEE ICCV, pages 144–158, 2015.
73. F. Wang, H. Han, S. Shan, and X. Chen. Multi-task learning for joint prediction of heterogeneous face attributes. In IEEE FG, pages 173–179, 2017.

74. H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. IEEE Trans. Pattern Anal. Mach. Intell., Aug. 2017.
75. M. Duan, K. Li and K.L. An ensemble CNN2ELM for age estimation. IEEE Trans. Inf. Forensics Secur, 13(3), 758–772, 2017.
76. B. Yoo, Y. Kwak, Y. Kim, C. Choi and J. Kim. Deep Facial Age Estimation Using Conditional Multitask Learning With Weak Label Expansion. In IEEE Signal Processing Letters, 25(6):808–812, 2020.
77. M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. Machine Learning, 37(2):183–233, 1999
78. G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In Proc. CVPR Workshops, pages 34–42, 2015.

謝辞

本論文は筆者が東京都市大学総合理工学研究科情報専攻の後期課程（2019年～2022年）に在学中に，視覚メディア研究室にて行った研究についてもとめたものであります。

本論文をまとめるにあたり，終始懇切なる御指導御教示を賜りました東京都市大学工学研究科の包躍教授に心から深く感謝申し上げます。3年間，画像処理に関する知識，原理をはじめ，動画表示や人工知能に関する知識など多岐に渡る分野において，厳しくも丁寧に御指導いただきました。研究を進めるにあたり、東京都市大学総合理工学研究科情報専攻の田口亮教授，塩本公平教授，神野健哉教授及び富山県立大学工学部情報システム工学科の中田崇行准教授には御多忙にもかかわらず御指導や有益な助言を頂き、深く御礼申し上げます。

また，研究を進めるにあたり，学術論文の作成や国際学会への参加など貴重な体験もさせていただいたことで，私の研究者としての心構えや考え方についても賜りました。

また，AI，画像処理，企業との共同研究などに関する研究を共に，行ってきた視覚メディア研究室の皆様にも深く感謝いたします。毎週欠かさずのゼミにおける様々な意見を交換することで，新たな発見なども多く学ぶことができました。

最後に，今までずっと私を支えてきた家族に感謝の意を表します。本当に，ありがとうございました。