Deep Convolutional Neural Network による顔器官点検出における最適なミニバッチ作成方法

山下　隆義† 木村　真稔†　福井　宏†　山内　悠嗣†　藤吉　弘亘†
† 中部大学
E-mail: yamashita@cs.chubu.ac.jp

Abstract

Deep Convolutional Neural Network (DCNN) の汎化性能は様々な分野で注目されているが，学習過程におけるパラメータの調整や学習サンプルの与え方など，まだまだ手探りなことが多く，DCNN の非常に高い汎化性能を充分に引き出すことは容易ではない．そこで，我々は DCNN を用いた顔器官点検出において，最適なネットワークを学習するために，学習サンプルの与え方について検討する．サンプルの与え方として，1）data augmentation を利用した方法，2）data augmentation したサンプルからランダムにサンプルを選択する方法，3）人物を固定して data augmentation を適用する方法，4）data augmentation を利用しない方法の４つを比較する．実験では，これらのサンプルの与え方による顔器官点検出性能を比較する．また，顔向きに対応した従来手法との比較も行い，DCNN による顔器官点検出手法の顔向きに対する頑健性を示す．

1　はじめに

顔器官点検出は，顔認証や表情推定等の顔画像処理における重要な前処理であり，画像認識分野で活発的に研究されている分野の１つである．顔認証や表情推定を高精度に行うために，顔器官点検出は，顔の向きや照明変化，表情や隠れが生じるような顔に対する頑健性が求められている．これまで，このような困難な条件に対応するために多くの手法が提案されており，それらは２つの方法に大別することができる．１つ目は識別問題として取り扱う方法であり [2][10][21] ，２つ目は回帰問題として取り扱う方法である [3][5][6][12][16][18]．識別問題の場合，各器官点の候補をスライディング・ウィンドウのアプローチにより探索する．探索範囲は各器官点の初期位置をもとに局所的な探索を行う．各器官点の候補の中から顔の形状の制約を利用して，最適な位置を推定する [2][12][21]．回帰問題の場合，探索を行わずに顔全体における各器官点の位置を回帰により推定する．各器官点は，収束条件を満たすまで繰り返し位置の更新を行うことで推定される．拘束条件として，位置の移動量が一定より小さくなることを条件とすることが多い．これらの方法は，疎密探索のフレームワークで組み合わせることが可能である．それにより顔器官点検出の精度を向上させる手法も提案されている [16][20][15]．従来手法において，これらの顔器官の位置を推定する方法とともに，どのような特徴量を用いるかも重要な要素の１つである．輝度や勾配に着目する特徴量が多く提案されているが，どの特徴量を採用するのが最適であるかは定かではない．

本稿では特徴量の設計を自動的に行うことが可能な Deep Convolutional Neural Network (DCNN) による顔器官点検出手法を提案する．DCNN の学習は，勾配降下法によりネットワーク全体のパラメータを繰り返し更新させ，誤差の収束を図っている．その際，学習サンプルはミニバッチという少量のサブセットとして与えられて誤差の算出を行っている．勾配降下法は局所解探索であり，学習サンプルの与え方が重要となる．我々は，最適なネットワークのパラメータを得るための学習サンプルの与え方について検討する．

2　関連研究

顔器官点検出は顔認証や表情推定などの顔画像処理の重要な前処理である．従来の一般的な手法として，顔全体の形状や見えを捉える Active Shape Model や Active Appearance Model がある [5][4]．これらの手法はあらかじめ学習した人物の顔器官点を頑健に検出できるが，未知の人物の顔器官点に対応することが困難である．識別アプローチの手法では，Vukadinovic らが Gentle Boost とガボールフィルタを利用して各器官点を独立に検出している [19]．Amberg らは，検出器を用いて顔全体から多くの顔器官候補を検出し，それらの候補から最適な形状を branch-and-bound 手法により抽出している [1]．これらの手法は，顔の見えの変化が大きい場合，器官候補を検出することができないことが多い．見えの変化に対応するために，Uricar らは Deformable Part Model [14] を用いた手法を提案している．また，Belhumerur らは，器官位置の全体モデルと各器官の検

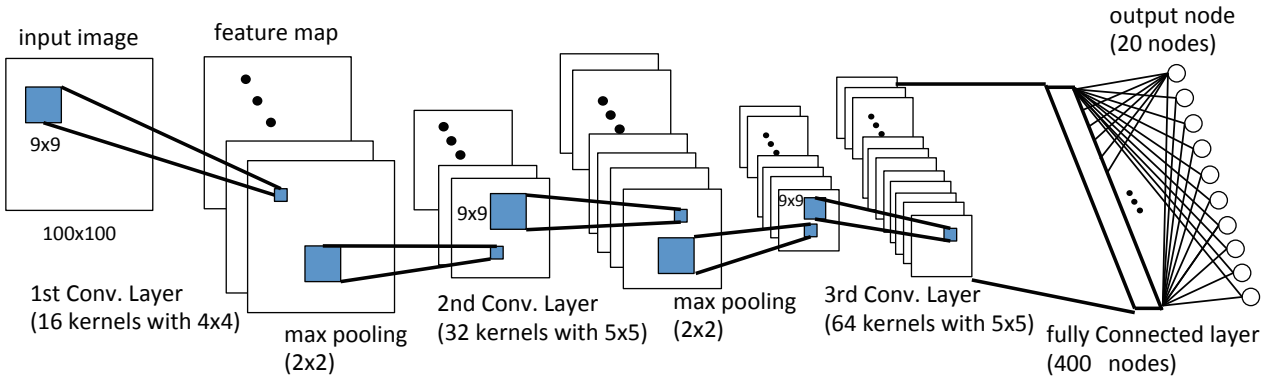

図 1　Deep Convolutional Neural Network の構造

出をベイジアンモデルにより統合する手法を提案している [2]．この手法は検出精度が高いが，高解像度の画像が必要となる．

回帰による手法として，Valstar らがサポートベクター回帰と条件付き Markov Random Field を用いた手法で全体の最適化を図る手法が提案されている [18]．また，Cao らは Random Ferms を用いた回帰手法を提案している [3]．顔の向きに頑健な手法として，顔向きごとに Regression Forests を用いた Conditional Regression Forests (CRF) が Dantone らにより提案されている [6]．CRF は 2 つの処理から構成されている．1 つは顔の向き推定であり，2 つ目が Regression Forests による顔器官の回帰である．この手法による顔器官点検出性能は，顔の向き推定の精度に依存している．また，眼鏡のフレームに対する誤認識といった問題もある．

Krizhevsky らが DCNN を一般物体認識に応用し，ベンチマークテストで優勝して以降，DCNN が非常に注目されている [7]．DCNN は LeCun らが提案した畳み込み処理をネットワークの初期層に導入したニューラルネットワークであり，平行移動等の変動に対する不変性を持っている [11]．DCNN の利点は，適用タスクに適した特徴を自動的に抽出できる点である．従来，特徴の選定や設計は人手で行っていたが，DCNN の場合は人手での特徴設計が不要であり，色やエッジ抽出のような単純な特徴からそれらを組み合わせたような複雑な特徴までを自動で抽出可能である．また，Sun らは DCNN をベースにカスケード構造を導入した顔器官点検出を提案している [17]．この手法では，顔全体の画像を入力し，各器官の大まかな位置を検出する．そして，その後，各器官について詳細な位置検出を行っている．この手法は非常に優れた検出精度であるが，複雑なネットワーク構成になっており，また人手による特徴設計が必要である．さらに，眼鏡などの装飾品がある場合，正しく検出できないことが多い．

従来の多くの手法は，顔向きへの対応を試みているが，複雑な構成になる場合が多い．それにより実時間での処理が困難となっている．本稿では，シンプルな構造の DCNN をベースに顔向きに対する頑健性を向上させた顔器官点検出を提案する．

3　提案手法

我々はシンプルな構成の DCNN をベースとし，最適なネットワークを学習するための学習サンプルの与え方について検討する．DCNN は勾配法により，ネットワークのパラメータは繰り返し更新しながら最適なパラメータを学習する．その際，学習セットを少量のサンプルを含むサブセットに分割して，学習を行う．勾配法は，局所解探索であり，学習サンプルの与え方が重要となる．そこで，どのようなサブセットの作り方が顔器官点検出に適しているかを検討する．

3.1　Deep Convolutional Neural Network

図 1 に示すように，DCNN は畳み込み層とプーリング層が階層的になっており，その後に全結合層が続いている．入力としては，RGB 画像やグレースケール画像，勾配画像，または正規化処理を施した画像が与えられる．与えられた画像に対して，畳み込み層ではサイズ $k_x \times k_y$ のフィルタを畳み込む．畳み込んで得られた値は活性化関数を通した後，特徴マップに格納される．畳み込み層のフィルタは M 個あり，各フィルタからそれぞれ特徴マップを作成する．その後，特徴マップはプーリング層で，縮小処理される．縮小処理は Max Pooling が用いられることが多い．Max Pooling は，2×2 のようなあらかじめ決めた領域における最大値により間引きを行う方法である．畳み込み層とプーリング層は階層的になっており，これらを繰り返すことで深いネットワーク構造を形成している．これらの階層的な処理の後に従来のニューラルネットワークと同様の全結合層が続いている．全結合層は，1 層前の特徴マップを 1 次元にして，入力層として用いる．入力された値は重み付きの全結合を行い活性化関数を通した後，次層の出力

となる．最終層は顔器官点の座標数のユニットがあり，各座標位置を出力する．識別問題に用いる DCNN の場合，出力層のユニット数は認識対象のクラス数となり，特定のユニットは 1 に近くなり，そのほかのユニットは 0 に近い出力値をクラス確率として出力する．一方，顔器官点検出の場合は回帰を行う DCNN であり，出力層のユニット数は，顔器官の座標数となる．すなわち，10 点の顔器官を回帰する場合，出力層のユニットは 10 点の器官点の x 座標と y 座標であるので 20 個となる．各ユニットは 0 から 1 の値を出力し，それらを画像サイズで乗じることで座標の実数値を得ることができる．

DCNN は教師ありの学習であり，各ネットワークのパラメータは逆誤差伝播法により繰り返し更新しながら最適な値を求める [13]．その際，初期値は乱数で初期化されている．逆誤差伝播法は式 (1) のように各学習サンプルの誤差 E_p を累積して全体の誤差 E を算出する．そして，勾配法により，全体の誤差 E を最小とするようにパラメータを更新する．各パラメータの更新量は式 (2) に示す誤差 E の偏微分から求める．

$$E = \frac{1}{2}\sum_{p=1}^{P} E_p \tag{1}$$

$$w_{ji}^{(l)} \leftarrow w_{ji}^{(l)} + \Delta w_{ji}^{(l)} = w_{ji}^{(l)} - \lambda \frac{\partial E_p}{\partial w_{ji}^{(l)}} \tag{2}$$

ここで，$\{p|1,...,P\}$ は学習サンプルであり，$\boldsymbol{o_p}$ は学習サンプル p に対する出力層の応答値 $\boldsymbol{t_p}$ は教師信号である．λ は学習における更新率，$w_{ji}^{(l)}$ は l 番目の層のノード i と次層のノード j との結合重みである．各学習サンプルの誤差は，出力層の応答値と教師信号との誤差から求めることができる．更新量は，式 (3) から求めることができる．

$$\Delta w_{ji}^{(l)} = -\lambda \delta_k^{(l)} y_j^{(l-1)} \tag{3}$$

$$\delta_k^{(l)} = e_k \phi(V_k^{(l)}) \tag{4}$$

$$V_k^{(l)} = \sum_j w_{kj}^{(l)} * y_j^{(l-1)} \tag{5}$$

$y_j^{(l-1)}$ は $(l-1)$ 番目の層のノード j であり，e_k はノード k の誤差，$V_k^{(l)}$ は前層の全てのノードからの重み付き累積の値である．また，式 (4) から勾配は算出される．ここで，ϕ は活性化関数であり，シグモイド関数や Rectified Liner Unit (ReLU)[7]， Maxout[8] などがある．ネットワーク全体のパラメータはあらかじめ決められた回数または収束条件を満足するまで繰り返し更新される．

誤差 E を算出するための学習サンプルの与え方として， *full-batch*, *online*, and *mini-batch* がある．*Full-batch* は，全ての学習サンプルを用いて 1 回の更新を行う方法である．この方法は，勾配の変化が大きいため収束しにくい．*Online* は学習サンプルを 1 枚ずつ与え逐次更新する方法である．この方法は，勾配の変化が小さいため最適解に得やすいが，非常に多くの更新回数を要するため処理時間がかかる．*Mini-batch* は，これらの中間的なアプローチであり，少量のサンプルを用いて 1 回の更新を行う．この方法は，更新に要する処理時間が比較的短く，繰り返し行えるため，十分な勾配の変化を得ることがでる．そのため，DCNN の学習において，*Mini-batch* は一般的な方法として利用されている．

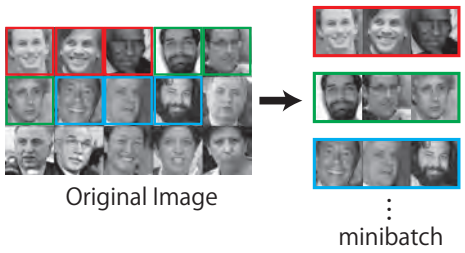

図 2　Data augmentation なしの mini-batch によるサブセット作成方法

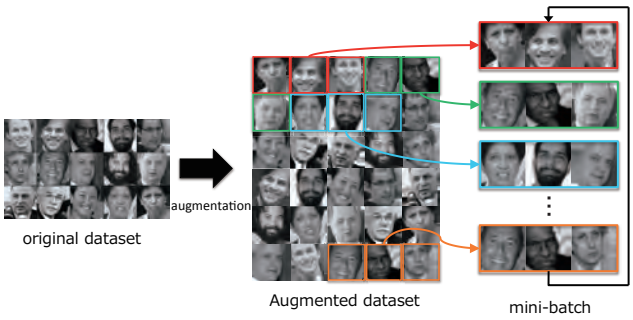

図 3　Augmentation mini-batch によるサブセット作成方法

3.2　Mini-batch 学習

DCNN のパラメータを学習するために，学習データセットは mini-batch と呼ぶサブセットに分割される．学習の各繰り返しにおいて，異なるサブセットを入力してパラメータを更新する．全てのサブセットを用いて更新を行うと，最初のサブセットを入力してパラメータを更新する．全てのサブセットを一度利用することをエポックと呼ぶ．そして，一定回数または収束条件を満たすまで繰り返しサブセットを入力し続ける．DCNN の学習は勾配法に基づいており，良いパラメータを得るためにはサブセットの与え方が重要となる．しかしながら，これまでどのようなサブセットを用意すれば良いのか十分な知見が得られていない．そこで，我々はサブセットの作成方法が性能に与える影響を調査し，

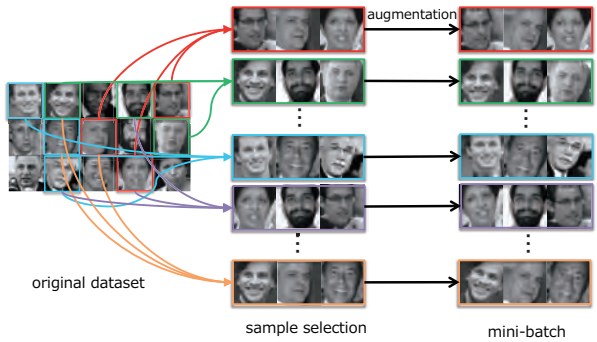

図 4　Random mini-batch によるサブセット作成方法

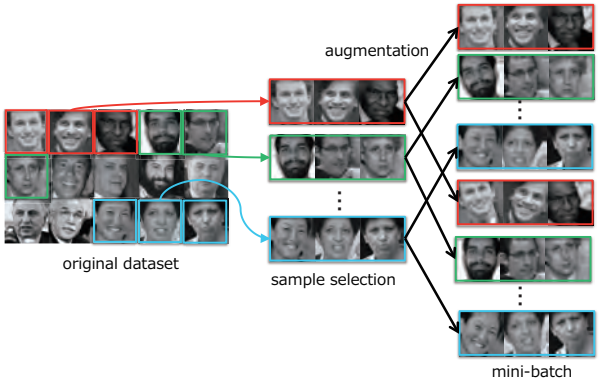

図 5　Fixed-person mini-batch によるサブセット作成方法

顔器官点検出において最も良いパラメータを得る方法を検討する．

1 つ目の作成方法は，学習サンプル数を data augmentation により増加させ，そこからサブセットを作成する augmentation mini-batch (以下，aug. mini-batch) である．Data augmentation は少ないデータセットの枚数を増加させる一般的な方法で，並進移動や回転，スケーリングを加えて数千枚の画像を数百万枚にする．図 3 に示すに, aug. mini-batch は，増加させたデータセットからランダムにサンプルを選択して，mini-batch に分割する．学習において，すべてのサブセットは繰り返し利用する．

2 つ目の作成方法は，サブセットを繰り返し利用しない random mini-batch である．図 5 に示すように，まず，random mini-batch はもとのデータセットからランダムにサンプルを選択する．そして，それらに data augmentation を利用して変形を加える．Aug. mini-batch は作成された mini-batch を繰り返し学習に利用するが，random mini-batch はランダムにサンプルを選択して mini-batch を作成する．これにより，学習中同じ mini-batch を利用することはなく，常に異なるサンプルが含まれる mini-batch により学習を行う．

3 つ目の作成方法は，サブセットに対して data augmentation を加える fixed-person mini-batch である．図 5 に示すように，random mini-batch と同様に，data augmentation を行う前に元のデータセットからランダムにサンプルを選択する．そして，選択されたサブセットに対して，data augmentation を行い学習を行う．全てのサブセットを利用して学習した後，random mini-batch と異なり，再度サブセットを利用する．その際，異なる data augmentation を加えることで，元の組み合わせは同じであるが変形の異なるサブセットを生成することができる．

4　実験

提案する mini-batch の作成方法による検出精度の比較を行い，顔器官点検出において最適な学習方法を検討する．また，従来法との比較を行い，DCNN を用いた顔器官点検出手法の有効性を示す．評価には，Labeled Faces in the Wild (LFW) データセットを用いる [9]. LFW データセットは，WEB から収集された撮影環境に制約がない画像である．Dantone らは LFW データセットの一部に，両目の目尻と目頭，鼻の下部 2 点，唇の両端と上下点の合計 10 点の器官点のアノテーションを行い，データセットを公開している??．アノテーションされているデータは，学習用に 1500 枚，評価用に 927 枚である．Dantone らとの比較を行うために，我々はこのサブセットを用いる．Aug. mini-batch と Random mini-batch は，学習データセットに対して，data augmentation を行い 20000 枚の画像を生成する．Data augmentation の範囲は，並進が ±10 ピクセル，回転が ±15° としている．学習する画像は 100×100 ピクセルのグレースケールとする．DCNN のネットワーク構成は図 1 に示すように，3 層の畳み込み層とプーリング層，1 層の全結合層，1 層の回帰層となっている．各畳み込み層のフィルタサイズは 9×9 であり，活性化関数に Maxout を用いる．また，フィルタの数はそれぞれ 16，32，64 としている．全結合層のユニット数は 400 個であり，汎化性を向上させるために Dropout を利用して学習する．回帰層は 20 個のユニットがあり，各ユニットは器官点の x 座標または y 座標に相当している．学習の更新率は 0.1，mini-batch のサイズは 10，更新回数は 30 万回としている．

評価方法は，従来の誤差計算方法と同様に，検出結果と教師信号との誤差を累積し，それを両目間の距離で正規化した値を誤差として算出する [6]. 両目間の距離で正規化することで，顔や画像の大きさに対する不変な尺度となっている．また，検出精度は，誤差が 10%以内であれば正解として求める．

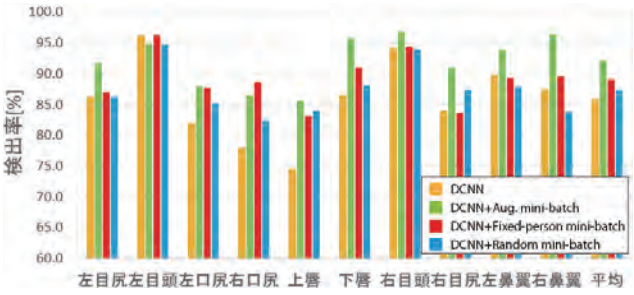

図 6　mini-batch 方法による性能比較

4.1　各 mini-batch 方法による比較

各 mini-batch 方法による性能比較を図 6 に示す．ここで黄色線の DCNN は data augmentation を用いていない場合の性能である．Data augmentation を利用することで，学習サンプルのバリエーションが増え，全ての mini-batch 作成方法の性能が上回っていることが分かる．Random mini-batch はサブセットに含まれる人物と変形のバリエーションが豊富であるが，他の mini-batch 作成方法と比べて性能が低くなっている．これは，DCNN の学習においてサンプル数は重要であるが，データのバリエーションの質を考慮する必要があることを示している．Fixed-person mini-batch は，random mini-batch と同様に変形のバリエーションは豊富であるが，人物のバリエーションに制約をもうけている．これにより，全器官点の平均性能において，Fixed-person mini-batch は，random mini-batch よりも約 2%精度が向上している．

一方，Aug. mini-batch は data augmentation により学習サンプルのバリエーションを増やしているが，同じサブセットを繰り返し利用する．それにより，人物と変形のバリエーションにより一層制約をもうけている．図 6 から Aug. mini-batch は random mini-batch より 4%，fixed-person mini-batch より 2%精度が向上し，最も良い性能となっている．Mini-batch の作成方法として，単純にサンプル数を増加させて学習するのではなく，十分なサンプルを繰り返し利用することが DCNN の学習に必要であることが分かる．

Data augmentation なしの DCNN と aug. mini-batch による DCNN の学習誤差について，図 7 に示す．Aug. mini-batch の学習誤差は data augmentation なしの DCNN に比べて，学習サンプルのバリエーションが多いため大きな値となっている．しかしながら，評価サンプルに対する検出性能は aug. mini-batch による DCNN の方が良い．これは，学習サンプルが少ないため，過学習を起こしていると考えられる．学習サンプルのバリエーションを増やした場合は，学習誤差が最小となるとは限らないので，一定の繰り返し回数ごとに評価を行うことが必要である．

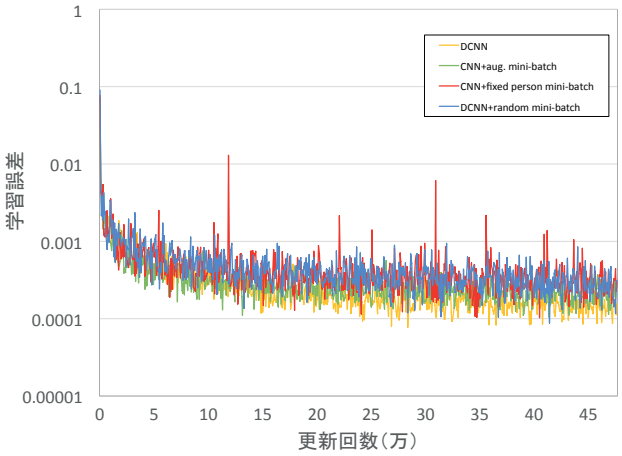

図 7　学習誤差の推移

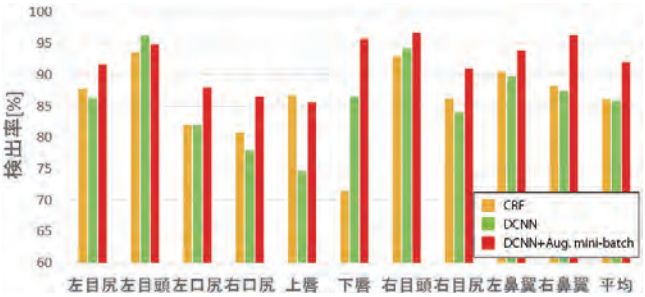

図 8　CRF と aug. mini-batch の比較

4.2　従来手法との比較

次に，提案手法と顔向きに頑健な従来手法である CRF との比較を行う．図 8 に CRF および data augmentation を用いた場合と用いない場合の DCNN による手法の結果を示す．Data augmentation を用いない DCNN は平均性能では CRF を僅かに上回っているが，多くの器官点で CRF より悪くなっている．これより，顔のバリエーションに対する汎化性能が不十分であることが分かる．

一方，data augmentation を用いた DCNN は CRF よりも平均性能において 6%精度が向上している．また，上唇は若干 CRF を下回っているが，それ以外の器官点の精度は大きく向上していることが分かる．下唇は口の開閉で大きく位置が変動するため，顔の器官点において，精度が低下しやすい位置である．このような変動の大きな下唇に対して，提案手法は，96%となっており，CRF の 72%を大きく上回っている．これより，学習サンプルのバリエーションを増やすことで顔の向きや見えの変化に対する頑健性を得ることができている．

顔器官点検出の結果例を図 9 に示す．提案手法により検出した器官点を緑色の点，教師信号を赤色の点で示している．これより，提案手法は異なる顔向きに対して頑健であることが分かる．また，シンプルな構造

図 9　顔器官点検出結果の例

の DCNN を用いているため，3.4GHz の CPU 上で 1 つの顔画像あたり約 10ms で検出することができる．

5　まとめ

我々は DCNN を用いた顔向きに頑健な顔器官点検出を提案した．また，顔器官点検出性能を向上させるために，DCNN の学習における学習サンプルの与え方について検討した．学習サンプルの与え方に関する実験により，data augmentation を行いサンプル数を増加させたデータセットから mini-batch を作成し，繰り返し利用する aug. mini-bath が顔器官点検出に最適な学習サンプルの与え方であった．従来手法との比較では，顔向きに頑健な CRF と比較して，提案手法は顔器官点検出精度を向上させることができた．

参考文献

[1] B. Amberg and T. Vetter. Optimal landmark detection using shape models and branch and bound. In ICCV, 2011.

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using an consensus of exemplars. In CVPR, 2011.

[3] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In CVPR, 2012.

[4] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models -their training and application. Computer vision and image understanding, 1995.

[5] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In ECCV, 1998.

[6] M. Dantone, J. Gall, G. Fanelli, and L. V. Gool. Real-time facial feature detection using conditional regression forests. In CVPR, 2012.

[7] A. Krizhevsky, I. Sutskever, and G. Hinton. magenet classification with deep convolutional neural networks. In NIPS, 2012.

[8] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. arXiv preprint arXiv:1302.4389, 2013.

[9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

[10] L. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via component based discriminative search. In ECCVV, 2008.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P.Haffner. Gradient-based learning applied to document recognition. In IEEE, 1998.

[12] X. Liu. Generic face alignment using boosted appearance model. In CVPR, 2007.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In Proc. Explorations in the Microstructures of Cognition, 1986.

[14] M, Uricar, V. Faranc, and V. Hlavac. Detector of facial landmarks learned by the structure output sum. In VISAPP, 2012.

[15] S, Ren, X. Cao, Y. Wei, J. Sun. Face Alignment at 30000 FPS via Regressing Local Binary Features. In CVPR, 2014.

[16] P. Sauer, T. Cootes, and C. Taylor. Accurate regression procedures for active appearance models. In BMVC, 2011.

[17] Y. Sun, X. Wang, and X. Tang. Deep Convolutional Network Cascade for Facial Point Detection. In CVPR, 2013.

[18] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using regression and graph models. In CVPR, 2010.

[19] D. Vukadinovic, and M. Panic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. In ICSMC, 2005.

[20] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas. Pose-free Facial Landmark Fitting via Optimized Parts Mixture and Cascaded Deformable Shape Model. In ICCV, 2013.

[21] X. Zhu, and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In CVPR, 2012.