LI-013

選択的モジュール起動を用いた Convolutional Neural Networks による変動にロバストな顔検出システム

Robust Face Detection System Based on Convolutional Neural Networks Using Selective Activation of Modules

御手洗 裕輔† 森 克彦† 真継 優和†
Yusuke Mitarai Katsuhiko Mori Masakazu Matsugu

1．はじめに

近い未来に到来する Ubiquitous 情報社会では、機械と人間のインタフェースのために、一般的な環境下で利用可能な画像認識機能、特に撮影された画像から人間の存在や状態を認知できる機能が必須となってくると思われる．このような、画像から人間の存在や状態を認知する機能を実現するためのステップとして、一般的な環境下で撮影された画像中から人間の顔を検出する技術は重要であり、様々な研究[1]が行われている．しかしながら、一般的な環境下で撮影された画像において想定される、認識対象のサイズ変動や回転変動等の様々な変動に対するロバスト性と、誤検出の低減を両立したものはほとんど報告されていない．そこで我々は、一般的な環境下で撮影された画像中からサイズや回転の変動に対してロバストに人間の顔を検出する手法として、Convolutional Neural Networks (以下 CNN)による変動にロバストな顔検出手法を提案し[2]、本研究では、変動ロバスト性と、誤(未)検出の低減を両立するために、選択的モジュール起動を用いた CNN ベースの顔検出システムを構築して、顔のサイズ変動 8 倍、且つ回転変動±30°を許容可能なロバスト性を有し、顔検出性能において FRR1%、FAR7%以下であることを実証した．

2．顔検出システムの概要

Fig.1 に本顔検出システムの顔検出手法の概要を示す．まず、HSV 表色系に変換した入力画像の V 成分のみを CNN の入力層に入力する．同時に HSV 表色系に変換した画像から、主に H 成分を用いて肌色画素の抽出を行い、それを CNN の中間層に肌色検出結果として挿入する．顔検出処理を行う CNN では、入力された V 成分画像を並列階層的に処理することにより、エッジ等の低次の特徴から、徐々に目や口等の高次の特徴を検出していき、目・口検出結果と、中間層に挿入した肌色検出結果とを用いて第 1 顔検出を行う．この第 1 顔検出結果から、入力画像中の顔候補位置を検出し、各顔候補位置近傍の CNN の中間層出力結果分布の重心等の複数の出力分布パラメータ(Distribution of Outputs Parameter：以下 DOP)を抽出する．次に抽出した複数の DOP (DOP ベクトル)に基づいて、特定の変動を与えた顔に特化して学習した複数の特定変動顔検出モジュール(以下モジュール)から起動するモジュールの選択を行い、選択されたモジュールを用いて、第 1 顔検出処理までに検出された特徴検出結果、及び肌色検出結果から第 2 顔検出を行う．上記処理を、入力画像の解像度を 1/2、1/4 に変更した画像に対しても実行する、解像度別 3 チャネル処理を行う．最終的な顔検出結果は、各チャネルの選択された各モジュールの検出結果を閾値処理して、その論理和を最終顔検出結果としている．

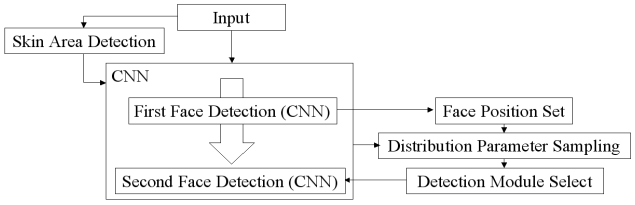

Fig.1：顔検出手法の概要

†キヤノン株式会社
神奈川県厚木市森の里若宮 5-1　キヤノン中央研究所

3．CNN による顔検出

3.1　顔検出 CNN モデル

CNN の基本構成は、前階層のニューロンに結合し、所定の特徴検出を行う特徴検出ニューロンにより構成される特徴検出層(FD 層)と、FD 層のニューロンに結合し、FD 層での特徴検出結果をプールする特徴統合ニューロンにより構成される特徴統合層(FP 層)とを 1 つのセットとし、それが並列階層的に構成されているものである．CNN では、FD 層において前階層の検出結果からより高次の特徴を検出し、それを FP 層でプールして、その階層の検出結果として次の階層に入力する、という処理を階層的に行うことで、低次の特徴から徐々に高次の特徴を検出していく処理を行う．CNN での特定画像パターン検出[3]では、変動に対するロバスト性が高い、位置依存性が無いといった特徴がある．

この CNN を基本とした、顔検出処理を行う新規 CNN モデルを Fig.2 に示す．

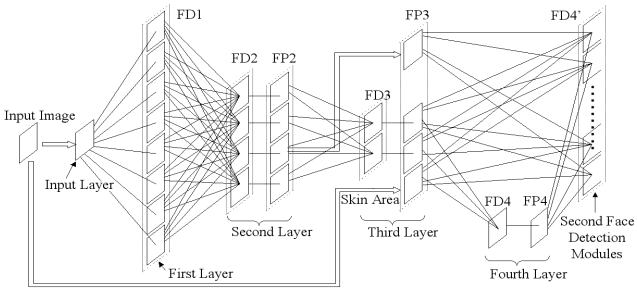

Fig.2：顔検出 CNN モデル

本手法の CNN では、FD1 層は入力層に結合し、入力画像から FD2 層での特徴検出において有効な 8 種類のエッジ等の初期特徴検出を行う．第 1 層は通常の CNN とは異なり特徴統合層を有さず、FD1 層の結果をそのまま FD2 層に入力する．FD2 層は FD3 層で検出する目、口を構成する局所的な特徴として、目尻、目頭、口の左右両端に対応する左右 2 種類の V 字型のエッジと、目や口の上辺、下辺に対応する 2 種類のエッジセグメントの、計 4 種類の特徴を検出する．FP2 層は FD2 層での検出結果を Gaussian Filter によりプールし、第 2 層の特徴検出結果として FD3 層に入力する．FD3 層は目と口の検出を行い、FP3 層で FP2 層と同様にプールする．また FP3 層には、入力画像から抽出した

肌色検出結果を挿入し、さらに特定変動顔検出モジュール層(FD4’ 層)の結合のため、FP2 層の目や口の上辺に対応するエッジセグメント検出結果をコピーする．FD4 層では、FP3 層の目、口検出結果と肌色検出結果から第 1 顔検出を行い、FP4 層でその結果をプールする．FD4’ 層は、特定の変動を与えた顔に特化して学習した複数のモジュールにより構成されている．モジュールは FP3 層と FP4 層に結合し、選択されたモジュールで、各モジュールに対応した特定変動の顔を検出する第 2 顔検出を行う．

3.2 学習手法

CNN の学習では、1 人の人物について、様々なサイズ、回転方向の 21 種類の変動(Table.1 参照)顔画像を用意し、汎化性能向上のためのノイズとして、各変動顔画像に対して、明度、コントラスト、色相の画質変動を与えたものを含む 7 種類の画質の顔画像 21×7=147 枚を作成し、それを学習用の画像として用いる．

	Front	R+30°	R-30°	P +30°	P -30°	T+30°	T -30°
No	1	3	6	9	12	15	18
S-Size 35*35							
No	1 , 2	4	7	10	13	16	19
M-Size 50*50							
No	2	5	8	11	14	17	20
L-Size 70*70							

Table.1：特定変動顔と対応モジュール番号

学習を行うのは FD 層のニューロンのみであり、中次の階層(FD2 層)から順に学習用パターン(Table.2 参照)を繰返し呈示して行う．

FD1 層は FD2 層とともに、Back Propagation (以下 BP)により学習し、FD2 層から FD4’ 層までは各画像の各位置で設定した教師信号に基づいて、一般化デルタルールにより学習した．式(1)に一般化デルタルール、式(2)に BP によるニューロンの結合係数更新式を示す．

(1) $$\Delta w_{n,u,v}^{S(l,m)}(p) = -\eta \cdot \left(y_{\xi,\zeta}^{S(l,m)} - t_{\xi,\zeta}^{S(l,m)}\right) \cdot f'\left(u_{\xi,\zeta}^{S(l,m)}\right) \cdot y_{u,v}^{C(l-1,n)} + \alpha \cdot \Delta w_{n,u,v}^{S(l,m)}(p-1)$$

(2) $$\Delta w_{n,u,v}^{S(l,m)}(p) = -\eta \cdot \left(y_{U,V}^{S(l+1,N)} - t_{U,V}^{S(l+1,N)}\right) \cdot f'\left(u_{U,V}^{S(l+1,N)}\right) \cdot w_{m,\xi,\zeta}^{S(l+1,N)} \cdot f'\left(u_{\xi,\zeta}^{S(l,m)}\right) \cdot y_{u,v}^{C(l-1,n)}$$

$w_{n,u,v}^{S(l,m)}$ は、l 層の第 m 特徴検出ニューロンの、l-1 層の第 n 特徴の相対位置 u,v に位置するニューロンに対する結合係数を示していおり、$\Delta w_{n,u,v}^{S(l,m)}(p)$ は、その結合係数の p 回目の修正量である．$y_{\xi,\zeta}^{S(l,m)}$、$u_{\xi,\zeta}^{S(l,m)}$、$t_{\xi,\zeta}^{S(l,m)}$は、絶対位置 ξ,ζ の、l 層の第 m 特徴検出ニューロンの出力値、内部状態、及び教師信号であり、$y_{\xi,\zeta}^{C(l,m)}$ は同位置の特徴統合ニューロンの出力値である．また α は慣性項係数（定数）、η は学習レートパラメータ（定数）、関数 f は活性化関数であり双曲正接関数を用いた．

	S-Size Front	M-Size Front	L-Size Front	M-Size R+30°	M-Size P+30°	M-Size T+30°	M-Size T-30°
‘<‘ Shaped Corner							
Upper-Edge Segment							
Eye							
Mouth							

Table.2：学習用パターンの例

4．選択的モジュール起動

CNN の第 1 顔検出レベルでは、Table.1 に示した変動顔を全て用いて学習しているため、それら全ての変動顔をカバー可能な顔検出性能を有するが、誤検出が多数発生する．一方、第 2 顔検出では、変動(サイズ、回転)を複数の範囲に分割し、モジュールごとに対応した特定変動顔に特化して学習するため、各モジュールがカバー可能な変動範囲は狭いが、誤検出数が低下する．具体的には、第 1 顔検出と第 2 顔検出である各モジュール単体の結果を比較すると、誤検出数が 1/10 程度に低下する．しかし、全モジュールを同時に起動すると、累積的に(モジュール数に比例するわけではないが)誤検出数が増加してしまい、誤検出数は 1/2 程度低下するだけに留まってしまう．

この問題に対し、CNN の中間層出力分布から、どのモジュールに対応する顔であるのかを推定し、選択的にモジュールを起動することで、累積的な誤検出数の増加を防ぎ、変動に対する高いロバスト性を保ちながら、顔検出性能の向上を行う．モジュールの選択は、CNN の中間層出力分布から DOP ベクトルを抽出して、抽出した DOP ベクトルと各モジュールの DOP ベクトルモデルとのマッチングを行い、マッチングのスコアによりモジュールを選択するというように行う．

5．結果

5.1 学習

CNN の学習には、ソフトピアジャパン提供の顔画像データベース 300 人中から無作為に選択した 100 人分の、1 人につき 147 変動顔画像、計 14700 枚を学習用の顔画像として用いた．またこれ以外に顔を含まない背景画像 147 枚を用意し、これらを合わせた計 14847 枚で学習を行った．

FD1 層から FD3 層までの学習は、学習用の顔画像のみ、FD4 層では背景画像も含めたものを用い、この学習用の画像から、Table.2 に例示した顔の部分画像を抽出して、それらを用いて学習を行った．FD4’ 層の学習では、Table.1 に示した各モジュール番号に対応した変動顔画像と背景画像を用いて学習した．Table.1 に示すように、本顔検出システムで用いているモジュール数は全 20 モジュールである．

肌色抽出に関しては、CNN の学習に用いた 14700 枚の顔画像から肌色として抽出する画素の H 成分、S 成分、V 成分の範囲を決定した．具体的には、H：-12.6° ～45.9° (±180°)、S：0.03～1.2 (0～3)、V：0.17～0.95 (0～1)の 3 つの範囲条件を満たしている画素を肌色として抽出するようにした．括弧内は各成分の値域である．このように、H の範囲は S や V と比較すると、値域に対する割合が小さく、肌色抽出における主要な要素となっている．これらの範囲は、一般的な顔検出を目的とする肌色抽出範囲としては十二分に広い設定であり、明らかに肌色とみなせない領域における顔検出を抑制する程度の補助的なものとしている．

5.2 DOP 抽出

DOP の抽出では、まず第 1 顔検出結果に対して所定の閾値処理を行い、その結果に基づいて顔候補位置を検出する．顔候補位置とは、存在すると思われる顔の候補の、顔の中心位置である．次にそれぞれの顔候補位置を基準とする所定の範囲において、中間出力結果の分布の分析を行い、顔候補位置ごとに複数の DOP を抽出する．この抽出した

DOP を用いて、顔候補の特定変動クラスを推定し、選択的なモジュール起動を行う。DOP 抽出の際に、中間出力結果の分布の分析を行う範囲は、検出した顔候補位置を中心とした、①顔候補上部、②顔候補下部、③顔候補全域である．①の範囲は目検出結果、②の範囲は口検出結果、③の範囲は、目検出結果、口検出結果、肌色検出結果、上辺エッジセグメント検出結果の、計 6 種の分布の分析を行う．分析により抽出する DOP は、各中間出力結果の分布の重心の、顔候補位置に対する相対座標と、範囲内の出力値総和である．上記 6 種の各分析により抽出した、重心相対座標(x,y)と、出力値総和 s の計 18 (6×3)の DOP を、それぞれの顔候補位置における DOP ベクトルとする．例として、面内回転±30° の変動顔画像の、①の範囲における目検出結果の分布の重心を Fig.3 に示す．

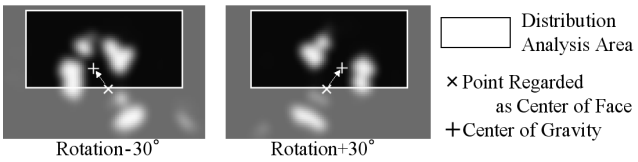

Fig.3：目検出結果の分布の重心位置比較

Fig.3 に示したように、抽出する DOP は変動顔画像ごとに特徴的であり、変動に適したモジュールの選択的起動が可能になる．

5.3 モジュールの選択

DOP ベクトルとマッチングを行う DOP ベクトルモデルは、各モジュールが対応する変動顔画像を用いて作成した．DOP ベクトルモデル作成では、モジュール番号 1、及び 2 (Table.1 参照)では、280 枚の変動顔画像、その他のモジュールでは 140 枚の変動顔画像を用い、まずその変動顔画像において 18 の DOP を抽出して、DOP それぞれの、モジュールに対応する変動顔画像ごとの平均、及び分散を算出し、これらモジュールごとの各 DOP の平均 $u_{i,j}$ と分散 $\sigma_{i,j}^2$ を各モジュールの DOP ベクトルモデルとして用いた．i はモジュール番号(1～20)、j は DOP の番号(1～18)である．

顔候補位置ごとの DOP ベクトルと、DOP ベクトルモデルのマッチングのスコアは、式(3)を用いて算出する．

$$(3)\quad Score_i = \sum_j \left[\frac{\sigma_{ALL,j}}{\sigma_{i,j}} \cdot \exp\left\{ \frac{-\left(DOP_j - u_{i,j}\right)^2}{2\sigma_{i,j}^2} \right\} \right]$$

式(3)中の $\sigma_{ALL,j}$ は、全顔候補位置の j 番目の DOP の標準偏差である．このスコアが所定値を超えた場合、そのモジュール番号のモジュールを選択的に起動する．

このように、顔候補位置ごとの DOP ベクトルと各モジュールの DOP ベクトルモデルとのマッチングにより、その顔候補位置において起動するモジュール(複数可)を選択することで、全モジュールを同時に起動した場合に発生する累積的な誤検出を低下させることができる．

5.4 性能評価

ソフトピアジャパン提供の顔画像データベースの内、学習に用いていない 100 人の 21 種の変動顔画像(2100 枚)と、同様に学習に用いていない、顔を含まない背景画像 100 枚に対して、第 1 顔検出(A)、全モジュール起動(B)、本手法(C)、の 3 手法における未検出率(FRR)と誤検出率(FAR)を評価した．評価結果を Table.3 に示す．Table.3 に示した数値は、FRR+FAR が最小となる閾値での結果を示している．FRR は 2100 枚の変動顔画像中で、顔が検出されなかった枚数の割合であり、FAR は 100 枚の背景画像中で、1 ヶ所でも誤検出が発生した画像の割合である．

	FRR [%]	FAR [%]
(A)	6.8	63.0
(B)	1.2	34.0
(C)	0.8	7.0

Table.3：FRR と FAR の評価

このように、第 1 顔検出(A)と比較して、複数モジュールを用いる(B)では、FAR を半減させることができ、さらに選択的モジュール起動により 34.0%から 7.0%と、大幅に FAR を低下させることができた．代表的な各方式での顔検出結果を Fig.4 に示す．

Fig.4：代表的な顔検出結果

(A)では誤検出、未検出ともに多く、(B)では(A)に比べ誤検出数は低下しているが、まだ誤検出があり、未検出も発生している．それに対し (C)では誤検出がなく、未検出も発生していない．このように(C)では誤検出が発生しにくいため、最終判定閾値を低く設定することができる．そのため(B)で未検出であった顔画像であっても、(C)では検出され、結果として FRR を低減することができる．

FRR と FAR の評価は、データベースの顔画像と背景画像を用いて行ったが、複雑な背景を有する一般的な環境下で撮影された画像においても、誤検出の少ない顔検出が可能であることを確認した．

複数のモジュールを用いることによる第 1 顔検出(A)に対する処理時間の増大は、(B)7%、(C)1%であり、第 2 顔検出による処理時間の増大はわずかであるといえる．また本手法の(C)方式では、選択的にモジュールを起動するため、第 2 顔検出での処理時間は、DOP を抽出する処理を含めても(B)の 1/7 程度である．

６．結論

撮影条件の変動ロバスト性が高く、且つ誤(未)検出率の低い顔検出手法を提案し、その有効性を確認した．各変動を複数のモジュールに分割対応させ、モジュール数の増加に伴う累積的に誤検出の発生を抑制するために、CNN の中間層出力分布に基づき、選択的にモジュールを起動する手法を導入した．これにより、顔のサイズ変動 8 倍、且つ回転変動±30° を許容可能なロバスト性を有し、顔検出性能において FRR1%、FAR7%以下であることを実証した．

文献

[1] Yang, M., Kriegman, D., Ahuja, N. “Detecting Faces in Images: A Survey”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.24, No.1, January 2002, pp.34-58
[2] Matsugu, M., Mori, K., Ishii, M., Mitarai, Y. “Convolutional Spiking Neural Network Model for Robust Face Detection”, Proceeding of 9th International Conference on Neural Information Processing, pp.660-664
[3] Le Cun, Y., Bengio, T. “Convolutional Networks for Images, Speech, and Time-Series”, In Michael A. Arbib, editor, The Handbook of Brain Theory and Neural Networks, MIT Press, Cambridge, Massachusetts, 1995, pp.255-258