2015 年度

早稲田大学大学院基幹理工学研究科情報通信専攻修士論文

マンガキャラクターを対象とした多視点顔検出の研究

2016.2.1

柳澤　秀彰

(5114F089-4)

所属　オーディオビジュアル情報処理研究室（渡辺裕教授）

Research on Multi-view Face Detection of Comic Characters

A Thesis Submitted to the Department of Computer Science and Communications Engineering,

the Graduate School of Fundamental Science and Engineering

of Waseda University

in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering

February 1st, 2016

By

Hideaki Yanagisawa

(5114F089-4) of

Advanced Multimedia Systems Laboratory

(Professor Hiroshi Watanabe)

目次

第1章　序論

1.1 研究の背景

近年，従来の紙媒体の書籍に代わって，タブレットやスマートフォンといったディジタル端末で購読する電子書籍の需要が増加している．2014年度における電子書籍市場規模は1411億円と推計され，前年度から398億円増加している[1]．その中でも電子コミックは重要なコンテンツであり，2014年度までの電子書籍市場の推移から電子書籍市場全体の約8割を占めるとされる．このことから，今後も電子書籍市場の規模の拡大が予想されるなかで電子コミックは大きな地位を担う存在といえる．

このような背景から，電子コミックの新たな機能について模索が行なわれている．従来の紙媒体のマンガにはない電子コミックのメリットとして，物理的な制約がないために，従来の書籍の枠にとらわれない表現が可能であるという点がある．例としては，マンガ内のキャラクター，台詞，コマ割りなどの情報をタグ付けすることによって，特定のキャラクターやシーンを基にアーカイブからマンガ作品の検索・掲示を行なう機能や，マンガ作品の要約を自動的に生成する機能が提案されている[2]．これらのユーザーセントリックな機能を提供することによって，今後の電子コミック市場に新たな価値を生み出すことが期待されている．このような機能の実現には，ディジタル化されたコミック画像のアーカイブにおいて，キャラクター・コマ割り・フキダシといったマンガの内容に関するメタデータを抽出し，それらを元の画像データと併せて保存することが必要となる．しかし，現状でこれらのメタデータを抽出するには，紙媒体のものをスキャンしてディジタル化した画像データより手作業で切り出してデータの抽出を行なう必要がある．このため，時間的なコストがかかることが実用化にあたって問題となっている．従って，作業を効率化するために，マンガ画像から自動的にメタデータを抽出する技術が必要である．

マンガにはコマやキャラクター，フキダシといった要素が重畳して構成されており，自然画像と比べて複雑性が高く，画像的な特徴は大きく異なっている．従って，マンガ画像に自然画像を対象とした一般的な画像処理手法をそのまま適応することは難しく，画像処理分野において独自な処理対象となっている．

現在，マンガ画像からコマ割りの情報を抽出する技術について，マンガの枠線を識別し，濃度勾配(intensity gradient)の方向を利用してコマの分割線を同定する手法[3][4][5]や，「マンガのコマは矩形であることが多い」という特徴を利用して，画像内から矩形領域を検出し，コマを特定する手法[6]が提案されており，いずれの手法でも80%を超える精度が報告されている．

また，フキダシを同定する技術について，画像内の文字領域をAda Boostを用いて特定し，その領域を基にフキダシの候補を検出し，SVMによってフキダシの形状を分

類する手法[7]が提案されており，この手法によって86%のフキダシを同定することが可能であると報告されている．

一方，マンガキャラクターの同定には，キャラクターの顔領域の候補を検出し，顔候補と予め作成したキャラクターの顔画像データベースとのマッチングを行なうことで，顔候補がどのキャラクターであるか同定する手法[8][9]が提案されている．マンガキャラクター顔検出に関して，従来研究よりHOG特徴量が特徴量記述子として有効であると報告されている．また，我々はマンガキャラクターのシーンごとの変化に対して，パーツに可変な検出モデルであるDPMの有効性を示した[10]．しかし，多様なマンガキャラクターに対して安定した検出を行なうことは未だに困難である．

近年の画像認識分野では，多層のニューラルネットワークを用いた機械学習手法であるディープラーニングが注目されている．2014年には，ディープラーニングのモデルの一つである畳み込みニューラルネットワーク(Convolutional Neural Network：CNN)を物体検出に応用したRegions with CNN features (R-CNN)が提案され，一般物体検出についてDPMなどの従来手法を上回る精度を示している．

本研究では，マンガ画像より横顔を含めた多視点顔検出を実現することを目的として，R-CNNと従来手法のDPMとの比較から，ディープラーニングのマンガ画像への有効性について検討する．

1.2 本研究の目的

本研究は，マンガ画像を対象とした高精度な顔検出・認識システムの実現を目的とする．画像から物体検出を行なうための基本的な操作は以下のようになる．まず，特徴量と呼ばれる特定の概念を特徴づける変数である画像から抽出する．次に，抽出された特徴量を機械学習によって生成された識別器に入力し，画像に対象物体が含まれるか否か判定する．

物体検出手法の代表例としては，認識率の低い弱識別器をCascade結合して一つの強力な識別器である強識別器を構成するViola-Jones法[11]や，物体を変形可能なパーツで構成されたモデルとして検出することで，物体の姿勢変化に頑健な検出を行なうDPM等が挙げられる．これらの手法では，Haar-Like特徴やHOG特徴といった，予め人間が設定した特徴量記述子によって特徴抽出を行なっている．

一方，ディープラーニングでは，入力されたデータを多層ニューラルネットワークに伝播させ，各層で学習を繰り返す過程でデータの識別に効果的な特徴量を自動的に決定する．動画像認識に一般的に利用されるモデルであるCNNは「画像データ全体から受け取れる意味は，これを構成する小さなパーツそれぞれが表す意味の組み合わせである」という概念に基づき，これらのパーツの中から元のデータをよく表すパーツ群・組み合わせを特徴量として導出する．R-CNNでは，CNNの特徴量を画像より切り出した物体の候補領域ごとに計算することによって，物体検出を行なう．しかし，マンガ画

像のような線画上の物体を対象とした実験は報告されておらず，マンガ画像に対しても自然画像と同様に有効な特徴量を導出できるかは未知である．

このような研究背景において，本研究では，R-CNN と DPM の比較から，ディープラーニングのマンガキャラクター検出に対する有効性の検討を目的とする．

1.3 本論文の構成

以下に本章以降の構成を示す．

第1章　本章であり，研究の背景およびその目的について述べている．

第2章　マンガキャラクター検出の現状について述べる．まず，画像処理におけるマンガ画像の特徴について述べる．次に，画像特徴量記述子である HOG 特徴の概要を述べる．そして，パーツに対して可変な物体検出手法である DPM の概要を述べる．そして，DPM の多視点顔検出への応用について述べる．最後に，マンガキャラクター検出におけるパートモデルの有効性を示す．

第3章　ディープラーニングを用いた物体検出法について述べる．まず，画像認識に用いられるニューラルネットワークのモデルである CNN の概要について述べる．次に，CNN を物体検出に応用した手法である R-CNN について述べる．そして，CNN の計算速度を改良した手法である Fast R-CNN について述べる．

第4章　マンガキャラクターを対象とした多視点顔検出手法の検討を行なう．まず，マンガ画像に最適な DPM の構成を実験より求める．次に，マンガキャラクターの多視点顔検出に対する DPM と R-CNN の検出率の比較を行い，R-CNN の優位性を示す．最後に，顔検出に有効な候補領域抽出手法について考察し，実験よりその性能を示す．

第5章　本研究の総括と今後の課題について述べる．

第2章　マンガキャラクター顔検出

2.1　まえがき

本章ではマンガキャラクター顔検出の従来手法について述べる．まず，画像処理におけるマンガ画像の特徴について述べる．次に，画像特徴量記述子である HOG 特徴の概要を述べる．そして，パーツに対して可変な物体検出手法である DPM の概要について述べる．そして，DPM の多視点顔検出への応用について述べる．最後に，マンガキャラクター検出におけるパートモデルの有効性を示す．

2.2　マンガ画像の特徴

マンガには極めて多様な形式が存在し，明確な定義を示すことは難しいが，大まかには以下のように定義される．

1. 視覚情報を絵として提示する(文章による説明ではない)．
2. 絵は話の展開を動的に描写し，情報の本質部分を占める(挿絵とは異なる)．
3. 視覚情報は人物のセリフは文字として，音が擬音として表現される．ただし，音楽は擬音ではなく絵やコマの行間のようなもので表現される場合が多い．
4. コマやフキダシなど独特の形式に沿っている．

本研究では，日本国内で出版される紙媒体のマンガを主な対象とする．マンガ画像の例を図 2.1 に示す．日本のマンガは基本的に「人物・背景・フキダシ・音喩・漫符・セリフ・その他の技法」から構成される．紙面はコマと呼ばれる枠によって分割されており，それぞれが一つの場面を現す．人物のセリフや思考はフキダシと呼ばれる枠の中に文字で書かれ，フキダシの形状や文字の書体によって語調を表す．擬音語・擬態語は，手書きの書き文字として絵の中に書かれることが多く，細々としたセリフなども書き文字で書かれることがある．漫符と呼ばれる一種の記号は，人物の心理や動作，ものの動きなどを明示的に表現する．

雑誌や単行本として刊行されるマンガは，カラーよりも 2 値のモノクロ画像のものが多い．このようなモノクロのマンガ画像は，白黒の 2 値からなる線画と，ベタと呼ばれる黒く塗りつぶされた領域，スクリーントーンと呼ばれる一定のパターンが印刷された領域の 3 つに分けることができる．

マンガ画像と自然画像との違いとして，マンガ画像では陰影の変化が省略されるため，画素間の輝度変化が大きい領域(エッジ成分)と輝度がほとんど変化しない平坦な領域が多い．また，マンガに登場するキャラクターは身体的な特徴や表情の変化などを誇張・強調し，簡易化・省略化して描かれることが多い．よって，マンガキャラクターは実際の人物よりも人物や登場シーンにおける形状的な変化が大きいといえる．

図 2.1: マンガ画像の例(文献[12]より引用)

2.3 Histograms of Oriented Gradients (HOG)

HOG 特徴量は，人物検出を目的として 2005 年に Dalal らによって提案された画像特徴量である[13]. 画像の局所領域の輝度の勾配方向をヒストグラム化した特徴量であり，幾何学的変換に強く，照明の変動に頑健であるという特長を持つ. 2012 年に石井らは，画像のエッジ成分に着目して特徴量抽出を行なう HOG 特徴量はエッジ成分を多く含むマンガ画像に対して有効であるとして，マンガキャラクター顔検出において Haar-Like 特徴量よりも高い検出精度を示したことを報告している[9]. HOG 特徴量の概要を図 2.2 に示す. HOG 特徴量の算出アルゴリズムは，1)輝度の勾配方向と勾配強度の算出，2)ヒストグラムの作成，3)ブロック領域による正規化の三つのステップからなる.

2.3.1 輝度の勾配方向と勾配強度の算出

画像の各ピクセルの輝度の値から勾配方向と強度を算出する. 勾配方向は式(2.1),

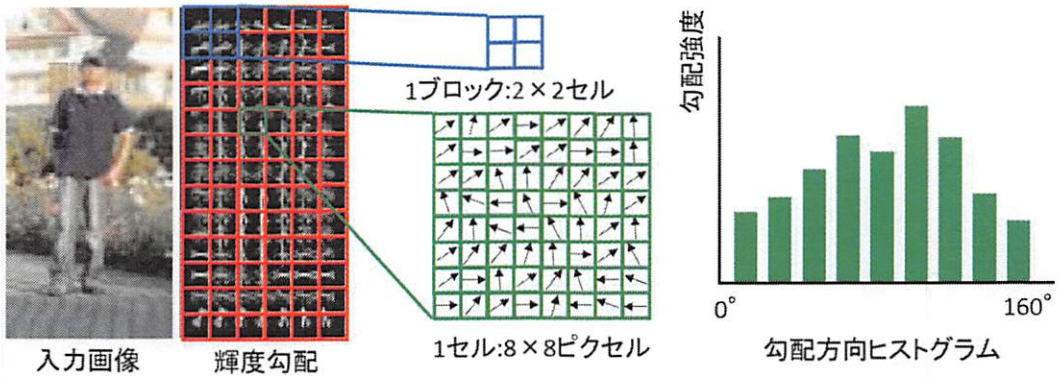


図 2.2: HOG 特徴量の概要(画像は文献[13]より引用)

勾配強度は式(2.2)より求められる．ここで，x, yはピクセルの座標を表し，$L(x, y)$は座標(x, y)のピクセルの輝度である．

$$m(x,y) = \sqrt{f_x(x,y)^2 + f_y(x,y)^2} \tag{2.1}$$

$$\theta(x,y) = \tan^{-1}\frac{f_y(x,y)}{f_x(x,y)} \tag{2.2}$$

$$\begin{cases} f_x(x,y) = L(x+1,y) - L(x-1,y) \\ f_y(x,y) = L(x,y+1) - L(x,y-1) \end{cases} \tag{2.3}$$

2.3.2 ヒストグラムの作成

勾配方向を 0°から 160°にかけて 20°ずつ 9 方向に分割する．次に，1 セルを 8×8 ピクセルからなる領域と設定し，1 セルにおける勾配強度のヒストグラムを作成する．

2.3.3 ブロック領域での正規化

1 ブロックを 2×2 セルからなる領域と設定する．あるn番目の HOG 特徴量を$v(n)$とすると，1 ブロックに含まれる HOG 特徴量の総和で正規化した値は式(2.4)によって表される．

$$v(n) = \frac{v(n)}{\sqrt{(\sum_{k=1}^{2\times2\times9} v(k)^2) + 1}} \tag{2.4}$$

2.4 Deformable Part Model

DPM は 2008 年に Felzenszalb らによって提案された物体検出手法である[14][15].

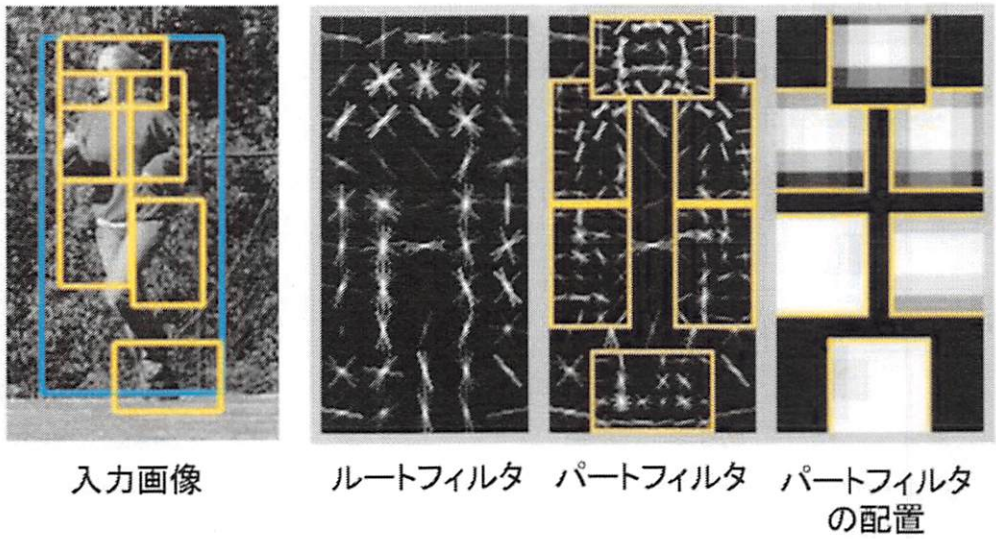


図 2.3: DPM の物体検出モデル(文献[15]より引用)

対象物体を複数のパーツから構成される検出モデルによって表現し，物体の全体および各パーツの HOG 特徴量とパーツの相対位置関係から物体を検出する．従来手法における検出器では，物体のパーツ位置は固定されており，人や動物といった物体を対象としたときに姿勢変化に対応できないといった問題があった．DPM では対象物体のパーツに対して可変であるため，姿勢の変化が大きい物体に対しても検出することが可能である．‘

2.4.1 検出モデル

DPM の物体検出モデルの例を図 2.3 に示す．DPM の物体検出モデルは，対象物体全体を捉えるグローバルなルートフィルタと，対象物体のパーツを捉える複数のパートフィルタから構成される．特徴量には HOG 特徴量を使用し，画像全体の検出ウィンドウをカバーするテンプレートにより算出される「疎な特徴」と，検出ウィンドウに対して可変なパートテンプレートにより算出される「密な特徴」の 2 つのスケールに対して適用される．

2.4.2 HOG ピラミッド

DPM では，ルートフィルタとパートフィルタについて，スケールの異なる HOG 特徴量を適用する．画像のスケールを変化させて解像度の異なる画像の集合であるイメージピラミッドを作成し，イメージピラミッドの各々のレベルの HOG 特徴量を計算することによって HOG 特徴ピラミッドを求める．イメージピラミッドと HOG ピラミッドの例を図 2.4 に示す．ここで，イメージピラミッドの上層では大域的に荒い HOG 特量

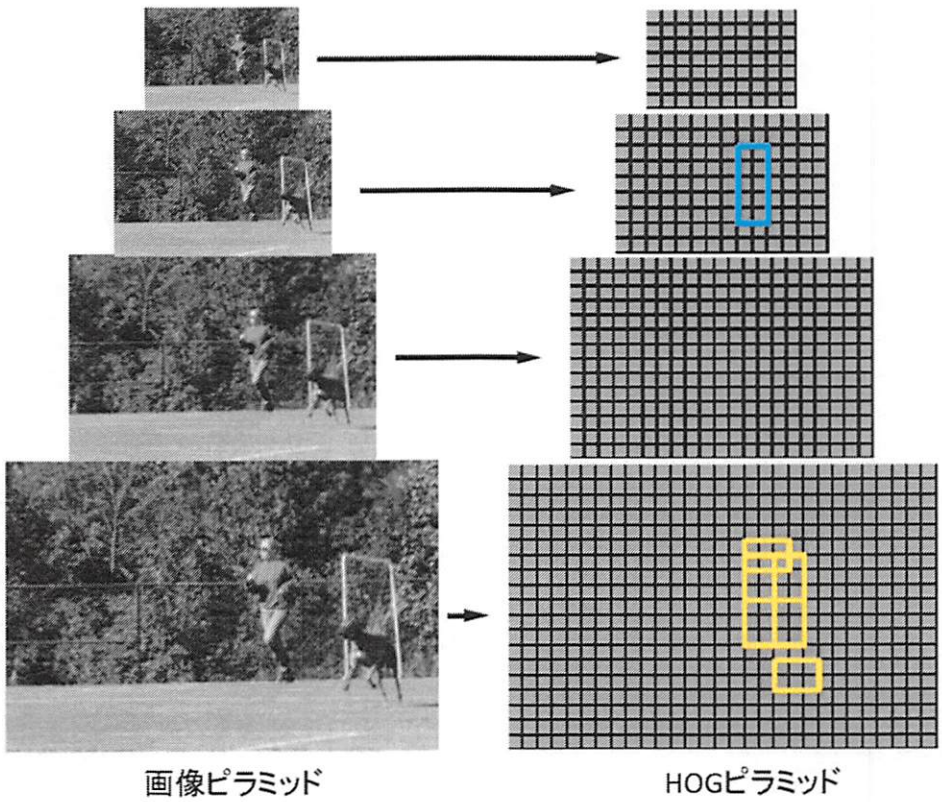


図 2.4: 画像ピラミッド(文献[15]より引用)

を捉え，下層では局所的に細かい HOG 特徴を捉える．

2.4.3 フィルタ

フィルタ Fは，入力された HOG 特徴に対する重みであり，w×h×9×4 個のベクトルで表現される．w と h は，サブウィンドウの高さと横幅を表す． HOG ピラミッドを H, セルの位置を $p=(x,y,l)$とする．ここで，lは，HOG ピラミッドのレベル（階層）である．取得された HOG 特徴量の強度は $\phi(H,p,w,h)$と示される．検出ウィンドウにおけるフィルタ Fのスコアは，重みを持ったベクトルと特徴量の内積 $F\cdot\phi(H,p,w,h)$によって表される．

2.4.4 可変パーツ

DPM の検出モデルにおいて，ルートフィルタは検出ウィンドウと同等と定義する．パートフィルタにおけるセルのサイズは，ルートフィルタのレベルにおけるセルのサイズの半分になるように設定する．このように，ルートフィルタのようなエッジを見るよりは，パートフィルタの高い解像度での特徴を見るほうが，局所的であり，高い認識性能を得ることができる．

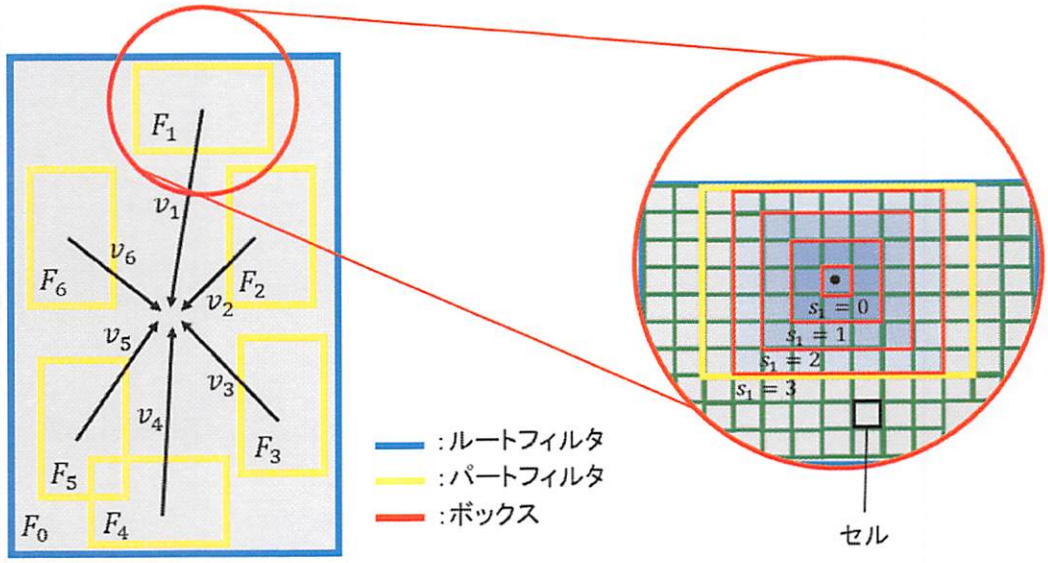


図 2.5: パートモデルの概要

n 個のパーツから構成される物体のモデルはルートフィルタF_0とパートモデル$(P_1,\cdots,P_n)$とで表される．このとき，$P_i = (F_i, v_i, s_i, a_i, b_i)$と表される．$F_i$は$i$番目のパートフィルタ，$v_i$はルートフィルタと$i$番目のパートフィルタの中心座標の相対的な位置関係を示す 2 次元ベクトル，s_iはi番目のパートフィルタの中心点を定める際の許容範囲を定める際の許容範囲を表すボックスのサイズ，a_i, b_iは，i番目のパートフィルタにおける 2 次元ベクトルによる係数を表す．パートモデルの概要を図 2.5 に示す．モデルの配置を潜在変数 zとし，$z = (p_1,\cdots,p_n)$とする．配置 zのスコアは，各フィルタのスコアと，パーツとルートの位置関係より，式(2.5)で与えられる．

$$score(z) = \sum_{i=0}^{n} F_i \cdot \phi(H, p_i) - \sum_{i=1}^{n} a_i \cdot (\tilde{x}_i, \tilde{y}_i) + b_i \cdot (\tilde{x}_i^2, y_i^2) \tag{2.5}$$

$$(\tilde{x}_i, \tilde{y}_i) = ((x_i, y_i) - 2(x, y) + v_i)/s_i \tag{2.6}$$

式(2.5)において，第一項目は，フィルタの重みと，HOG 特徴ベクトルの内積をとったフィルタのスコアの合計を表し，第二項目は，パートフィルタとルートフィルタの相対的な位置関係と距離を表す．式(2.4)は，i 番目のパートフィルタの中心座標(x, y)と，ルートフィルタの中心座標(x_i, y_i)と，v_i ，s_iを用いてパートフィルタの配置を表す$(\tilde{x}_i, \tilde{y}_i)$ を算出する．このとき，パートフィルタはルートフィルタの 2 倍の解像度を持っているので，距離関係を元に戻すために，(x, y)を 2 倍にして計算する．ここで，$\tilde{x}_i$と$\tilde{y}_i$は共に-1 から 1 の値をとる．また，式(2.5)での配置 z のスコアは，次式のβとψ(H, z)

の内積$\beta \cdot \psi(H,z)$で表すことができる．

$$\beta = (F_0, \cdots F_n, a_1, b_1, \cdots, a_n, b_n) \tag{2.7}$$

$$\psi(H,z) = (\phi(H,p_0), \phi(H,p_1), \cdots, \phi(H,p_n), \tilde{x}_1, \tilde{y}_1, \tilde{x}_1^2, \tilde{y}_1^2, \cdots, \tilde{x}_n, \tilde{y}_n, \tilde{x}_n^2, \tilde{y}_n^2) \tag{2.8}$$

2.4.5 検出

画像全体にかけてスライディングウィンドウを走査し，各ルート位置においてスコアを計算する．このうち，式(2.5)のスコアを最大化するパートフィルタの組み合わせを求め，スコアの値が閾値以上になった箇所を物体として検出する．

$$score(p_0) = \max_{p_1, \cdots, p_n} score(z) \tag{2.9}$$

$$score(p_0) = \max_{p_1, \cdots, p_n} \left(\sum_{i=0}^{n} F_i \cdot \phi(H,p_i) - \sum_{i=1}^{n} a_i \cdot (\tilde{x}_i, \tilde{y}_i) + b_i \cdot (\tilde{x}_i^2, \tilde{y}_i^2) \right) \tag{2.10}$$

また，各パートフィルタのスコアは独立に求めることができるため，それぞれのパートフィルタについて最大値を求めることによってルート位置のスコアの最大値を計算できる．

$$\text{score}(p_o) = F_0 \cdot \phi(H,p_i) + \sum_{i=1}^{n} \max_{p_i} F_i \cdot \phi(H,p_i) - \left(a_i \cdot (\tilde{x}_i, \tilde{y}_i) + b_i \cdot (\tilde{x}_i^2, \tilde{y}_i^2) \right) \tag{2.11}$$

2.4.6 学習

DPM の学習には，対象物体の位置をバウンディングボックスによって指定したポジティブサンプル画像と，検出対象を含んでいないネガティブサンプル画像によるデータセットを用いる．学習データセットを$D = (\langle x_1, y_1 \rangle, \cdots, \langle x_n, y_n \rangle)$として，$x_i$をサンプル画像，$y_i \in \{-1,1\}$はサンプル画像に対するラベルとする．また，HOG ピラミッドを$H(x_i)$，ルートフィルタとパートフィルタの有効な配置の範囲を$Z(x_i)$と示す．$Z(x_i)$はポジティブサンプルにおいて指定されているバウンディングボックスを元に決められ，最低でも50%の領域がバウンディングボックスに重なるように定義される．ポジティブサンプルは，パートの位置ごとと対象物体そのものの位置の両方を学習する．

2.4.7 Latent SVM

ポジティブサンプルおよびネガティブサンプルxのスコアは，式(2.11)で表される．

$$f_\beta(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x, z) \tag{2.12}$$

ここで，βはモデルのパラメータ，zはモデルの配置を表す潜在変数である．$\Phi(x, y) = \psi(H(x), z)$と置き換えることができるので，式(2.11)は式(2.6)と式(2.7)で示した配置のスコアの最大値をとることと等しい．よって，式(2.11)を最大化するようなβをポジティブサンプルの学習から得る．βやzといった潜在的な変数を使い，学習を行なうアプローチを Latent SVM と呼ぶ．Latent SVM では，まずβを固定し，$f_\beta(x)$を最大化するzを求める．次にzを固定して,通常の SVM のアルゴリズムよりβの最適化を行なう．この操作を繰り返すことによって，最適なβの値を求める

2.4.8 ハードネガティブの抽出

一般物体認識において，膨大な量の学習サンプルの大多数はネガティブサンプルとなる．一度に全てのネガティブサンプルを学習することは難しいため，ネガティブサンプルの中からより識別しづらいものをハードネガティブサンプルとして選び，ポジティブサンプルと，ハードネガティブサンプルから成る学習データを作成する．ハードネガティブサンプルの作成はDとβを用いて次(2.12)で表される．ハードネガティブサンプルは間違って識別されたネガティブサンプルの集合となる．

$$\mathrm{M}(\beta, D) = \{\langle x, y \rangle \in D | y f_\beta(x) \leq 1\} \tag{2.13}$$

2.4.9 学習の詳細

DPM の学習プロセスは以下のようになる．

1. ルートフィルタの初期化
   学習用サンプルで設定されたバウンディングボックスのアスペクト比をもとに，ポジティブサンプルをm個のグループに分類し，対応するルートフィルタの寸法を自動的に決定する．
2. ルートフィルタの初期化
   m個のグループに分類されたポジティブサンプルについて，左右の方向にクラスタリングを行う．潜在変数を持たない通常の SVM を用いて，1 つのグループについて対称となる 2 枚のルートフィルタF_0を学習する．ネガティブサンプルはデータセット内のネガティブ画像の中からランダムに決定したものを使用する．
3. ルートフィルタの更新
   学習された対称のルートフィルタを 1 つのコンポーネントとして扱い，バウンディングボックスに重なるように，式(2.5)のスコアが最も高くなるルートフィルタの位

置を探し出して更新する．その後，F_0を再学習する．

4. パートフィルタの初期化

   2.で学習されたルートフィルタより，ルートフィルタの80%以上を占めるようにn個のパートフィルタを配置する．パートフィルタの位置はHOG特徴量の値が最も高い位置から順に決定される．パートフィルタの初期の移動コストは，パートフィルタの係数a_i, b_iの初期値$a_i = (0,0)$，$b_i = -(1,1)$より求める．

5. 検出モデルの更新

   Latent SVMによって，新しいモデルを更新するため，学習データDにモデルの配置zを追加して，$\langle x_i, z_i, y_i \rangle$という形に再構築する．その後，バウンディングボックスに50%以上重なるように画像から検出を行なう．この中でバウンディングボックスの配置と最も一致しているものを採用する．これによってβも更新される．ハードネガティブサンプルには，対象物体ではないのに高いスコアを出したものを使用する．この学習をファイルサイズの限界まで10回繰り返し行なう．学習の過程において，ハードネガティブサンプルを記録し，メモリの限界の範囲内においてできるだけ多くの新しいハードネガティブサンプルを追加していく．

2.5 多視点顔検出へのDPMの適用

DPMを顔検出に適用した例として，2015年にOrozcoらはDPMを用いた多視点顔検出手法を提案している[16]．多視点顔検出は，顔の向きや隠れの存在に関係なく顔検出を行なう方法である．2004年にWuらは，Viola-Jones法による顔検出器を顔の向きや傾きに応じて複数作成し，それらを組み合わせることによって多視点顔検出を行なうといった手法を提案している[17]．また，2014年にZhuらは木構造モデルを使用した手法を提案している[18]．この手法では，目や鼻といった顔パーツを検出し，その位置情報をもとに顔の内側の構造をモデル化することによって顔の検出を行なう．しかし，顔パーツからモデルを生成するための計算量が膨大になることや，顔パーツを正確に検出するために解像度の高い画像が必要であるといった点が問題となっている．Orozcoらはこの問題に対して，より簡易な検出モデルとしてDPMを適用した方が多視点顔検出に有効であると主張している．

論文では，多視点顔検出に有効なDPMのルートフィルタ数と，パートフィルタ数について検討を行なっている．まず，正面・横に分類した4枚のルートフィルタを持つ検出器と，より詳細な角度によって分類した8枚，13枚のルートフィルタを持つ検出器の比較を行なった．Annotated Facial Landmarks in the Wild (AFLW)とFace Detection Database (FDDB)の2種類のデータセットについて検出を行なった結果を図2.6に示す．この実験結果から，ルートフィルタ数が4枚のとき最も検出率が高くなることが確認できる．また，パートフィルタを6枚使用した検出器と，20枚使用した検出器では，パートフィルタが6枚の方の検出率が高くなった．この結果について，

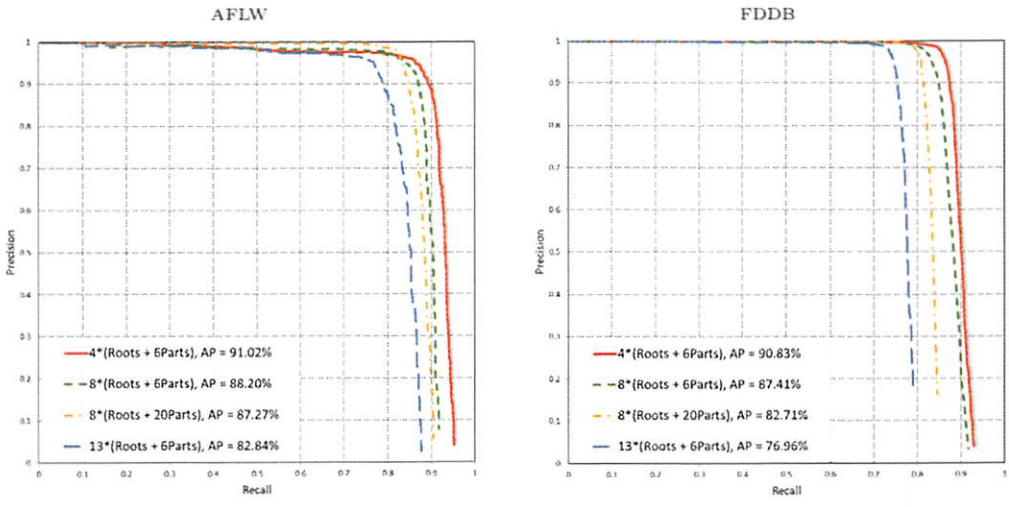


図 2.6: 多視点顔検出におけるフィルタ数の影響(文献[16]より引用)

Orozco らはルートフィルタとパートフィルタの数が多くなるほど Recall が上がり，Precision が下がるトレードオフの関係にあると説明している．

2.6 マンガキャラクター検出における DPM の有効性の検討

本節では，HOG 特徴量と SVM を使用する従来手法と DPM との比較から，マンガ画像からのキャラクター顔検出における DPM の有効性を示す．本実験では，従来手法は DPM の検出モデルのうちルートフィルタのみを使用する検出器と同等であるとして，パートフィルタも使用した検出器との比較を行なった．また，DPM のアルゴリズムには voc-release5[19]を使用した．

2.6.1 学習・テストに使用するデータセット

本節では，DPM の学習およびテストに使用したデータセットについて説明する．本実験では，複数のマンガ作品についてキャラクターの検出が行なえる検出器の作成を目的として，「ドラえもん」[20]，「ブラック・ジャック」[21]・「名探偵コナン」[22]・「SLAM DUNK」[23]の 4 作品に登場するキャラクターを無作為に選択したものを検出対象とした．元のマンガ画像 1 ページには大量のキャラクター顔領域が含まれる，アノテーションの指定が複雑になるため，本実験では顔領域と非顔領域について切り出した画像を使用した．ポジティブサンプルは，キャラクターの顔領域周辺を切り出して 200×200 ピクセルにリサイズした画像を使用し，顔領域のバウンディングボックスを記述するアノテーションを作成した．また，顔領域のうち両目が描かれている角度のものを「正面顔」，片目のみが描かれている角度のものを「横顔」，コマやオブジェクトによって顔の一部が隠れているものを「隠れ顔」と定義した．文献[12]のマンガ画像より，ポジティブサンプルの正面顔，横顔および隠れ顔の例を図 2.7，図 2.8，図 2.9 に示す．ここで図 2.7，

表 2.1: DPM 評価実験の学習に使用するマンガ画像

タイトル	ポジティブサンプル	ネガティブサンプル
	正面顔	
"ドラえもん"	100	
"ブラック・ジャック"	100	1000
"名探偵コナン"	100	
"SLAM DUNK"	100	
合計	400	1000

表 2.2: DPM 評価実験のテストに使用するマンガ画像

タイトル	ポジティブサンプル		ネガティブサンプル
	正面	隠れ	
“ドラえもん”	90	10	
“ブラック・ジャック”	90	10	800
“名探偵コナン”	90	10	
“SLAM DUNK”	90	10	
合計	360	40	800

図 2.8，図 2.9 において，赤枠で示された領域はバウンディングボックスで指定した領域を示している．ネガティブサンプルは，先述のマンガ作品からキャラクターの顔を含まない領域を無作為に切り出して 200×200 ピクセルにリサイズした画像を使用した．文献[12]のマンガ画像より，ネガティブサンプルの例を図 2.10 に示す．

本実験では，既存手法との比較を目的として，正面顔および隠れ顔を検出対象とした．学習およびテストに使用したデータセットの内容を表 2.1，表 2.2 に示す．学習セットは正面顔のみを含むポジティブサンプル 400 枚，ネガティブサンプル 1000 枚，テストセットは正面顔と隠れ顔を含んだポジティブサンプル 400 枚，ネガティブサンプル 800 枚とした．

2.6.2 DPM の設定

DPM のルートフィルタ数は，正面顔の左右に対応する 1 枚と設定した．DPM のパラメータは，パートフィルタの枚数を 8 枚，NMS を 0.5 として，その他のパラメータは voc-release5 のデフォルトの値を使用した． 実験に使用した DPM の各パラメータを表 2.3 に示す．

本実験における検出器の評価には，PASCAL VOC の Precision-Recall プロトコル[24]を適用した．顔として検出された領域と，アノテーションに記載されたバウンディ

ングボックスが 50%以上オーバーラップしているとき Ture Positive と判定される．また，検出された領域とバウンディングボックスとのオーバーラップが 50%未満のとき False Positive と判定される．さらに，バウンディングボックスで指定された顔領域のうち検出されなかったものは False Negative となる．Precision と Recall の値は，True Positive，False Positive，False Negative の個数より，それぞれ式(2.13)，式(2.14)から求められる．

$$Precison = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{2.14}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2.15}$$

式(2.10)の検出スコアに対する閾値を変動させて，テストセットに対する顔検出結果から Precision，Recall の値を算出し，Precision と Recall の変動を図示する．また，Precision の平均値から Average Precision (AP)を算出する．

2.6.3 実験結果

学習によって生成された DPM の検出モデルを図 2.11 に示す．図 2.11 において，(a)はルートフィルタの HOG 特徴量に対する応答を可視化したものを表す．また，(b)は各パートフィルタの 2 倍の解像度における HOG 特徴量に対する応答を可視化したものを表す．そして，(c)は 8 枚のパートフィルタの検出モデル内のデフォルトの配置を表す．さらに，DPM と従来手法との比較を図 2.12 に示す．横軸が Precision，縦軸が Recall の値を示し，実線が DPM の検出結果，破線が従来手法の検出結果を表している．実験結果より，DPM は従来手法を Precision，Recall ともに上回っており，AP において 11.7%上昇していることが確認できた．このことから，マンガキャラクター顔検出における DPM の有効性が示された．

2.7 むすび

本章では，マンガキャラクター顔検出の従来手法について述べた．まず，画像処理におけるマンガ画像の特徴について述べた．次に，画像特徴量記述子である HOG 特徴の概要を述べた．そして，パーツに対して可変な物体検出手法である DPM の概要を述べた．さらに，DPM の多視点顔検出への応用について述べた．最後に，DPM と従来手法との比較実験から，マンガキャラクター検出に対する DPM の有効性を示した．

第 3 章では，近年の機械学習にて注目を集めている手法であるディープラーニングについて述べ，ディープラーニングの物体検出法への適用について言及する．

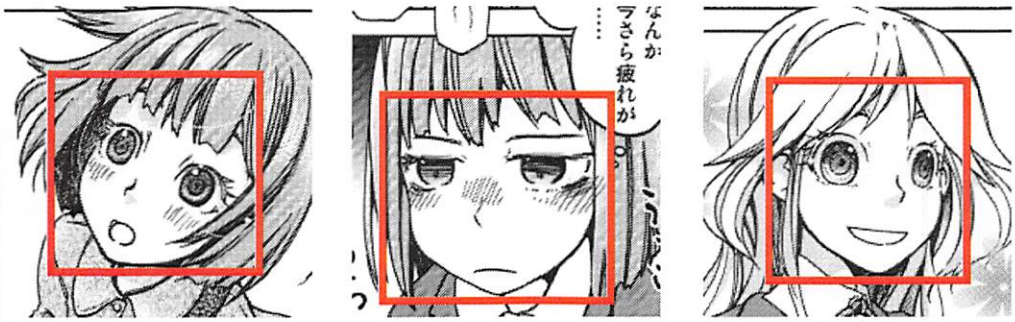


図 2.7:　正面顔の例(画像は文献[12]より引用)

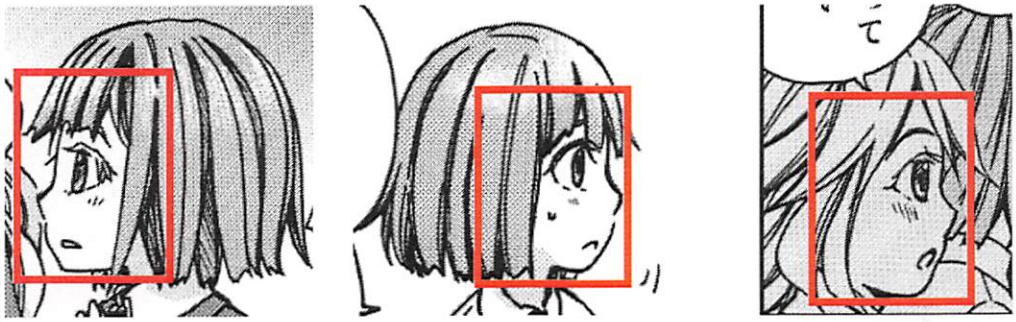


図 2.8:　横顔の例(画像は文献[12]より引用)

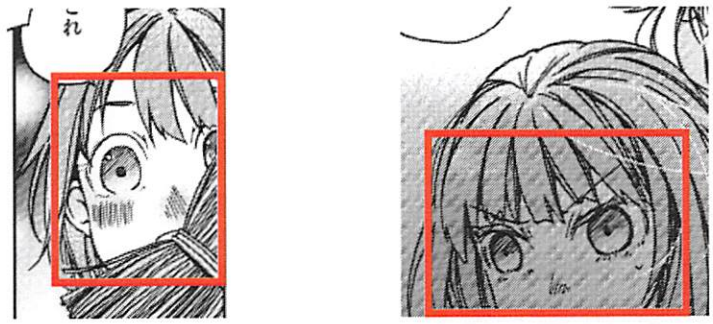
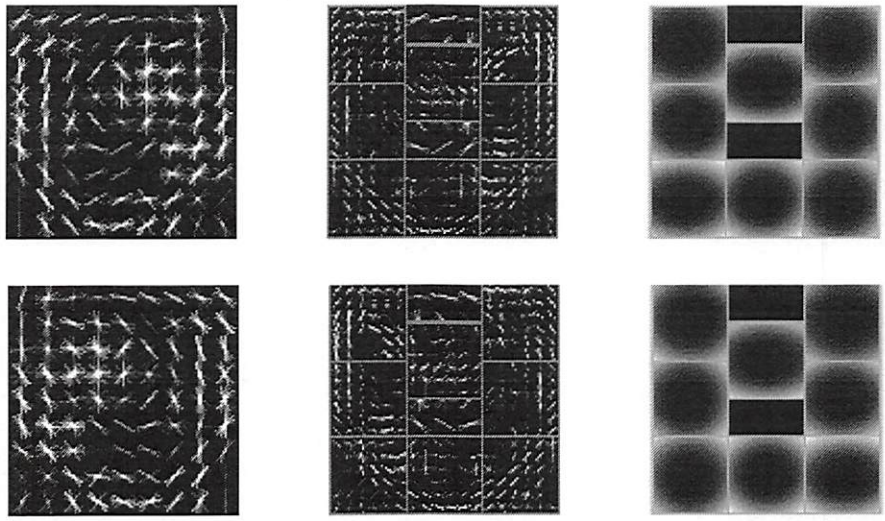


図 2.9:　隠れ顔の例(画像は文献[12]より引用)

図 2.10:　ネガティブサンプルの例(画像は文献[12]より引用)



(a) ルートフィルタ　(b) パートフィルタ　(c) パートフィルタの配置

図 2.11: マンガキャラクターの検出モデル

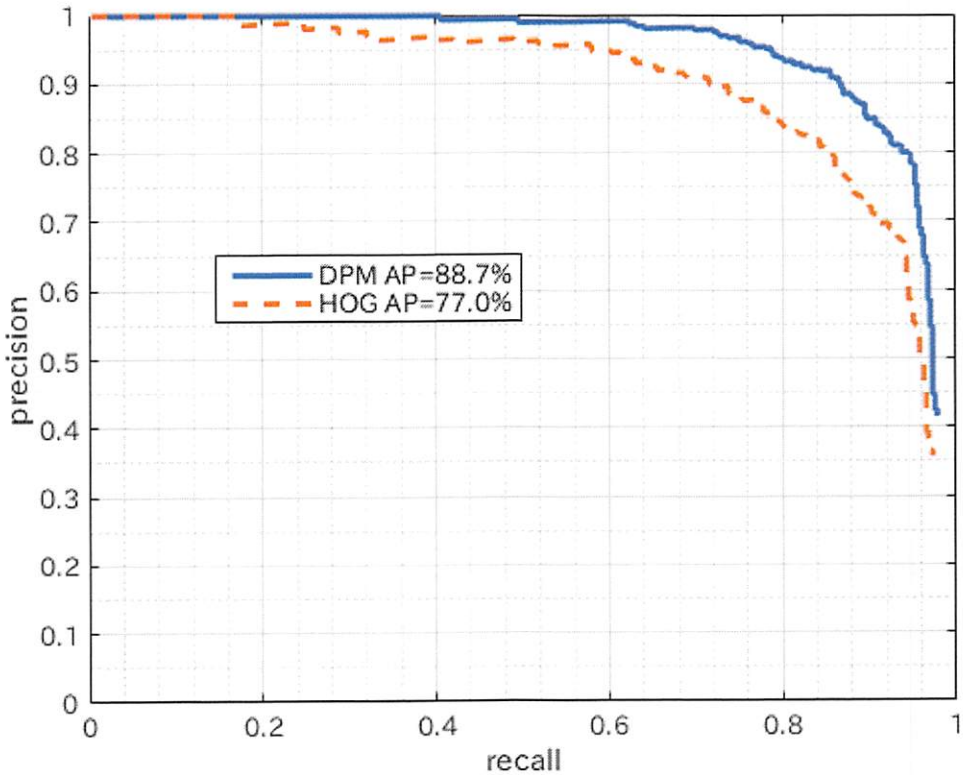


図 2.12: HOG と DPM の比較

第3章　ディープラーニングを用いた物体検出手法

3.1　まえがき

第2章では，マンガキャラクター検出における既存手法について述べた．本章では，ディープラーニングを用いた物体検出法について述べる．まず，ディープラーニングの考えの基となるニューラルネトワークについて述べる．次に，動画像認識に用いられるニューラルネットワークのモデルである CNN の概要について述べる．そして，CNN を物体検出に応用した手法である R-CNN とその改良手法である Fast Regions with CNN feature (Fast R-CNN)について述べる．最後に，CNN を多視点顔検出に適用した例である Deep Dense Face Detector (DDFD)について述べる．

3.2　ニューラルネットワーク

3.2.1　ニューロンモデル

ディープラーニングの考えの元となっているニューラルネットワークは，人間の神経細胞の学習のメカニズムをモデルに作られたアルゴリズムである．ニューラルネットワークのユニットの構造を図 3.1 に示す．あるニューロン$x_1, \cdots x_d$から 0 か 1 の入力信号を受け取り，その値に何らかの重み$w_1, \cdots w_d$を付加して総和を求める．出力yは式(3.1)のように表される．

$$y = f\left(\sum_{i=1}^{d} w_i x_i\right) \tag{3.1}$$

ユニットの出力yは，活性化関数fに入力信号x_iと重みw_iの総和を入力することで計算される，活性化関数には様々な種類があり，ニューラルネットワークには一般的にシグモイド関数が使われる．シグモイド関数はどんな入力に対しても 0 か 1 の値を出力する関数である．入力信号と重みの総和をXと表したとき，シグモイド関数は式(3.2)のように示すことができる．

$$f(X) = \frac{1}{1 + \exp(-gX)} \tag{3.2}$$

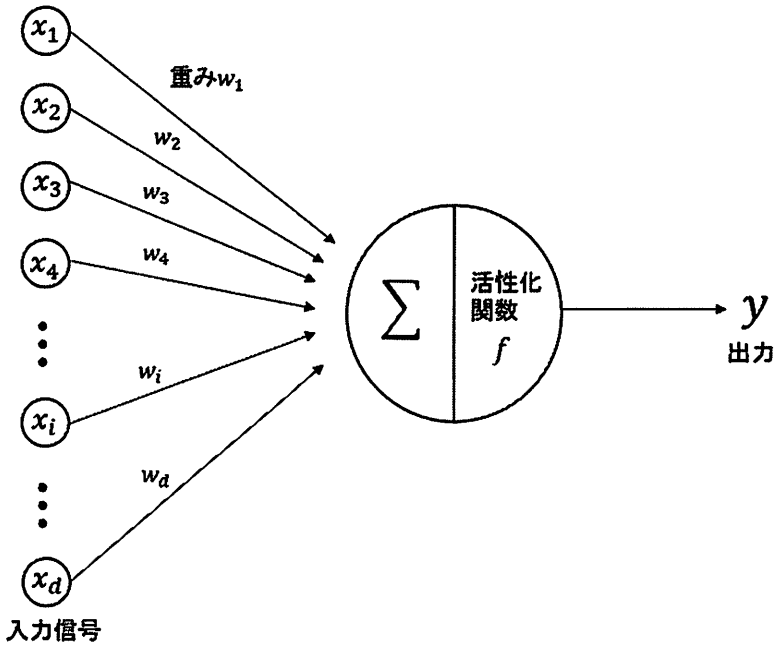


図 3.1: ニューラルネットワークのユニットの構造

ここで，式(3.2)のgはゲインを示す．ゲインgは，シグモイド関数の曲線の緩急を制御する関数である．シグモイド関数は比較的単純な非線形関数であり，微分の計算も容易である．一連の流れの中で重要になるのが重み付けであり，学習の過程で重みw_iを変化させ，最適な値を出力するように更新していくことで，精度を高めていく．

3.2.2 単純パーセプトロン

単純パーセプトロンは 1957 年に提案されたパーセプトロンモデルである．入力層と隠れ層，出力層の 3 層構造となっているが，入力層から中間層への重みの値は固定されているため，実質的には 2 層構造と見なすことがでる．単純パーセプトロンの構造を図 3.2 に示す．単純パーセプトロンでは通常のユニットと重みに加えてバイアスθを設定する必要があり，θの値も学習によって更新する．d次元の入力層のユニット$x = [x_1, x_2, \cdots, x_i, \cdots, x_d]$があるとき，各成分をノードとして見て，これらを重みベクトル$w = [w_1, w_2, \cdots, w_i, \cdots, w_d]$で線形結合して出力$y$を得る．よって出力$y$は式(3.3)で表すことができる．

$$y = f\left(\sum_{i=1}^{d} w_i x_i - \theta\right) \tag{3.3}$$

重みとバイアスの更新は，出力yと教師信号tを用いて式(3.4)と式(3.5)によって表さ

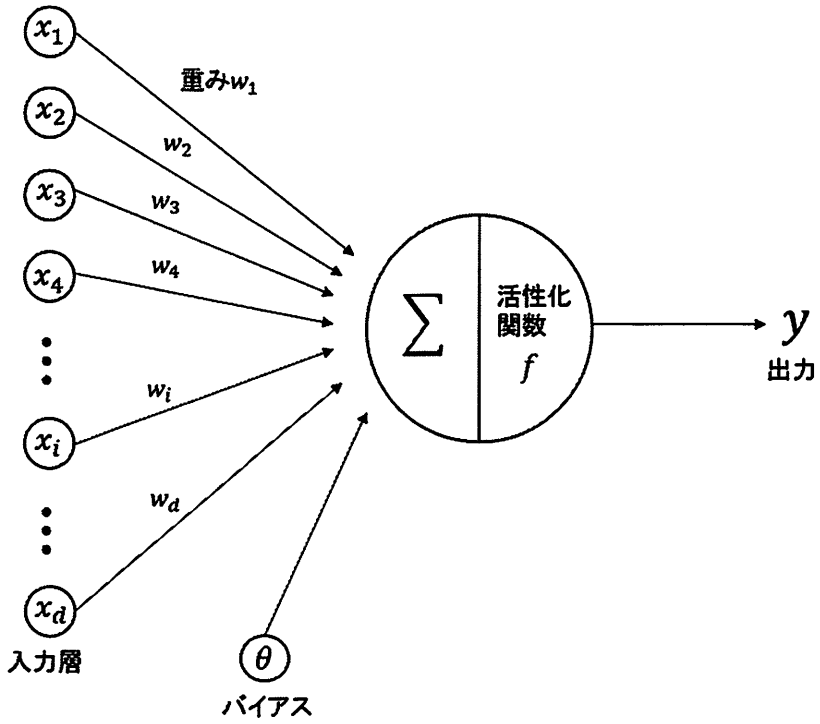


図 3.2: 単純パーセプトロンの構造

れる.

$$w_{t+1} = w_t + \eta(t - y)x \tag{3.4}$$

$$\theta_{t+1} = \theta_t + \eta(t - y) \tag{3.5}$$

式(3.4)と式(3.5)のtは更新回数を示し,ηは学習係数と呼ばれる重みの更新量を定める定数であり，$0 < \eta \leq 1$である．単純パーセプトロンでは式(3.4)と式(3.5)の更新を全てのサンプルに対して行なう.単純パーセプトロンの学習は学習の終了条件を満たすまで繰り返される．学習の終了条件は，一般的に更新回数が指定した回数に達したときや，誤識別率が一定の値を下回ったときに設定される.

3.2.3 多層パーセプトロン

多層パーセプトロンは非線形の多クラス識別器であり，入力層，隠れ層，出力層の 3 層で構成される．多層パーセプトロンの構成の例を図 3.3 に示す．それぞれの層を構成する各ユニットは結合するユニットの重みを通して前の層からの入力を受けとる.ユニットに繋がる全ての入力に対応する重みの総和を出力関数に通したものがそのユニットの出力となる.多層パーセプトロンと単純パーセプトロンの大きな違いは 2 点挙げら

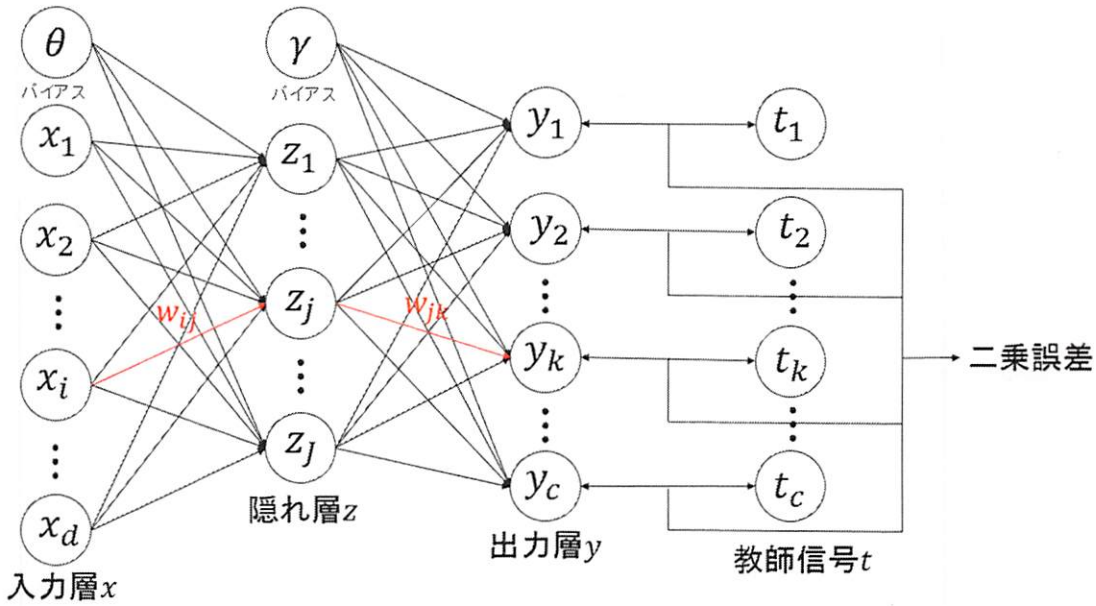


図 3.3: 多層パーセプトロンの構造

れる. まず, 単純パーセプトロンでは入力層と隠れ層の重みは一定の値となっているが, 多層パーセプトロンでは全ての重みに対して学習で更新を行なう. また, 多層パーセプトロンでは多クラス識別を行なうために出力層のユニットはクラス数cだけ用意する. 入力層のユニット数は入力次元数dと同じに設定し, 隠れ層のユニット数は任意の数Jとする. 多層パーセプトロンの学習は, 教師付き学習による誤差逆伝播法を用いた勾配降下最適化法によって行なわれる. 勾配降下最適化法は, 以下の 3 種類に分類することができる.

1. 最急降下法

   最急降下法では, 全ての学習サンプルを一度に用いてパーセプトロンの各パラメータの更新を行なう. まず, 全てのサンプルの学習誤差を求める. 誤差関数として二乗誤差を用いる場合, データセット数をNとしたとき, 学習誤差は式(3.6)で表される.

$$E_N = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{c}(y_k - t_k)^2 \tag{3.6}$$

   この誤差関数E_Nを用いて, 式(3.9)よりパーセプトロンの各パラメータの更新を行なう. 重みの更新量は誤差関数E_Nの勾配を算出して学習係数ηをかけたものを重みの更新量とする. 最急降下法では, 全ての学習サンプルを一度に用いるため, 誤差

関数の減少値が最大となる方向にパラメータが更新される.

$$w^{t+1} = w^t - \eta \frac{\partial E_N}{\partial w^t} \tag{3.7}$$

2. 確率的勾配降下法

最急降下法では，全ての学習サンプルについて誤差関数の総和を計算するため，学習サンプルの量が増えると計算量が増加することが問題となる．確率的勾配降下法は，一つの学習サンプルを用いてパーセプトロンの核パラメータを更新する手法である．学習サンプルが増えても計算量が増加しないため，ニューラルネットワークのような大量の学習サンプルを使用する検出器に対して有効である．確率的勾配降下法の誤差関数E_nは式(3.10)より得られる．

$$E_N = \frac{1}{2}\sum_{k=1}^{c}(y_k - t_k)^2 \tag{3.8}$$

重みの更新量は，最急降下法と同様に誤差関数E_nの勾配を算出して学習係数ηをかけたものを重みの更新量とする．各パラメータの更新式は式(3.11)で表される．

$$w^{t+1} = w^t - \eta \frac{\partial E_n}{\partial w^t} \tag{3.9}$$

3. ミニバッチ学習法

ミニバッチ学習法は，1度に複数の学習サンプルを用いて学習を行なう手法である．ミニバッチ学習法は，確率的勾配降下法と比べてパラメータの更新回数を削減することが可能であり，最急降下法と比べて計算量を削減できるメリットがある．バッチサイズをMとしたとき，学習誤差E_mは式(3.12)，パラメータの更新式は式(3.13)のようになる．

$$E_M = \frac{1}{2}\sum_{m=1}^{M}\sum_{k=1}^{c}(y_k - t_k)^2 \tag{3.10}$$

$$w^{t+1} = w^t - \eta \frac{\partial E_M}{\partial w^t} \tag{3.11}$$

パーセプトロンの学習では，学習誤差を用いて各層の重みおよびバイアスの更新量を求める．この更新量を求める方法として誤差逆伝播法を用いる．誤差逆伝播法のアルゴリズムは以下のようになる．始めに，入力ベクトルを順伝播し，隠れ層と出力層の出力を求める．次に，求めた出力と教師信号から誤差を求める．そして，算出した誤差から各パラメータの更新量を求め，勾配降下最適化法によって各パラメータを更新する．パーセプトロンに入力される特徴次元数をn，識別するクラス数をmとする．入力層のユニットをx_i，隠れ層のユニットをz_j,出力層のユニットをy_k，教師信号をt_kとし，隠れ層と出力層の活性化関数fはシグモイド関数を使用する．また，入力層と隠れ層の重みをw_{ij}，隠れ層と出力層の重みをw_{jk}とする．さらに，隠れ層のバイアスをθ_j，出力層のバイアスをγ_kとしたとき，隠れ層のユニットの出力は式(3.6)，出力層のユニットの出力は式(3.7)となる．

$$z_j = f\left(\sum_{i=1}^{d} w_{ij}x_i + \theta_j\right) \tag{3.12}$$

$$y_k = f\left(\sum_{j=1}^{c} w_{jk}z_j + \gamma_k\right) \tag{3.13}$$

今回の例では，確率的勾配降下法によって各パラメータの更新を行なう．誤差関数E_nは，式(3.10)より，式(3.16)のように表される．また，誤差関数の出力y_kによる微分は，式(3.17)のような出力層における教師信号t_kとの誤差δ_kで表される．

$$E_n = \frac{1}{2}\sum_{k=1}^{c}(y_k - t_k)^2 \tag{3.14}$$

$$\begin{aligned}\frac{\partial E_n}{\partial y_k} &= (y_k - t_k) \\ &= \delta_k\end{aligned} \tag{3.15}$$

出力層と隠れ層の誤差関数E_{njk}の勾配∇E_{njk}は，出力層のユニットy_kの内部ポテンシャルを$p_k = \sum_k w_{jk}z_j + \gamma_k$として，偏微分の連鎖法則より式(3.18)のように求めることができる．

$$\begin{aligned}\nabla E_{njk} &= \frac{\partial E_n}{\partial w_{jk}} \\ &= \frac{\partial E_n}{\partial y_k} \cdot \frac{\partial y_k}{\partial w_{jk}} \\ &= \frac{\partial E_n}{\partial y_k} \cdot \frac{\partial y_k}{\partial p_k} \cdot \frac{\partial p_k}{\partial w_{jk}} \\ &= \delta_k \cdot y_k \cdot (1 - y_k) \cdot z_j\end{aligned} \tag{3.16}$$

また，隠れ層から入力層の誤差の勾配∇E_{nij}も，同様に偏微分の連鎖法則を用いて式(3.19)のように求めることができる．

$$\begin{aligned}\nabla E_{nij} &= \frac{\partial E_n}{\partial w_{ij}} \\ &= \frac{\partial E_n}{\partial z_j} \cdot \frac{\partial z_j}{\partial p_j} \cdot \frac{\partial p_j}{\partial w_{ij}} \\ &= \frac{\partial E_n}{\partial y_k} \cdot \frac{\partial y_k}{\partial p_k} \cdot \frac{\partial p_k}{\partial z_j} \cdot \frac{\partial z_j}{\partial p_k} \cdot \frac{\partial p_j}{\partial w_{ij}} \\ &= \left(\sum_k \delta_k \cdot y_k \cdot (1 - y_k) \cdot w_{jk} \right) \cdot z_j \cdot (1 - z_j) \cdot x_i\end{aligned} \tag{3.17}$$

そして，式(3.18)と式(3.19)を用いて各層間のパラメータの更新式を，確率的勾配降下法によって設計する．出力層と隠れ層の重みの更新式は，式(3.11)に式(3.18)を代入することで，式(3.20)のように求められる．またバイアスの更新式は，式(3.21)のようになる．

$$w_{jk}^t = w_{jk}^t - \eta \cdot \delta_k \cdot y_k \cdot (1 - y_k) \cdot z_j \tag{3.18}$$

$$\gamma_k^t = \gamma_k^t - \eta \cdot \delta_k \cdot y_k \cdot (1 - y_k) \tag{3.19}$$

同様にして，隠れ層と入力層の重みとバイアスの更新式は式(3.22)と式(3.23)のように求められる．

$$w_{ij}^t = w_{ij}^t - \eta \cdot \left(\sum_k \delta_k \cdot y_k \cdot (1 - y_k) \cdot w_{jk} \right) \cdot z_j \cdot (1 - z_j) \cdot x_i \tag{3.20}$$

$$\theta_j^t = \theta_j^t - \eta \cdot \left(\sum_k \delta_k \cdot y_k \cdot (1 - y_k) \cdot w_{jk} \right) \cdot z_j \cdot (1 - z_j) \tag{3.21}$$

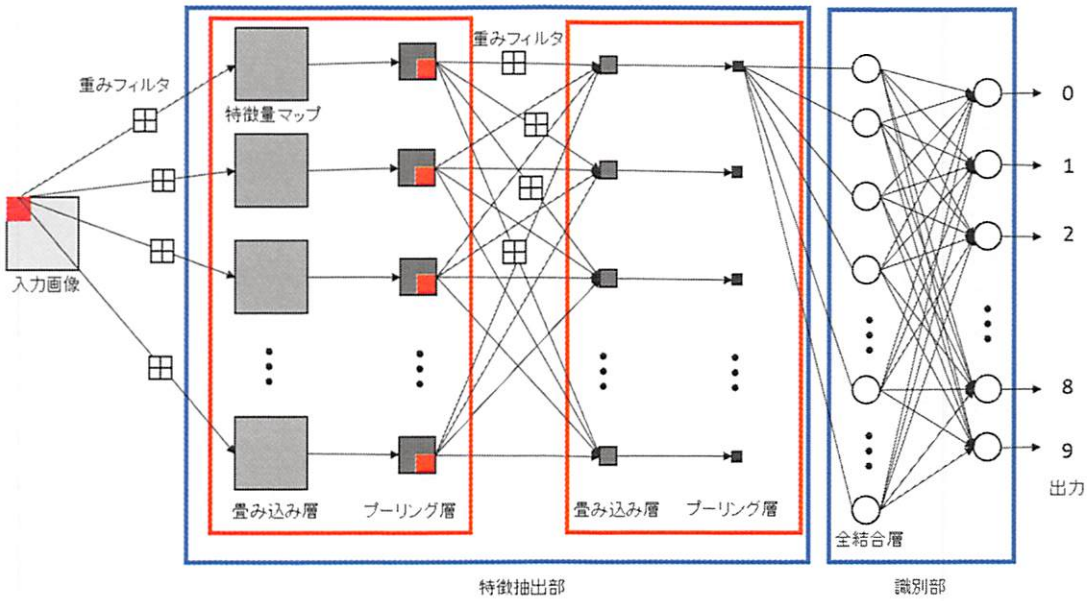


図 3.4: CNN の処理の流れ

多層パーセプトロンの学習では，各パラメータの学習を全ての学習サンプルに対して行なう．そして，全ての学習サンプルに対して各パラメータを更新したとき，学習の終了条件を満たしている場合には学習を終了し，満たさない場合には初めから学習サンプルの学習を行なう．

3.3 Convolutional Neural Network

Convolutional Neural Network (CNN)は，多層パーセプトロンの一つで，脳の視覚情報処理を模した構造のニューラルネットワークである．CNN は，複数の隠れ層を用意して畳み込みとプーリングの処理を繰り返し行なうことによって，特徴量を自動的に取得する．従来の多層パーセプトロンでは，各層間で重みが全結合しているため，隠れ層が増えると誤差の勾配が拡散してしまうという問題がある．この問題に対して，CNN ではユニット間の結合を局所に限定し，層間の結合を疎にすることで，複数の隠れ層がある場合にも学習を行なうことを可能にしている．

CNN の学習は，教師付き学習を前提とし，誤差逆伝播法を用いた勾配降下最適化法で学習する．図 3.4 に CNN の処理の流れを示す．CNN の処理は多段接続された複数の処理ユニットを通して行なわれる．各ユニットの入出力は，特徴量マップと呼ばれる複数枚の二次元画像となる．まず，入力画像に対して重みフィルタの畳み込み処理を行い，特徴量マップとして出力する．次に，出力された特徴量マップを入力としてプーリング処理を行い，新たな特徴量マップを得る．この処理を繰り返すことによって特徴量を自動生成する．入力に近い層ではエッジや線などの単純なパーツが抽出され，それら

が畳み込みとプーリングを繰り返すことで特徴同士がまとめ上げられ，顔や物などの複雑で抽象的な特徴量が生成される．最後に得られた特徴量マップを入力として識別部に入力し，識別を行なう．

3.3.1 畳み込み層

畳み込み層では，入力画像または特徴マップに対して重みフィルタとの内積をとり，重みフィルタをスライドさせて繰り返し畳み込みを行なうことで複数の特徴マップを出力する．フィルタの重みは，誤差逆伝播法による勾配降下最適化法によって自動的に学習される．畳み込み処理において，画像と重みフィルタのサイズをそれぞれ$n_x \times n_y$，$n_w \times n_w$としたとき，出力される特徴マップのサイズn'_x，n'_yは式(3.22)のようになる．

$$\begin{aligned} n'_x &= n_x - 2[n_w/2] \\ n'_y &= n_y - 2[n_w/2] \end{aligned} \tag{3.22}$$

また，畳み込み層で複数のフィルタを使用することによって入力画像のさまざまな特徴を捉えることが可能となる．

3.3.2 プーリング層

プーリング層は，畳み込み層の直後に置かれ，入力された特徴量マップの小領域から値を出力して新たな特徴量マップに変換する処理を行なう．プーリングを行なう目的は二つある．まず，プーリングによってユニット数が削減されるため，調整するパラメータを削減することができる．また，ある小領域から応答を出力するため，画像のどの位置でフィルタの応答が強かったこという情報を一部捨てることで，画像内に現れる特徴の微小な位置変化に対する応答の普遍性を得ることができる．プーリング処理は畳み込み層の隣接している2×2ユニットについて行なう．プーリング処理の流れの例を図3.5に示す．プーリング処理には以下の3種類がある．

1. 最大プーリング
   最大プーリングは小領域$(p, q) \in P_{ij}$内のユニットの出力y_{pq}の最大値を出力するプーリングである．最大プーリングの出力$\widetilde{y_{ij}}$は，式(3.23)から求めることができる．

$$\widetilde{y_{ij}} = \max_{(p,q)\in P_{ij}} y_{pq} \tag{3.23}$$

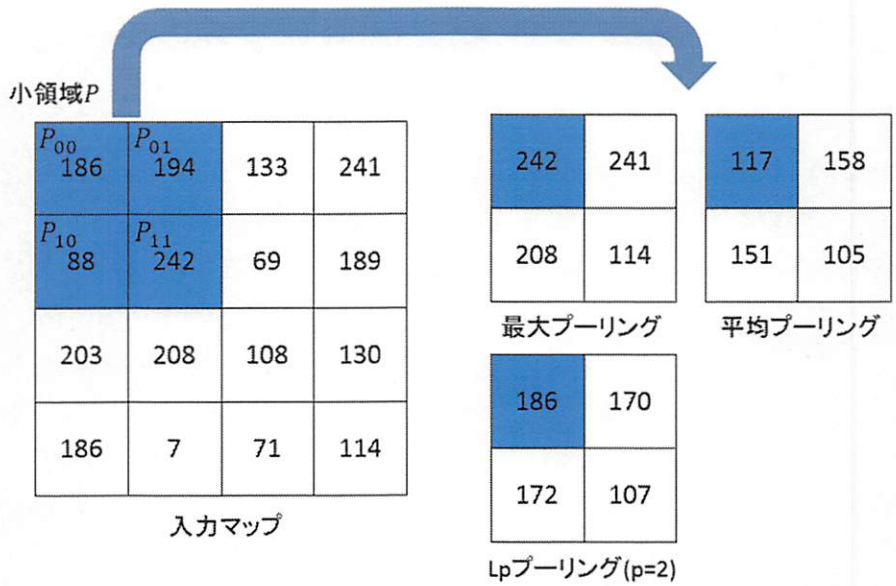


図 3.5: プーリング処理の流れ

2. 平均プーリング
平均プーリングは小領域P_{ij}内のユニットの出力y_{pq}の平均値を出力するプーリングである．平均プーリングの出力$\widetilde{y_{ij}}$は，式(3.24)から求めることができる．

$$\widetilde{y_{ij}} = \frac{1}{|P_{ij}|} \sum_{(p,q)\in P_{ij}} y_{pq} \tag{3.24}$$

3. Lp プーリング
Lp プーリングは，最大プーリングと平均プーリングを中間的な存在であり，小領域P_{ij}内のユニットのp乗平均偏差を出力するプーリングである．例えば$p = 2$のときは二乗平均値が出力される． Lp プーリングの出力$\widetilde{y_{ij}}$は，式(3.25)で表現される．

$$\widetilde{y_{ij}} = \left(\frac{1}{|P_{ij}|} \sum_{(p,q)\in P_{ij}} y_{pq}^{p} \right)^{\frac{1}{p}} \tag{3.25}$$

プーリング処理によって出力される特徴量マップのサイズは式(3.26)のようになる．

$$\begin{aligned} n'_x &= n_x/2 \\ n'_y &= n_y/2 \end{aligned} \tag{3.26}$$

3.3.3 全結合層

CNN では最後に全結合した多層パーセプトロンを配置した識別部によって認識を行なう．特徴抽出部の畳み込みとプーリングの処理より自動生成された特徴量マップを，全結合層のユニットに入力する．このとき，最終的に生成された特徴量マップと全結合層のユニットの層間の重みは全結合されている．その後，従来の多層パーセプトロンと同じように出力層のユニットに応答値が入力して識別する．

3.3.4 ユニットの構成

CNN のユニットの構成について説明する．まず，入力層と出力層のユニット数は多層パーセプトロンと同様である．隠れ層では，畳み込みとプーリングによってユニット数がそれぞれ式(3.22)と式(3.26)のように変化する．畳み込みでは$n_w \times n_w$の重みフィルタで畳み込み処理をするため，$n_w \times n_w$の範囲のユニットから 1 つのユニットに応答値を出力する．その後，その後，2×2の小領域でプーリングを行い 1 つのユニットに応答値を出力する．識別部では，特徴抽出部で抽出した特徴から識別部の全結合層のユニットと全結合し，識別する．このような構造から，CNN の層間は結合が疎であるといえる．

3.4 Regions with CNN feature

CNN を用いた物体検出アルゴリズムとして，2015 年に Girshick らは R-CNN を提案した[25]．R-CNN は，入力画像から物体の候補領域の抽出を行い，抽出されたそれぞれの候補領域を CNN に入力することで特徴量の計算を行い，物体の判定を行なう手法である．R-CNN の検出処理の流れを図 3.6 に示す．

3.4.1 Selective Search

画像から物体認識を行なう領域を切り出すために，従来ではスライディングウィンドウと呼ばれる手法が用いられる．スライディングウィンドウは，様々なサイズ・アスペクト比の矩形領域について，画像全体をスライドされていき，領域の切り出しを総当り的に行なう手法である．しかし，スライディングウィンドウには処理対象となる領域が非常に多くなることや，対応できる形状やサイズに制限があるといった問題がある．そこで，画像から物体領域の候補となる場所を検出するアルゴリズムによって領域の切り出しを行なうことにより，計算量を削減させる手法が提案されている．R-CNN は Uijlings らによって提案された Selective Search[26]と呼ばれるアルゴリズムを利用して候補領域の抽出を行なう．Selective Search によって複数のスケールの画像についてセグメンテーションと候補領域抽出を行なった結果を図 3.7 に示す．図 3.7 において，上の画像はセグメンテーションの結果を示し，下の画像は画像から候補領域として抽出

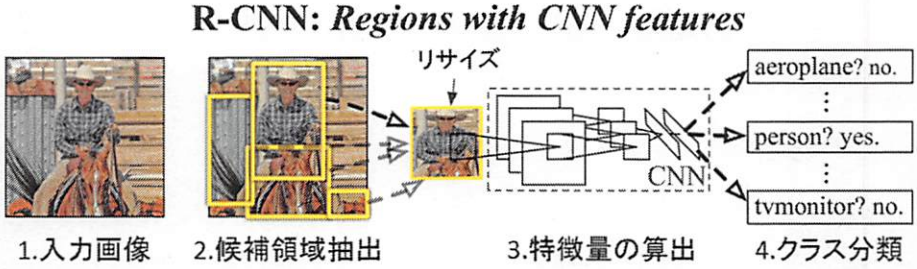


図 3.6: R-CNN の検出処理の流れ(文献[25]より引用)

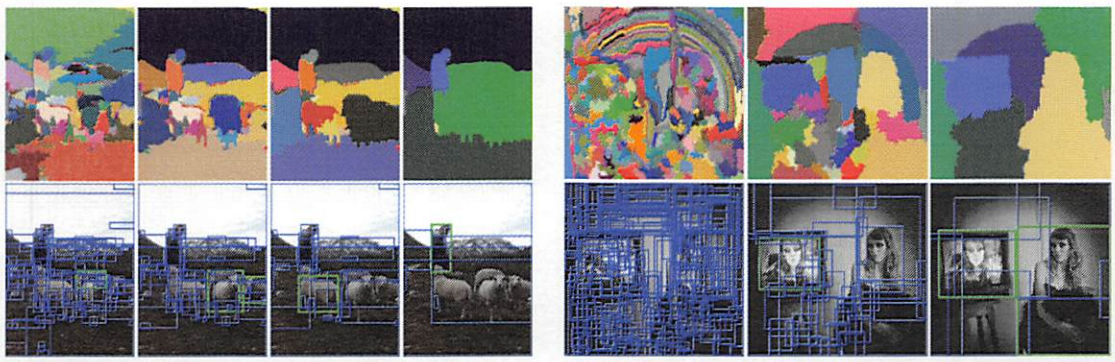


図 3.7: Selective Search によるセグメンテーションと候補領域抽出(文献[26]より引用)

された領域を示す．また，下の画像の緑枠は抽出された候補領域のうち，正しく物体の物体を検出している領域を示す．

Selective Search は，ボトムアップ型の階層的セグメンテーションによって，あらゆる位置やスケールに対応した候補領域の抽出が可能である．まず，Efficient Graph-Based Image Segmentation[27]と呼ばれるアルゴリズムによって初期のセグメンテーションを行なう．このアルゴリズムは，画像中の各画素を 1 つのノードとした木から，輝度が類似なノードを纏めていくことでセグメンテーションを行なう．次に，セグメンテーションによって作成された各小領域について，色特徴・テクスチャ特徴・小領域の面積・小領域の外接矩形の四つの特徴を複合した特徴量を算出する．そして，特徴量の類似度が最も高い近接領域を統合し，小領域の外接矩形を候補領域として取り出す．この統合処理を 1 枚の画像となるまで繰り返す．最終的に，2000 個程度の候補領域が画像から抽出される．

3.4.2 特徴量の抽出

候補領域を CNN に入力して特徴量の計算を行なう．周辺領域の情報を付け加えるために，3.3.1 節で検出された候補領域より少し大きい領域（リサイズ後のサイズで周囲 16 画素分）を 227×227 画素にリサイズし，CNN に入力する．CNN の 7 層目の全

結合層から出力される特徴量は4096次元の特徴ベクトルとなる.

ディープラーニングでは，予め大規模なデータセットについて学習済みの状態から目的とする別のデータセットへ学習し直すことによって，過学習を防ぐ効果があることが知られている．そこでR-CNNでは，ImageNetで教師付き事前学習を行なったニューラルネットワークモデルをベースに，実際に評価に利用するデータベースについて詳細な学習を行なう.

3.4.3 SVMによる物体検出

R-CNNでは，全結合層でクラス識別を行なう代わりに，生成された特徴量を線形SVMに入力して識別を行なう．ニューラルネットワークでクラス分類を行なうためには大規模な学習データを必要とするが，特徴量のクラス分類に線形SVMを用いることで少量の学習データからでも高精度な分類ができる[28]．多クラス物体の識別には，物体のクラスごとに学習した複数の線形SVMを使用する．識別結果が複数のクラスについてオーバーラップした場合には，non-maximum suppression (NMS)によってSVMのスコアが小さい方を除去する．Selective SearchとCNNの特徴量は複数のクラスに共通して計算できるため，クラス依存の計算は線形SVMの識別とNMSだけで効率的に計算できる．線形候補領域が物体として認識された後，CNNによって計算された特徴量から境界ボックス回帰を行なうことで，検出された領域がよりバウンディングボックスの配置に近づくように修正する.

3.4.4 Fast R-CNN

R-CNNは入力された全ての候補領域についてCNNの計算を行なうため，冗長な計算が多数発生して学習のための計算量が非常に大きくなるという問題がある．この問題に対して，計算量を削減して高速化を行なったアルゴリズムとしてFast R-CNNが提案されている[29]．Fast R-CNNでは初めに，CNNの畳み込み層までを使って任意サイズの入力画像の特徴量マップを計算する．次に，Selective Searchによって求めた候補領域を特徴量マップ上に射影し，候補領域についてプーリングを行なう．そして，特徴量の計算を行なった後，物体クラスの分類問題と境界ボックス回帰問題を同時に解く．また，学習の際には誤差逆伝播によって重みを更新する.

3.5 Deep Dense Face Detector

CNNを顔検出に適用した例として，FarfadeらはDeep Dense Face Detectorを提案した[30]．DDFDは，向きや目印といったアノテーションを使わずに，一つの検出器で多視点からの顔検出を行なうことを目的としている．また，検出器の構造を単純にすることで，計算の複雑さを最小化している．DDFDの基本的な構成はR-CNNと同様に，画像から切り出した領域をCNNに入力し，特徴量の計算を行なって物体の判定を

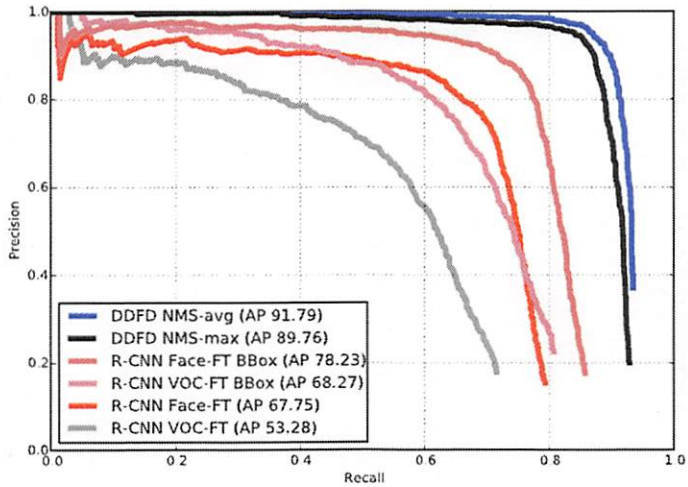


図 3.8: DDFD と R-CNN の比較(文献[30]より引用)

行なう．多視点顔検出を行なうために，DDFD は約 20 万枚の大規模な顔画像データセットについて学習している．DDFD と R-CNN との構造上の違いとして 3 点が挙げられる．まず，DDFD では画像からの領域の切り出しに Selective Search の代わりにスライディングウィンドウを使用する．この理由としては，検出器の構造を単純化する目的のほか, Selecti ve Search よりもスライディングウィンドウを用いた方が良い検出結果が得られたためであると著者は説明している．また，境界ボックス回帰についても，構造の単純化と横顔に対する検出率の低下を理由に DDFD では使用しない．さらに，構造を単純化するために顔領域の分類に SVM を使用せず CNN によって分類を行なう．顔検出における DDFD と R-CNN との比較を図 3.8 に示す. 図 3.8 において, NMS-max は顔と判定されたウィンドウが重なったときにスコアが最も大きいウィンドウの位置を検出する処理で，NMS-avg は平均化した位置を検出する処理である．また，Face-FT と VOC-FT はそれぞれ学習に使用したデータセットを意味し, BBox は境界ボックス回帰を意味する．この結果より，DDFD が R-CNN を上回る検出率を示すことが確認できる．これは，先述のように Selective Search と境界ボックス回帰が顔検出に不適であるためと考察されている．

3.6 むすび

本章では，ディープラーニングを用いた物体検出法について述べた．まず，ディープラーニングの考えの基となるニューラルネトワークについて述べた．次に，動画像認識に用いられるニューラルネットワークのモデルである CNN の概要について述べた．そして，CNN を物体検出に応用した手法である R-CNN とその改良手法である Fast R-CNN について述べた．最後に，CNN を多視点顔検出に適用した例である DDFD に

ついて述べ，顔検出における性能を示した．

第４章では，本章の内容に基づき，マンガキャラクターの多視点顔検出手法についての検討を行なう．

第4章　マンガキャラクターの多視点顔検出

4.1　まえがき

第3章において，ディープラーニングによって自動生成される特徴量が自然画像における多視点顔検出で高い性能を示すことを述べた．本章では，マンガキャラクターを対象とした多視点顔検出手法の検討を行なう．まず，本研究に使用するDPMの検出システムの概要を示す．次に，マンガ画像に最適なDPMの構成を実験より求める．次に，マンガキャラクターの多視点顔検出に対するR-CNNの適用について，DPMとの比較と，Selective Searchの有効性を実験より検証する．

4.2　マンガ画像に最適なDPM検出モデルの検討

本節では，マンガキャラクターの多視点顔検出を対象とした最適なDPMの構成について検討する．DPMはポジティブサンプルをアスペクト比から分類し，複数のルートフィルタの学習を行なうことができる．また，物体のパーツを捉えるパートフィルタについても任意の枚数に設定できる．従来のDPMは，一般物体全般を検出対象としてパラメータが設定されているが，この構成をマンガ画像に最適化させることで，更なる検出率の向上が期待できる．DPMのアルゴリズムはvoc-release5 [19]を使用した．

4.2.1　DPM最適化の学習・テストに使用するデータセット

本実験において，学習・テストに使用したデータセットについて説明する．ポジティブサンプルおよびネガティブサンプルは，2.6.1節にて定義したものと同様とする．

本実験では，マンガキャラクターの多視点顔検出を目的として正面顔，横顔，隠れ顔を検出対象とする．学習およびテストに使用したデータセットの内容を表4.1，表4.2に示す．学習セットには正面顔と横顔を含んだポジティブサンプル600枚・ネガティブサンプル1000枚を使用し，テストセットには正面顔，横顔と隠れ顔を含んだポジティブサンプル600枚・ネガティブサンプル1000枚を使用する．

4.2.2　ルートフィルタ数の最適化

2.4.9節で述べたように，DPMはバウンディングボックスのアスペクト比によって，ポジティブサンプルを複数のコンポーネントに分類して学習することが可能である．2.4節にて述べたDPMの多視点顔検出では，正面および横の左右方向について分類した4枚のルートフィルタを使用したとき検出率が最も高くなると報告している．一方，マンガではディフォルメ表現によって，あるキャラクターが他のキャラクターよりも面長に描かれるなど，正面顔の中でもアスペクト比が極端に異なるケースが考えられる．従って本実験では，正面顔についてさらに分類を行なった6枚のルートフィルタを用い

表 4.1: DPM 最適化の学習に使用するマンガ画像

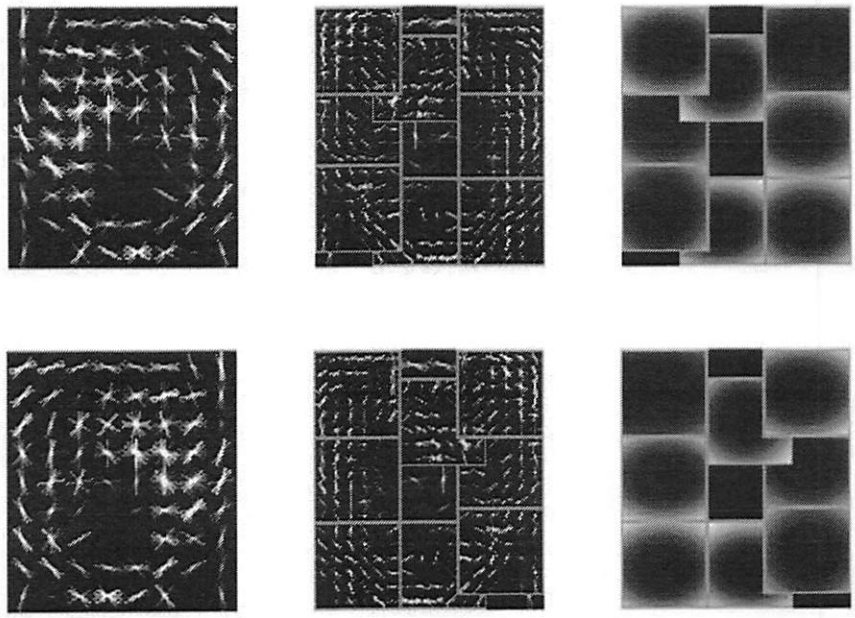
作品タイトル	ポジティブサンプル		ネガティブサンプル
	正面顔	横顔	
"ドラえもん"	100	50	1000
"ブラック・ジャック"	100	50	
"名探偵コナン"	100	50	
"SLAM DUNK"	100	50	
合計	400	200	1000

表 4.2: DPM 最適化のテストに使用するマンガ画像

作品タイトル	ポジティブサンプル			ネガティブサンプル
	正面顔	隠れ顔	横顔	
“ドラえもん”	90	10	50	800
“ブラック・ジャック”	90	10	50	
“名探偵コナン”	90	10	50	
“SLAM DUNK”	90	10	50	
合計	360	40	200	800

た場合についても検討を行なった．ルートフィルタ数をそれぞれ 2 枚，4 枚，6 枚と設定した検出器を作成し，検出率の比較を行なった．DPM のパラメータは，パートフィルタ数を 8 枚，NMS を 0.5 に設定した．

学習によって生成された検出モデルを図 4.1，図 4.2，図 4.3 に示す．図 4.1 はルートフィルタ数が 2 枚のときの検出モデル，図 4.2 は 4 枚のときの検出モデル，図 4.3 は 6 枚のときの検出モデルを表している．3 種類の検出器による Precision-Recall 曲線を図 4.4 に示す．図 4.4 より，ルートフィルタ数が 2 枚の検出器と 4 枚の検出器を比較すると，Precision，Recall において，ルートフィルタ数が 4 枚の方が全体的に高い値が得られている．一方，ルートフィルタ数が 6 枚の検出器は 4 枚の場合と大きな変化は見られなかった．AP はルートフィルタ数が 4 枚のとき 88.0%となり，最も高くなった．



ルートフィルタ　　パートフィルタ　　パートフィルタの配置

図 4.1: 2 枚のルートフィルタから構成される DPM 検出モデル

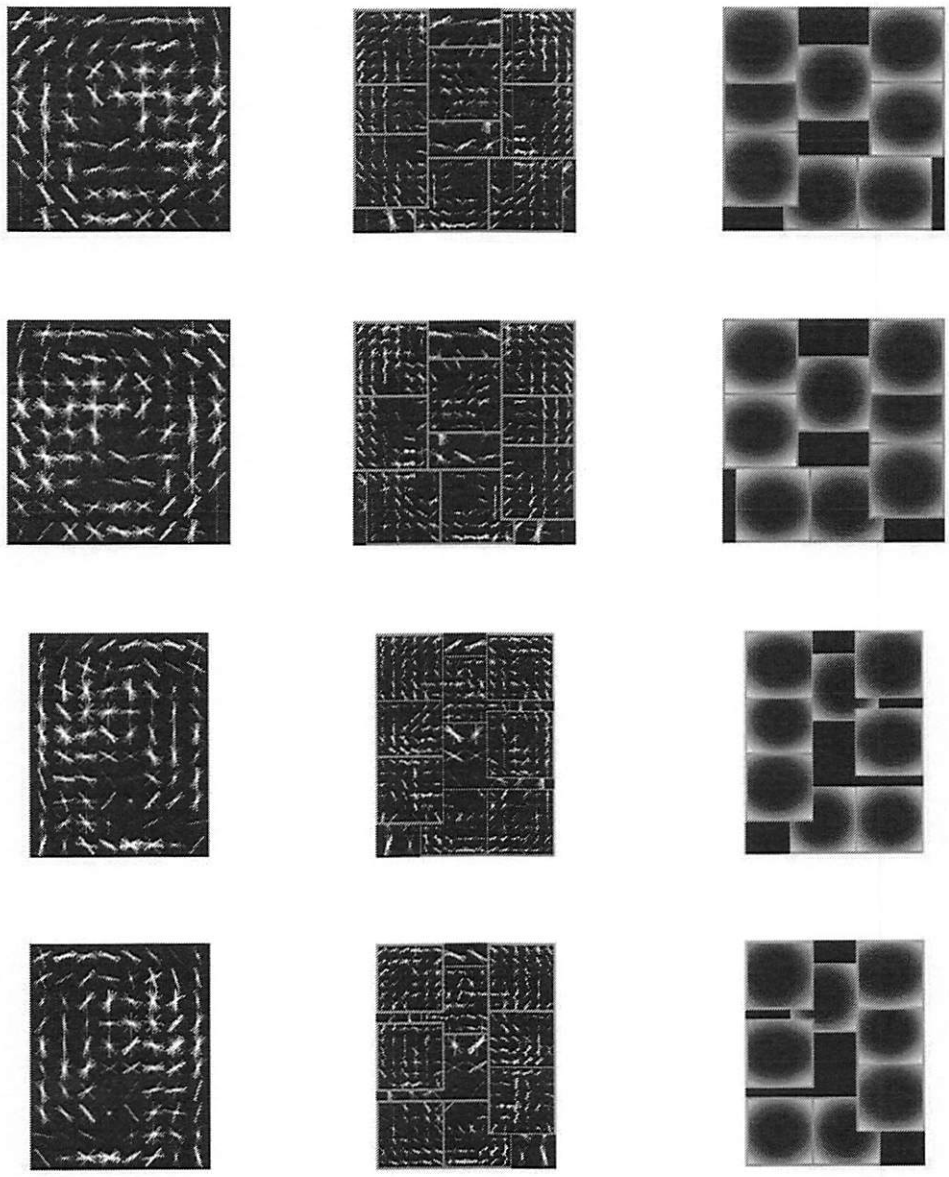


図 4.2: 4 枚のルートフィルタから構成される DPM 検出モデル

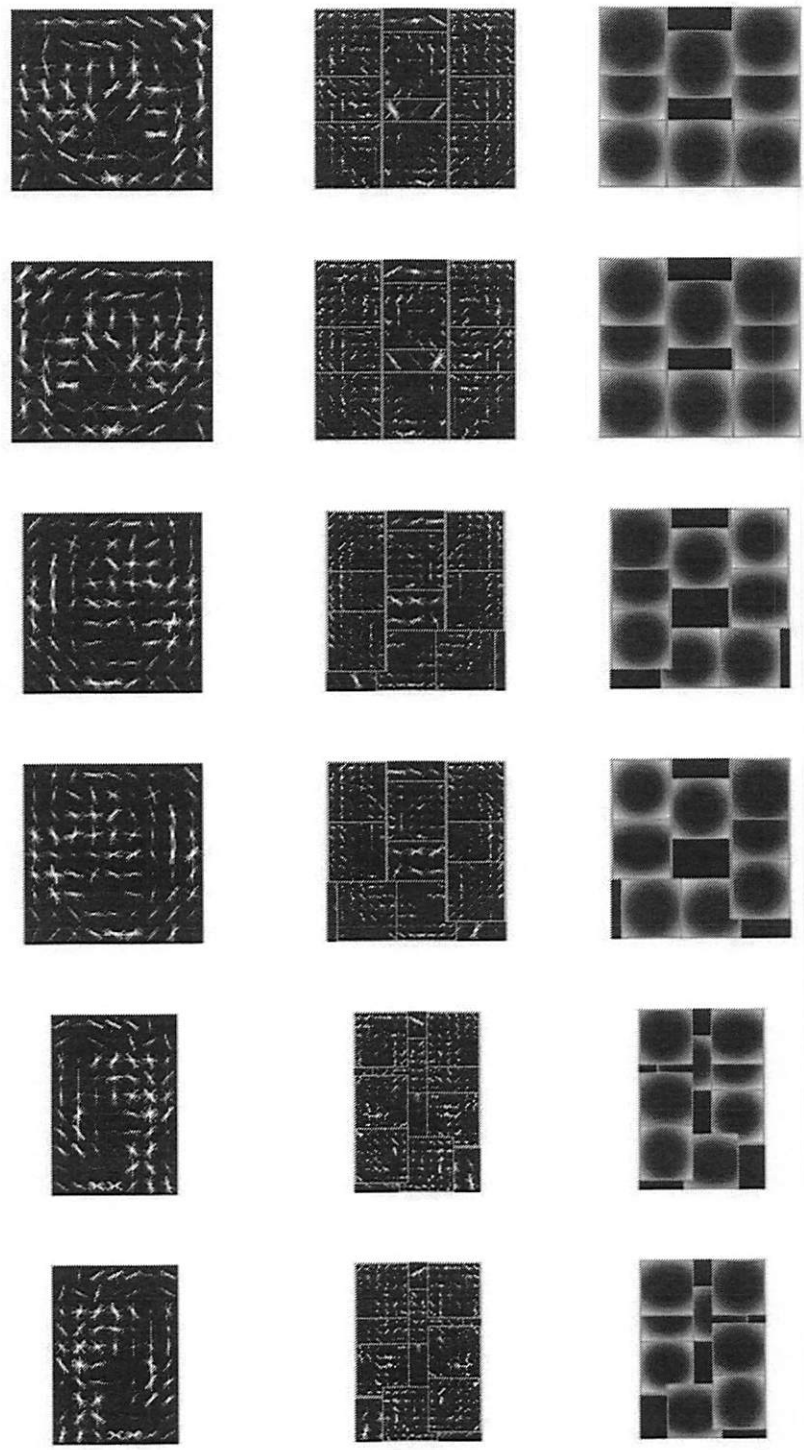


図 4.3: 6 枚のルートフィルタから構成される DPM 検出モデル

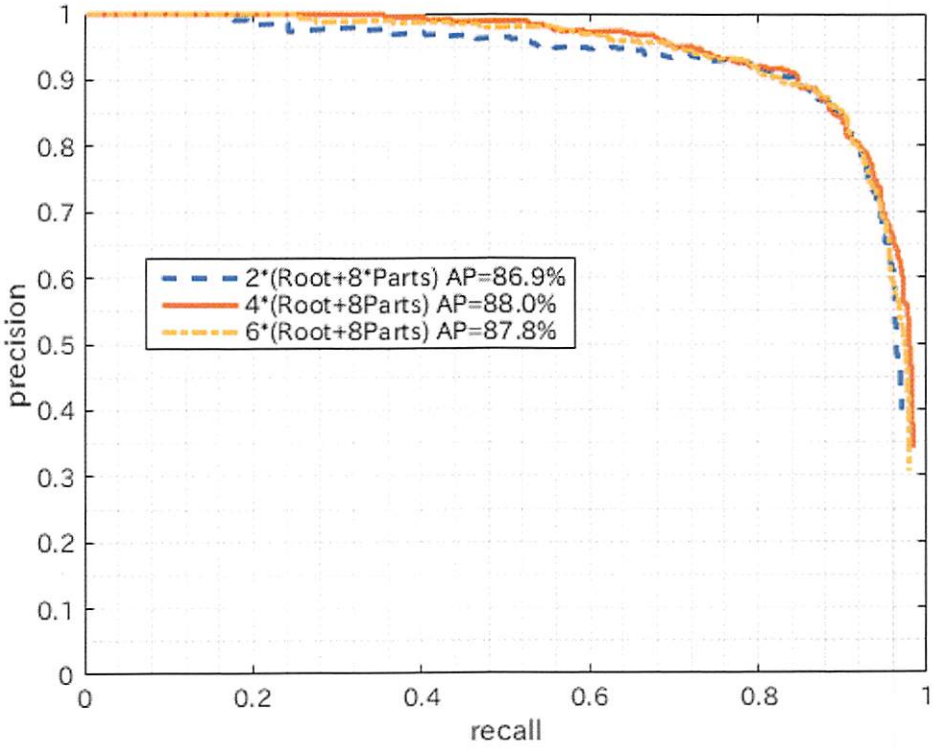


図 4.4: ルートフィルタ数による DPM の検出率変化

4.2.3 パートフィルタ数の最適化

2.4.2 節より，マンガキャラクター検出において 4 枚のルートフィルタが有効であることが分かった．この結果を踏まえて，マンガ画像に最適なパートフィルタ枚数について検討を行なった．ルートフィルタ数を 4 と設定し，パートフィルタ枚数を 2，3，4，5，6，8 枚に設定した検出器を比較した．DPM のその他のパラメータは，第 4.2.3 節と同様に設定した．

学習より生成された検出モデルのパートフィルタの応答と検出モデル内のパートフィルタの配置を図 4.4 に，6 種類の検出器による Precision-Recall 曲線の比較を図 4.5 に示す．適合率・再現率はどちらも，パートフィルタ数が 2 枚から 4 枚まで増えるごとに上昇し，パートフィルタ数が 4 枚以上増えた場合には大きな変動は見られなかった．AP は，パートフィルタ数が 4 枚のとき 88.2%となり，最も高くなった．

4.2.4 DPM 最適化の考察

以上の実験結果に基づき DPM のマンガキャラクター検出への最適化の考察を行なう．まず，第 4.2.1 節において述べたルートフィルタ数の最適化について述べる．ルートフィルタ数を 2 枚から 4 枚に上昇させたとき，検出率の増加が見られた．一方で，6

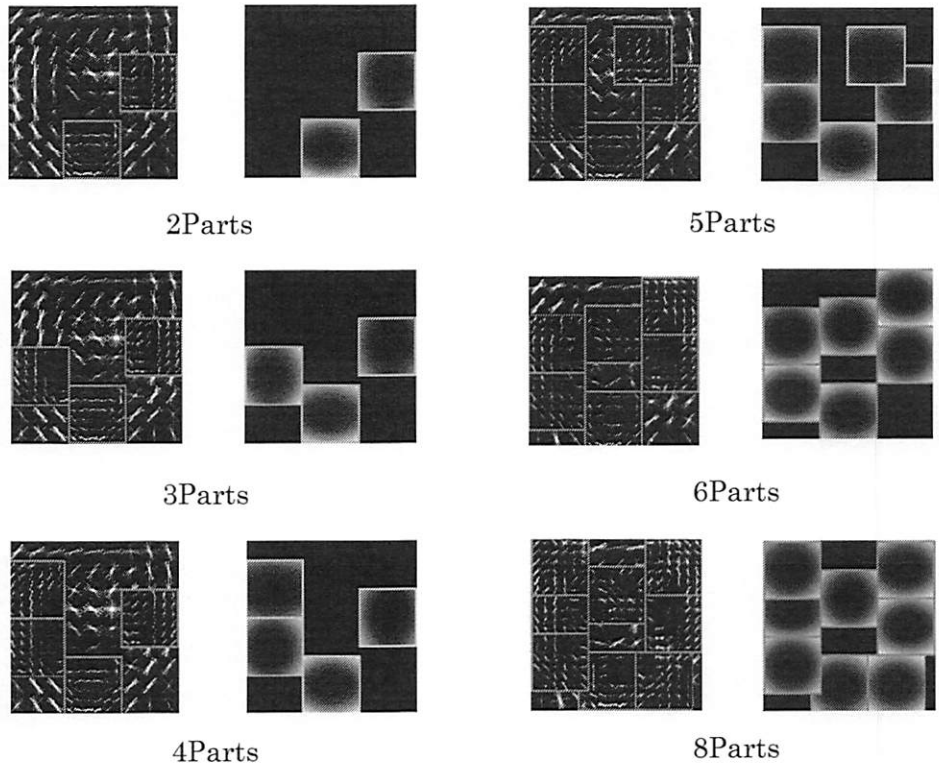


図 4.5: 検出モデルのパートフィルタの応答と配置

枚のルートフィルタを使用した場合には 4 枚のときより検出率が僅かに低下することが確認できた．従って，キャラクターに対応した検出器の細分化を行なうより，キャラクター全体について検出できる検出器を使用した方が全体的な検出率は高くなると考えられる．

次に，第 4.2.2 節において述べたパートフィルタ数の最適化について述べる．DPM を用いた人検出では，6 枚のパートフィルタが頭・両肩・両手・足の 6 パーツに対応する．図 4.5 に示した検出モデルより，マンガキャラクターの検出では 4 枚のパートフィルタが左右の輪郭と顎に対応していることが確認できる．パートフィルタ数が 4 枚以上増加した場合に検出率の大きな変動が見られなかった原因は，先述した 4 箇所のパーツがマンガキャラクターにおいて形状的な変動が少ないパーツであり，顔検出に大きく貢献しているためであると考えられる．

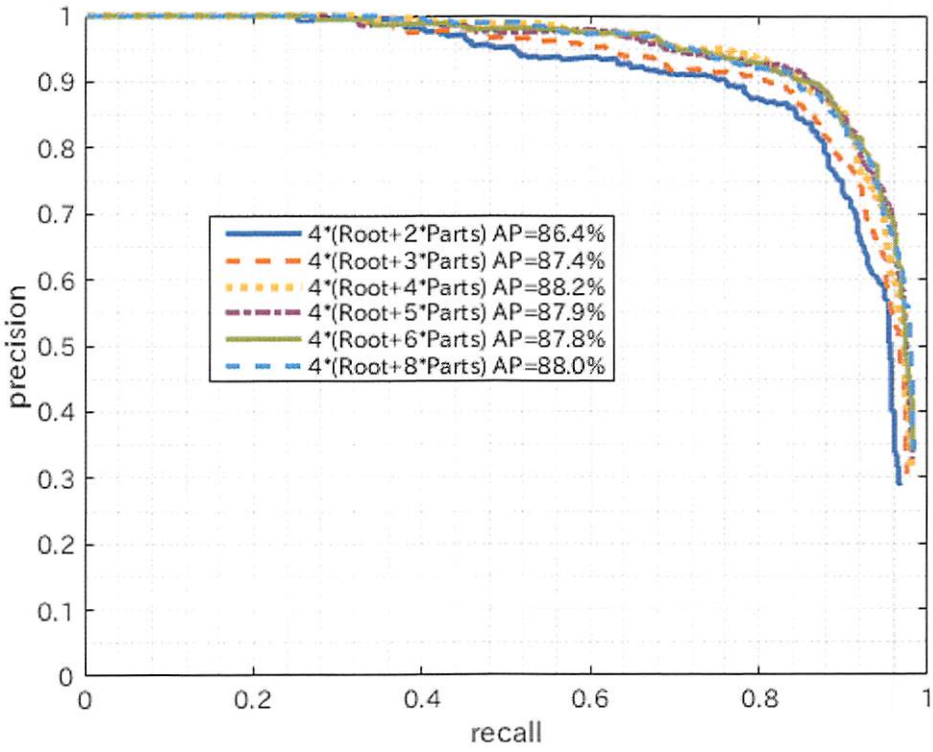


図 4.6: パートフィルタ数による DPM の検出率変化

4.3 R-CNN のマンガ画像への適用

R-CNN のマンガキャラクター多視点顔検出への有効性について，実験により検討する．

4.3.1 R-CNN と DPM の学習・テストに使用するデータセット

本実験では 4.2 節と同様に，マンガキャラクターの多視点顔検出を目的として，正面顔，横顔，隠れ顔を検出対象とした．ただし，DPM と R-CNN の 2 種類の検出器においてより正確な比較を行なうためにテストに使用するネガティブサンプルの枚数を増加させた．学習・テストに使用するデータセットの内容を表 4.3，表 4.4 に示す．学習セットは正面顔と横顔を含むポジティブサンプル 600 枚，ネガティブサンプル 1000 枚を使用した．テストセットは正面顔，横顔および隠れ顔を含むポジティブサンプル 600 枚，ネガティブサンプル 2000 枚を使用した．

4.3.1 マンガキャラクター検出における DPM と R-CNN の比較

マンガキャラクターの多視点顔検出について，DPM と R-CNN の比較を行なった．DPM の設定は 4.2.2 節の実験結果より，ルートフィルタ数を 4 枚，パートフィルタ数

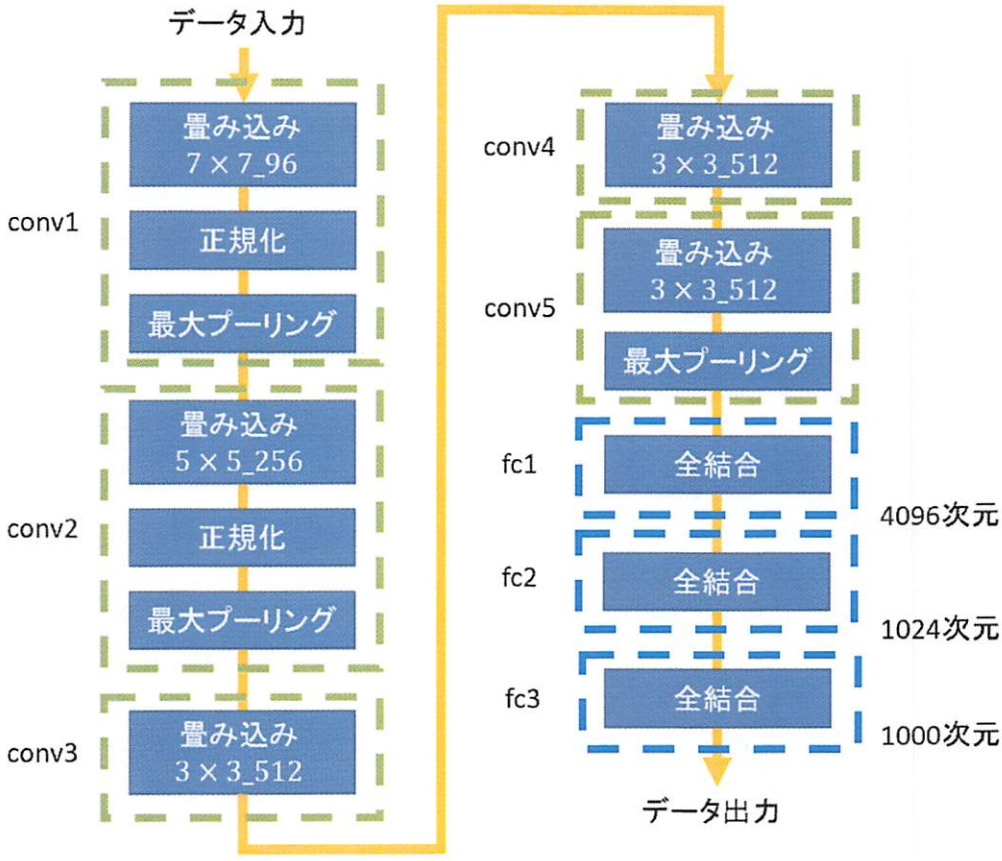


図 4.7: vgg_cnn_m_1024 の概要

を 4 枚と設定した．また，NMS を 0.1 として，その他のパラメータは 4.2.3 節と同様に設定した．R-CNN のアルゴリズムは，girshickICCV15fastrcnn[29]を使用し，ニューラルネットワークのアークテクチャには vgg_cnn_m_1024[31]を使用した．
vgg_cnn_m_1024 の概要を図 4.7 に示す．vgg_cnn_m_1024 は 5 層の畳み込み層と 3 層の全結合層から構成される 8 層の CNN である．第 7 層で出力される特徴量を 1024 次元とすることで，学習時間を削減している．R-CNN のパラメータは，NMS を 0.1 とし，学習の反復回数を 40000 回，バッチサイズを 128 に設定した．

DPM と R-CNN の Precision-Recall 曲線を図 4.8 に示す．AP について，R-CNN は DPM を 2.2%上回る結果が得られた

表 4.3: R-CNN と DPM の学習に使用するマンガ画像

作品タイトル	ポジティブサンプル		ネガティブサンプル
	正面顔	横顔	
"ドラえもん"	100	50	1000
"ブラック・ジャック"	100	50	
"名探偵コナン"	100	50	
"SLAM DUNK"	100	50	
合計	400	200	1000

表 4.4: R-CNN と DPM のテストに使用するマンガ画像

作品タイトル	ポジティブサンプル			ネガティブサンプル
	正面顔	隠れ顔	横顔	
“ドラえもん”	90	10	50	2000
“ブラック・ジャック”	90	10	50	
“名探偵コナン”	90	10	50	
“SLAM DUNK”	90	10	50	
合計	360	40	200	2000

4.3.3 Selective Search の有効性

3.4 節より，自然画像における多視点顔検出では，Selective Search による候補領域抽出が検出率を低下させることを述べた．本節では，Selective Search のマンガ画像に対する有効性について検討した．4.3.2 節で使用した R-CNN について，候補領域の切り出しに従来の Selective Search を使用した検出器と，スライディングウィンドウを使用した検出器の検出率を比較した．

二つの検出器の Precision-Recall 曲線を図 4.9 に示す．Selective Search を使用した検出器は，スライディングウィンドウを使用した場合と比べて AP が 0.02%高くなった．

4.3.4 R-CNN を用いたマンガキャラクター検出の考察

以上の実験結果より，R-CNN を用いたマンガキャラクター検出の考察を行なう．まず，第 4.3.2 節で述べた R-CNN と DPM の比較について述べる．実験結果より，マンガキャラクターの多視点顔検出において，R-CNN の検出率は DPM を上回った．このことから，ディープラーニングによる画像特徴抽出はマンガ画像に対しても有効であるといえる．

次に，4.3.3 節で述べた Selective Search のマンガ画像への有効性について述べる．実験結果より，マンガ画像ではスライディングウィンドウより Selective Search を使用した方が検出率は高くなるという結果が得られた．自然画像において Selective Search

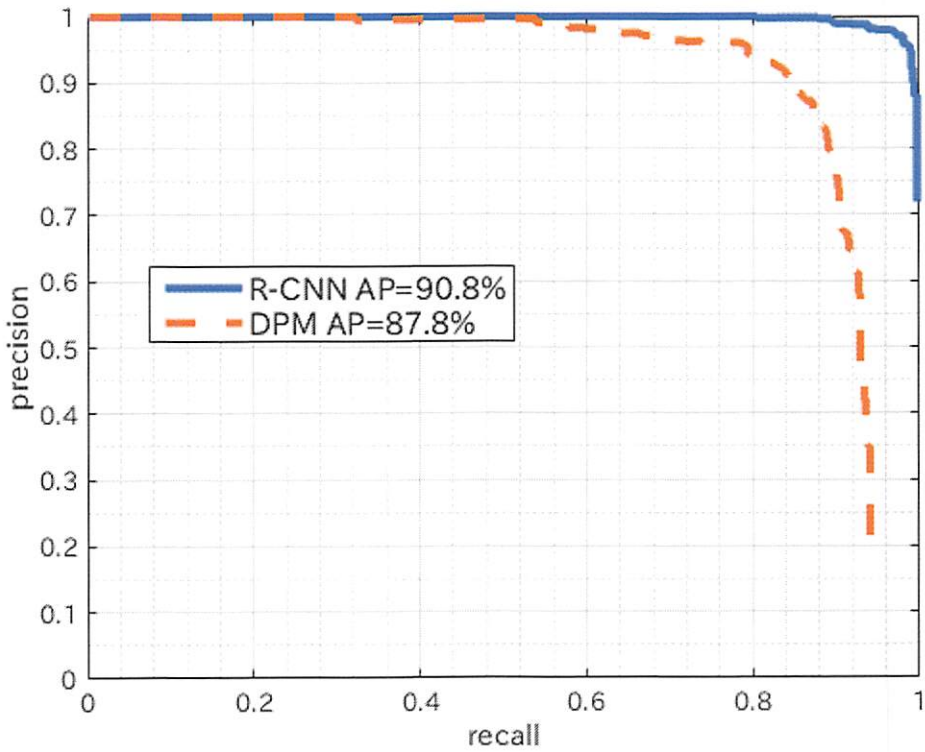


図 4.8: R-CNN と DPM の比較

によるセグメンテーションが不適である理由として，照明変化や画質によって色特徴やテクスチャ特徴が影響を受けやすいことが考えられる．これに対して，マンガ画像は白黒の均一なテクスチャで構成されるので，Selective Search によってセグメンテーションが正確に行なえるため検出率が低下しないと考察できる．

4.4 むすび

本章では，マンガキャラクターを対象とした多視点顔検出手法の検討を行なった．まず，マンガ画像に最適な DPM の構成を実験より求めた．次に，マンガキャラクターの多視点顔検出に対する R-CNN の適用について，DPM との比較と，Selective Search の有効性を実験より検証した．

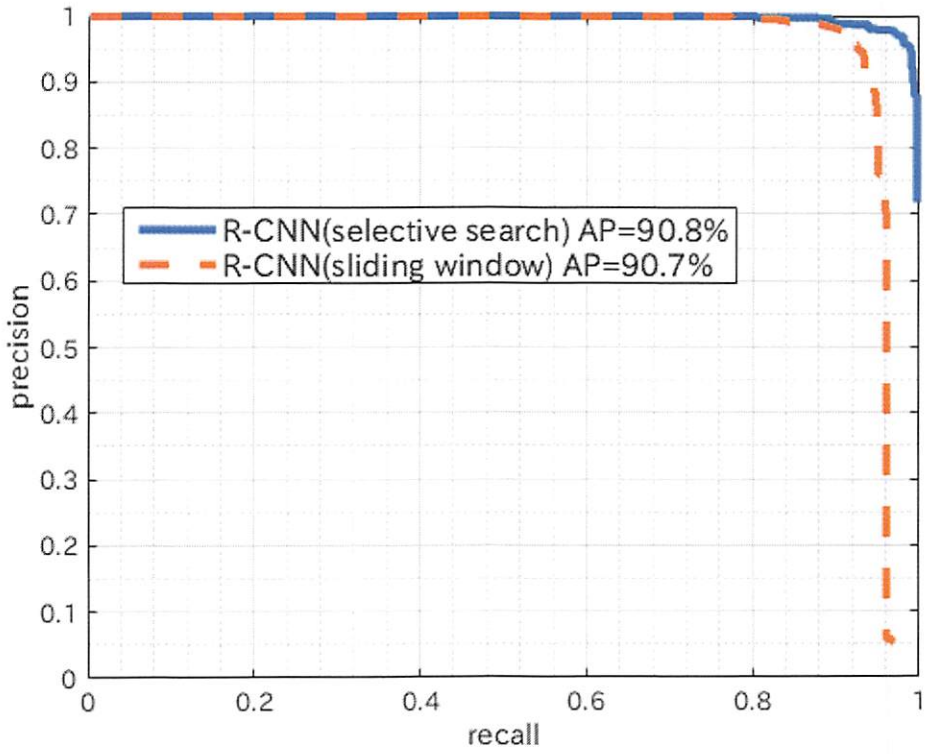


図 4.9: Selective Search とスライディングウィンドウの比較

第5章　結論

5.1　総括

本研究では，マンガキャラクターを対象とした多視点顔検出手法の検討を行なった．従来のマンガキャラクター検出では正面顔画像のみを検出対象としており，その検出には既定の画像特徴量を用いた手法が使われている．これに対して，近年ではディープラーニングによって自動生成される特徴量を用いた物体検出手法であるR-CNNが提案されている．本研究では，マンガ画像より横顔を含めたマルチビュー顔検出を実現することを目的として，R-CNNと従来手法のDPMとの比較から，ディープラーニングのマンガ画像への有効性について検討した．

本研究では，まず，DPMによるマンガキャラクターの多視点顔検出について，4枚のルートフィルタと4枚のパートフィルタを使用する検出モデルが最も有効であることを示した．そして，R-CNNとDPMとの比較から，R-CNNの優位性を示した．さらに，Selective Searchのマンガ画像への有効性を示した．

第1章では，本研究の背景と目的，および本論文の構成について述べた．

第2章では，マンガキャラクター顔検出の従来手法について述べた．まず，画像処理におけるマンガ画像の特徴について述べた．次に，画像特徴量記述子であるHOG特徴の概要を述べた．そして，パーツに対して可変な物体検出手法であるDPMの概要を述べた．さらに，DPMの多視点顔検出への応用について述べた．最後に，マンガキャラクター検出に対するパートモデルの有効性を示した．

第3章では，ディープラーニングを用いた物体検出法について述べた．まず，ディープラーニングの考えの基となるニューラルネトワークについて述べた．次に，動画像認識に用いられるニューラルネットワークのモデルであるCNNの概要について述べた．そして，CNNを物体検出に応用した手法であるR-CNNとその改良手法であるFast R-CNNについて述べた．最後に，CNNを多視点顔検出に適用した例であるDDFDについて述べ，顔検出における性能を示した．

第4章では，マンガキャラクターを対象とした多視点顔検出手法の検討を行なった．まず，マンガ画像に最適なDPMの構成を実験より求めた．次に，マンガキャラクターの多視点顔検出に対するR-CNNの適用について，DPMとの比較と，Selective Searchの有効性を実験より検証した．

第5章は結論であり，本論文の総括および今後の課題について述べている．

5.2　今後の課題

今後の課題として以下がある．

5.2.1 マンガ画像に適したニューラルネットワークの設計

本研究では，ニューラルネットワークのアーキテクチャとして一般物体認識を目的として設計された vgg_cnn_m_1024 [31]を適用している．よりマンガ画像に適したニューラルネットワークを使用することで，検出精度の更なる改善が期待できる．

5.2.2 少量のデータセットからの学習

本研究は，メタデータの自動抽出による，マンガ画像へのタグ付け作業の効率化を目的としている．本研究で使用した検出手法は，学習のために 1000 枚以上の学習セットが必要となる．実用的な顔検出システムを構築するために，少量のデータセットから自動学習を行なうといったアルゴリズムが必要である．

謝辞

本研究の機会及び素晴らしい実験環境を与えて下さり，貴重な時間を割いてご指導頂いた渡辺裕教授に心から感謝いたします．

日頃から研究のアドバイスや議論をして頂いた研究室の皆様に心から感謝いたします．

本研究を行なうにあたって，コミック画像の提供および論文への掲載を許可いただいた木野陽様に心から感謝いたします．

最後に，私をここまで育てて下さった家族に深く感謝します．

平成 28 年 2 月 1 日

参考文献

[1] インプレスビジネスメディア，“電子書籍ビジネス調査報告書 2015”，株式会社インプレスビジネスメディア，(2015).

[2] 松下光範，“コミック工学の可能性”，第 2 回 ARG WEB インテリジェンスとインタラクション研究会，pp.63-68，(2013).

[3] 石井大祐，河村圭，渡辺祐，“分割線選択によるコミックのコマ分割に関する検討”，情報科学技術フォーラム一般講演論文集，Vol.5，NO.3，pp. 263-264，(2006).

[4] 石井大祐，河村圭，渡辺祐，”コミックのコマ分割処理に関する一検討”，情報処理通信学会研究報告，Vol.2012-AVM-76，No.1，pp.1-5，(2012).

[5] Tanaka, T., Shoji, K., Toyama, F. and Miyamichi, J. “Layout Analysis of Tree-Structured Scene Frames in Comic Images,” Proc. 20th International Joint Conference on Artificial Intelligence , pp. 2885-2890, (2007).

[6] 野中俊一郎，野沢拓也，羽場典久，“コミックスキャン画像からの自動コマ検出を可能とする画像処理技術「GT-Scan」の開発”，FUJIFILM RESERCH & DEEVELOPMENT, No.57，pp. 46-49，(2012).

[7] 田中孝昌，外山史，宮道壽一，東海林健二，“マンガ画像の吹き出し検出と分類”，映像情報メディア学会誌，VOL.64, No.12，pp. 1933-1939，(2010).

[8] 新井俊宏，松井佑介，相澤清晴，“漫画画像からの顔検出”，電子情報通信学会総合大会，pp.161，(2012).

[9] 石井大祐，渡辺祐，“マンガからの自動キャラクター位置検出に関する一検討”，情報処理学会研究報告，Vol.2012-AVM-76, No.1, pp. 1-5，(2012).

[10] H. Yanagisawa, D. Ishii, H. Watanabe, “Face detection for comic images with deformable part model”, In The 4th International Workshop on Image Electronics and Visual Computing 2014, 4A-1, (2014).

[11] M. Viola and P. Viola, “Fast multi-view face detection”, Mitubishi Electric Research Lab TR-200003-9 3, (2003).

[12] 木野陽，”ベリーベリークリームショコラ　ふたつのベリー”，(2010).

[13] N. Dalal, B. Triggs, “Histograms of Oriented Gradients for Human Detection,” IEEE CVPR, pp. 886-893, (2005).

[14] P. Felzenszalb, R. Girshick, D. McAllester, D. Ramanan, “Object Detection with Discriminatively Trained Part Based Models”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.32, No.9, pp. 1627-1645 (2010).

[15] P. Felzenszalb, D. McAllester, D. Ramanan, “A Discriminatively Trained, Multiscale, Deformable Part Model”, IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, (2008).

[16] J. Orozco, B. Martineza, M. Pantic, "Empirical Analysis of Cascade Deformable

Models for Multi-View Face Detection", Image and Vision Computing, Vol.42, pp.47-61, (2015).

[17] B. Wu, H. Ai, C. Huang, S. Lao, "Fast rotation invariant multi-view face detection based on real adaboost", In IEEE International Conference on Automatic Face and Gesture Recognition, pp. 79-84 (2004).

[18] X. Zhu, D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild", In CVPR, IEEE, pp. 2879–2886, (2012).

[19] P. Felzenszwalb, R. Girshick, D. McAllester, "Discriminatively Trained Deformable Part Models Version 5", http://people.cs.uchicago, (2012).

[20] 藤子・F・不二雄，藤子プロ，"ドラえもん"，小学館.

[21] 手塚治虫，"ブラック・ジャック"，秋田書店.

[22] 青山剛昌，"名探偵コナン"，小学館.

[23] 井上雄彦，"SLAM DUNK"，集英社.

[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, "The PASCAL VOC2012 Results", (2012).

[25] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", In IEEE conference on Computer Vision and Pattern Recognition, pp. 580-587, (2014).

[26] J. Uijlings, K. van de Sande, T. Gevers, A. Smeulders, "Selective Search for Object Recognition", International Journal of Computer Vision, vol.102 (2), pp. 154-171, (2013).

[27] P. Felzenszwalb, D. Huttenlocher, "Efficient Graph-Based Image Segmentation", International Journal of Computer Vision, 59, pp. 167–181, (2004)

[28] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, "A Deep Convolutional Activation Feature for Generic Visual Recognition", arXiv:1310.1531, (2013).

[29] R. Girshick, "Fast R-CNN", International Conference on Computer Vision, (2015).

[30] S. Farfade, M. Saberian, "Multi-view Face Detection Using Deep Convolutional Neural Networks", International Conference on Multimedia Retrieval, arXiv:1502.02766, (2015).

[31] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, "Return of the Devil in the Details: Delving Deep into Convolutional Nets", British Machine Vision Conference, (2014).

図一覧

表一覧

研究業績

	題名	発表年月	発表掲載誌	連名者
(1)	マンガ画像からの顔検出におけるパーツ特徴量の一検討	2014年9月	映像情報メディア学会年次大会，17-9	石井　大祐 陳　明 渡辺　裕
(2)	Face detection for comic images with deformable part model	2014年10月	The 4th International Workshop on Image Electronics and Visual Computing (IEVC2014)，4A-1	石井　大祐 渡辺　裕
(3)	マンガの複数キャラクターに対する顔検出率について	2015年3月	電子情報通信学会総合大会，D-12-24	石井　大祐 渡辺　裕
(4)	R-CNNを用いたマンガキャラクター検出に関する一検討	2015年11月	映像メディア処理シンポジウム，I-4-12	渡辺　裕
(5)	マンガキャラクター検出における学習画像枚数の影響	2015年12月	映像情報メディア学会冬季大会，23B-5	渡辺　裕
(6)	マンガキャラクターのマルチビュー顔検出に関する検討	2016年3月予定	電子情報通信学会総合大会，D-12-12	渡辺　裕

マンガ画像からの顔検出におけるパーツ特徴量の一検討

A Study on Parts Features for Face Detection from Comic Images

柳澤 秀彰　† 石井 大祐　† 陳 明　渡辺 裕
Hideaki YANAGISAWA　Daisuke ISHII　Ming CHEN　and　Hiroshi WATANABE
早稲田大学大学院　基幹理工学研究科
† 早稲田大学大学院　国際情報通信研究科
Graduate School of Fundamental Science and Engineering, Waseda University
Graduate School of Global Information and Telecommunication Studies, Waseda University

Abstract　In recent years, studies of extracting meta-data from comic images has been focused on for query and search application of e-comics. In this paper, we propose to apply Deformable Part Model, which is a detection method using parts placement of an object, to detect character' faces in comic images.

1. はじめに

近年，マンガ書籍の電子化が進むに伴って，検索サービスの利便性向上のためにマンガ画像からメタデータを自動的に抽出する技術について研究が行なわれている．その中で登場人物の抽出については，マンガ画像より顔領域を検出して登場人物の認識を行なう手法が提案されているが[1]，現在のところ顔領域の安定した検出手法は確立されていない．

本稿では，物体のパーツ配置を利用した検出手法である Deformable Part Model (DPM)[2],[3]のマンガ画像に対する応用について検証する．また，DPM のパーツ検出に使用する特徴量を変化させた場合に，検出結果に対して与える影響について検討を行う．

2. Deformable Part Model

Deformable Part Model (DPM)は Felzenszwalb らが提案した物体検出手法であり，物体のモデルをパーツの集合として表現している．物体全体と各パーツにおける形状の妥当性に加えて，物体とパーツの相対位置情報を隠れ変数とすることで評価を行なう．DPM のスコアは以下の手順で求められる．

1. 解像度の異なる画像についてそれぞれ特徴量マップを求めた特徴量ピラミッドを作成する．
2. ある解像度の特徴量マップに対して，物体の全体を捉えるルートフィルタを配置し，フィルタと特徴量との内積を算出して，ルートのスコアを求める．
3. ルートフィルタの 2 倍の解像度の特徴量マップに対して，検出物体のパーツを捉えるパートフィルタをそれぞれルートフィルタに重なるように配置し，各パートフィルタと特徴量との内積を計算する．
4. パートフィルタとルートフィルタの相対的な位置関係から、各パーツの歪み情報を算出する．
5. 3, 4 を繰り返し，パートフィルタのスコアから歪み情報を引いた値が最大となるパートフィルタの位置を求める．このルート位置における最終的なスコアは，ルートと各パーツのスコアの合計値として求められる．
6. スコアの合計値が閾値を超えた場合，そのルート位置を検出対象として検出する．

通常の DPM では，画像特徴量として局所領域における輝度勾配を利用した手法である Histograms of Oriented Gradients (HOG)[4]特徴量を使用している．しかしマンガ作品において，目や輪郭といった顔を構成するパーツは，実画像における物体の構成パーツと比較して，形状変化が大きい傾向にある．本稿では，パートフィルタに対する特徴量を，局所領域における周辺画素の輝度値の大小を利用した手法である Local Binary Pattern (LBP)[5] に変更することで，コミック画像からの検出にどのような影響を及ぼすかについて検討を行った．

3. 実験

本実験では DPM の学習と検出に公開プログラム[6]を使用した．一つのマンガ作品より，正面を向いた顔領域を手動で切り出した画像 100 枚を正例，顔領域を含まない部分を切り出した画像 206 枚を負例として DPM によって検出器を作成し，既知画像 15 枚，未知画像 11 枚に対して顔検出を行った．また，DPM のパートフィルタに利用する特徴量を LBP 特徴量に変更したものについて，同様のデータから学習を行い，検出器を作成した．

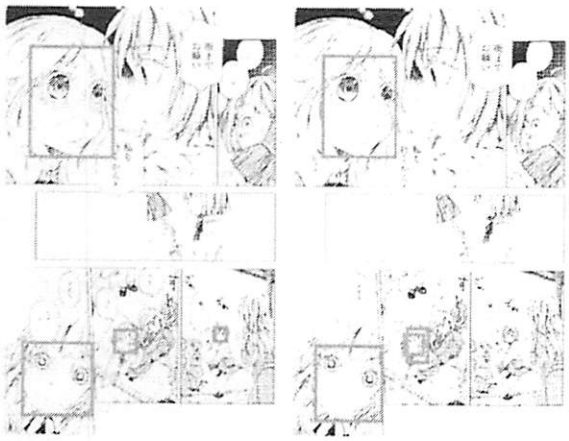


図 1. マンガ画像[7]に対する顔検出結果
Fig1. Result of face detection for comic images.

左：通常の DPM　右：パートフィルタに LBP を使用

図 1 にマンガ画像[7]からの検出例を示す．通常の DPM，LBP 特徴を利用した DPM ともに，顔領域について検出が行われていることを確認できた．

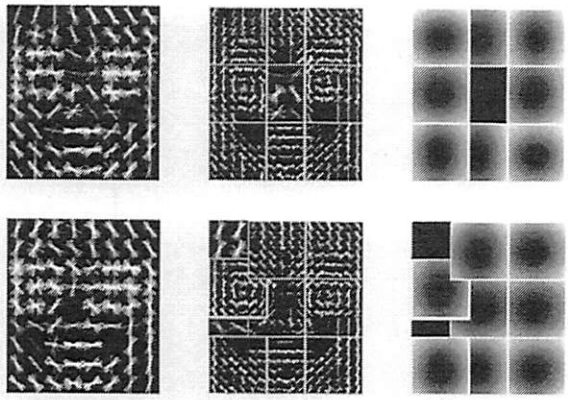


図 2. DPM の検出モデル
Fig2. Detection model of DPM

上段：通常の DPM　下段：パートフィルタに LBP を使用
左より順にルートフィルタ，パートフィルタ，パートフィルタの配置を表す．

図 2 に DPM の検出モデルを示す．通常の DPM と，LBP 特徴量を利用した DPM においてパートフィルタの配置が変化していることが確認でき，パートフィルタの特徴量の変化が，検出モデルに影響を与えることが確認できた．表 1 に DPM による顔領域の検出率，表 2 にパートフィルタに LBP を利用した DPM による検出率を示す．通常の DPM の未知画像に対する検出成功率が 94.6%であるのに対して，LBP を使用した DPM による未知画像への検出率は 88.6%に低下していた．

表 1. 通常の DPM による顔検出結果

	検出対象数	検出数	誤検出数
既知画像	52	52	0
未知画像	37	35	0

表 2. LBP を使用した DPM による顔検出結果

	検出対象数	検出数	誤検出数
既知画像	52	50	0
未知画像	37	31	0

4. まとめ

実験より，DPM を用いた検出がマンガ画像に対して有効であることが分かった．LBP 特徴量を利用した DPM については，通常の DPM と比較して検出率は低下したが，パートフィルタの配置が変化することが確認できたため，今後改善の余地があると考えられる．

謝辞

本稿では，マンガ家の木野陽様 http://www.etheric-f.com/ より学術目的の為に使用許可を頂いたマンガを実験に使用した．マンガ画像の提供および原稿への掲載を許可いただいた木野陽様に深く感謝する．本研究は JSPS 科研費 25330137 の助成を受けたものである．

文　献

[1] 石井大祐，渡辺裕，“マンガからの自動キャラクター位置検出に関する一検討”，情報処理学会研究報告 Vol.2012-AVM-76, No.1, pp.1-5, (2012)
[2] P.E Felzenszalb, et al. : “Object Detection with Discriminatively Trained Part Based Models,” PAMI, vol.32, No.9, pp.1627-1645, (2009)
[3] P.E Felzenszalb, et al. : “A Discriminatively Trained, Multiscale, Deformable Part Model,” CVPR, (2008)
[4] N. Dalal and B. Triggs: “Histograms of oriented gradients for human detection,” IEEE Computer Society Conference on Computer Vision & Pattern Recognition, pp.886–893 (2005).
[5] T. Ojala, M. P. Ainen and D. Harwood: “A comparative study of texture measures with classification based on featured distributions,” Pattern Recognition, Vol. 29, pp.51–59 (1996).
[6] Girshick, R. B. and Felzenszwalb, P. F. and McAllester, D., “Discriminatively Trained Deformable Part Models, Release 5”, <http://people.cs.uchicago.edu/~rbg/latent-release5/>, <最終アクセス 2014/2/7>
[7] 木野陽: ベリーベリークリームショコラ ふたつのベリー, (2010)

†早稲田大学　基幹理工学研究科
〒169-0072 東京都新宿区大久保 3-14-9 早大シルマンホール 401
TEL.03-5286-2509　E-mail:bule-cosmo@ruri.waseda.jp

Proceedings of the Fourth IIEEJ International Workshop
on Image Electronics and Visual Computing
Koh Samui, Thailand, October 7-10, 2014

FACE DETECTION FOR COMIC IMAGES WITH DEFORMABLE PART MODEL

Hideaki Yanagisawa, Daisuke Ishii and Hiroshi Watanabe

Department of Computer Science and Communication Engineering,
Graduate School of FSE, Waseda University

ABSTRACT

Comic images include several kinds of picture elements, such as lines, dots, characters and sound effects. Therefore, they form quite complex structure compared with natural images. We have been trying to improve the convenience of e-comics by retrieving metadata elements, such as names of characters and positions of the characters. To extract characters from comic images, a method detecting characters' face using the Histograms of Oriented Gradients (HOG) features and discriminating them has been proposed. However, this method does not provide stable face detection. In this paper, Deformable Part Model (DMP), which is originally proposed to detect natural objects, is applied to comic images in order to improve accuracy of face detection. As a consequence, it is turned out that we can obtain 85.5 % detection rate for unknown images. Thus, DPM can be regarded an effective method to detect objects in comic images.

1. INTRODUCTION

According to a survey in 2011, sales of e-comics accounted for 81.7% of the e-book market [1]. It shows that e-comics have an important presence while the e-book market is expected to grow in the future.

In archived comic images, it is possible to provide more convenient services by providing metadata such as character names, and balloons and panel layout. For example, such services can be considered to change the image size to fit the screen of a terminal, to search target images from archived data based on the information of characters or particular scenes.

Currently, several approaches have been proposed for the extraction of the balloon and panel layout with high accuracy [2] [3].

On the other hand, to extract the characters, a method for detecting a face area of the characters using HOG features [4] and SVM have been proposed [5] [6]. Also, in this method, it is suggested that by limiting the iris portion of the face subjected to detection, it is possible to improve the accuracy of detection. However, the false positive rate is quite high, and does not provide stable detection.

In this paper, we apply Deformable Part Model, which is originally proposed to detect objects for a natural image, to comic images. Through the experiment of face detection to comic images, highly accurate detection rate can be obtained.

2. DETECTION METHOD

2.1 Histograms of Oriented Gradients (HOG)

Most of the comic images are painted by lines, which is basically binary representation. Comparing with a natural image, a comic image contains a lot of edge components. At the edge, the change of the intensity is large and at the flat area, the change of the intensity is small. Basically, HOG features use the information about the edge direction. Therefore, it is regarded to be a desired feature descriptor for comic images. HOG features can be calculated by the following procedure.

1. Gradient direction and gradient strength are calculated from the intensity of each pixel in the image. Next, the gradient direction is quantized to 9 discrete directions. Namely, separated into each 20° range of up to 160° from 0°.

2. The local area is split into cells, where each cells consists of 8×8 pixels.

3. In each cell area, create a gradient direction histogram of intensity. Then, obtain 9 dimensional vector.

4. The 2×2 cells are regarded as one block. To each block, normalization is performed by combining the vectors of the cells. The feature vector that is finally obtained is composed of multi-dimensional vector from the vectors of all blocks.

Fig. 1 Samples of face area.

Fig. 2 Example of non-face area.

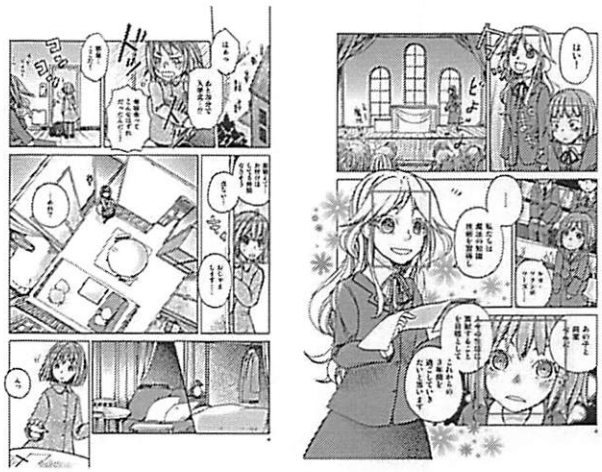


Fig. 3 Positive samples of learning images.
(Frames represent points specified in the bounding box as face areas.)

Fig. 4 Negative samples of learning images.
(Does not include the face area that defined in positive samples.)

2.2 Deformable Part Model (DPM)

Deformable Part Model (DPM) is a method of object detection proposed by Felzenszalb et al. [7] [8]. This method expresses the object model as a set of parts, and evaluates it by the validity of each part and relative position relationship thereof. In the conventional method, part locations of the object are fixed. However, the part location is variable in DPM, and it is possible to respond to pose changes of the object. The score of detection window is calculated from the next equation.

$$\text{score} = \sum_{i=0}^{n} F_i \cdot \phi(H, p_i) - \sum_{i=1}^{n} d_i \cdot \phi(dx_i, dy_i) + b \quad (1)$$

In the first term of Eq. (1), it calculates the score of filters. The detection model of DPM is constructed form one route filter, which captures the entire image of object, and n-part filters, which capture several parts of the object. First, the score of the route filter is calculated form the inner product of route filter (F_0) and HOG feature map of the image ($\phi(H, p_0)$). Second, scores of part filters are calculated form the inner product of part filters ($F_1 \sim F_n$) and HOG feature map at the twice higher resolution ($\phi(H, p_1) \sim \phi(H, p_n)$).

In the second term of Eq. (1), it calculates the distortion of each part placement. Then, $\phi(dx_i, dy_i)$ suggests relative positions of the route filter and part filters, and d_i is distortion parameter.

By training, DPM sets the value of F_i and d_i, and it creates object detector. Detection by the DPM has been applied to various natural objects so far. However, detection results by DPM to objects represented by line drawings, such as the comic images, have not yet been shown. Therefore, we try to apply DPM to comic images, and check the performance of this approach through the experiment described in the next section.

3. FACE DETECTION

By giving the training data in DPM, we create a detector for the face area of the comic image. We also create a detector by HOG features and SVM. Two detectors are used to detect face area from comic images, and these performances are compared. In this experiment, we used the algorithm in the reference [9] for the learning model of DPM and detection.

Comic images, used in this experiment, are 26 pages, size of 1342 × 1877 pixels. We use 9 of these as a training sample. Face areas, which should be detected and learned, are assumed to satisfy the three conditions below.

1. Both eyes are included.
2. Size is 60×60 pixel or more.
3. Area from the top of eyes to chin is included.

Fig. 1 shows the example images of face area assumed by above conditions. Fig. 2 shows the example images of non-face area.

Positive samples, specifying the face areas, are taken and shown by bounding box from the learning image. Fig 3 shows the example of positive samples.

Negative samples are those to cut out a region that does not include the face area from learning images. Fig 4 shows the example of negative samples. As a result, positive samples are 28 locations in 9 images, and negative samples are 114 sheets of images.

4. RESULT

Fig. 5 shows the detection model of the face region generated by DPM after learning. From these figures, relative locations can be seen. Further, Fig. 6 shows an example image of the result of performing the detection of the face area from a comic image using two detectors. Table 1 shows the result of performing detection from known and unknown images using the two detectors. From the experiment, it is turned out that DPM greatly outperforms HOG from 17.4% to 61.3%.

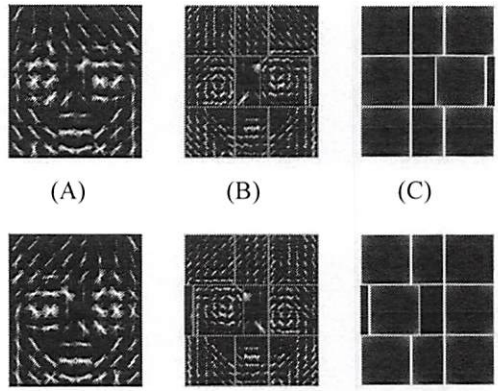


Fig. 5 Face detection model generated by DPM.
(A) HOG features of root filter. (B) HOG features of part filters. (C) Location of part filters.

Table. 1 Results of face detection from known and unknown images with DPM and HOG.

	Known images			Unknown images		
	precision	recall	F-measure	precision	recall	F-measure
DPM	92.9%	100%	96.3%	85.5%	100%	92.2%
HOG	75.0%	82.6%	78.6%	32.7%	38.7%	35.4%

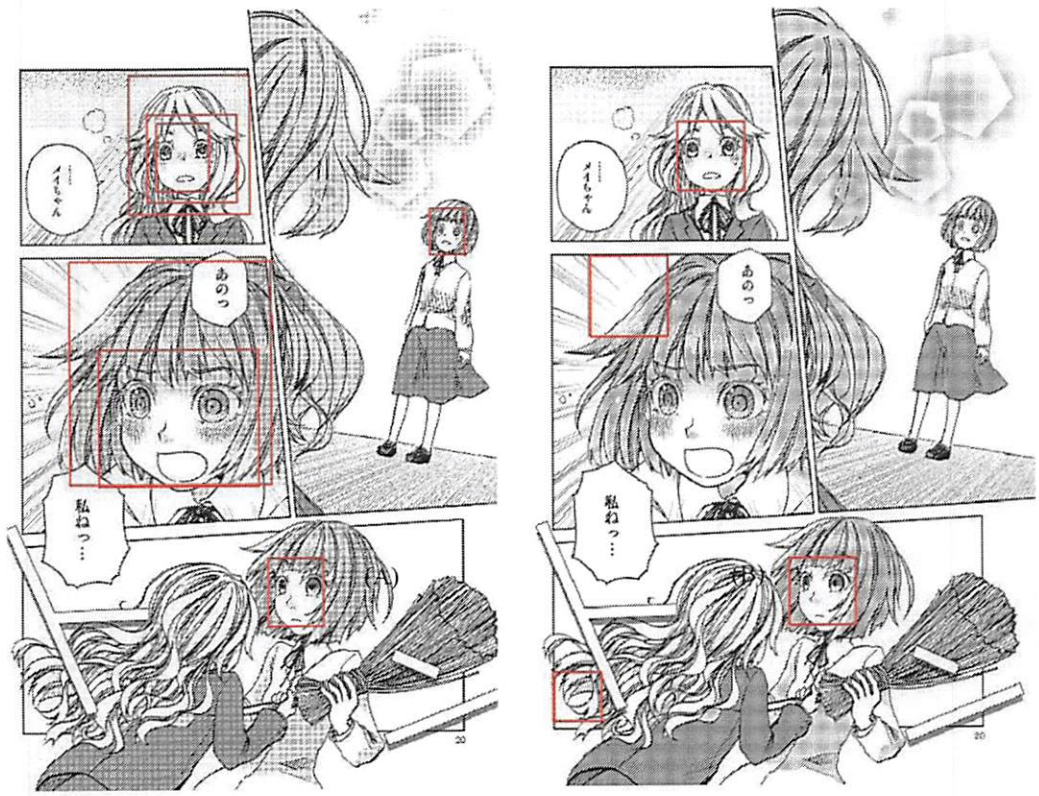


Fig. 6 (A) Result of face detection with DPM. (B) Result with HOG from same image. (Frames suggest detected area as character's face).

5. CONCLUSION

In this study, we have tried to improve the detection accuracy of the character's face area in comic images by using the DPM. The experimental results showed that the DPM improves both precision and recall for unknown images compared with the method that uses HOG only. Therefore, we could conclude that DPM is valid for detecting objects in comic images.
On the other hand, DPM could not detect radically distorted face images as shown in Fig. 7 It seems difficult to detect all of such images by using only DPM. To solve this problem, combination of DPM and some other features such as texture pattern may be effective.

Fig. 7 Samples of face images that DPM could not detect faces.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number 25330137. All comic images used in Fig.1-Fig.7 except for Fig 5 in this paper are provided by Hinata Kino (http://www.etheric-f.com/).

REFERENCES

[1] Internet Media Research (Ed), "e-Comic Business Report 2012", Impress Business Media Corporation, 2012.
[2] S. Nosaka, T. Sawano and N. Haneda, "Development of "GT-Scan", the Technology for Automatic Detection of Frames in Scanned Comic", FUJIFILM RESARCH & DEVELOPMENT, No. 57, pp. 46-49, 2012.
[3] T. Tanaka, F. Toyama, J. Miyamichi and K. Shoji, "Detection and Classification of Speech Ballons in Comic Images," ITE Vol. 64, No. 12, pp. 1933-1939, 2010.
[4] N. Dalal, and B. Triggs,"Histograms of Oriented Gradients for Human Detection," IEEE Conference on Computer Vision and Pattern Recognition CVPR, pp.886-893, 2005.
[5] D. Ishii and H. Watanabe "A Study of Automatic Human Detection for Comic Image," IPSJ SIG Technical Report Vol.2012-AVM-76, No.1, pp.1-5, Feb, 2012.
[6] T. Tanaka, K. Shoji, F. Toyama, and J. Miyamichi, "Layout Analysis of Tree-Structured Scene Frames in Comic Images," Proc. 20th International Joint Conference on Artificial Intelligence , pp.2885-2890, 2007.
[7] P. Felzenszalb, R. Girshick, D. McAllester, D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol.32, No.9, pp.1627-1645, 2010.
[8] P. Felzenszalb, D. McAllester, D. Ramanan, "A Discriminatively Trained, Multiscale, Deformable Part Model," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.
[9] R. B. Girshick, P. Felzenszwalb, and D. McAllester, " Discriminatively Trained Deformable Part Models, Release 5", http://people.cs.uchicago.edu/~rbg/latent-release5/, <accessed in 7.2.2014>.

マンガの複数キャラクターに対する顔検出率について

On face detection rate for characters in comic images

柳澤 秀彰[*1]　石井 大祐[*2]　渡辺 裕[*1,2]
Hideaki Yanagisawa　Daisuke Ishii　Hiroshi Watanabe
[*1] 早稲田大学大学院基幹理工学研究科　[*2] 早稲田大学大学院国際情報通信研究科
[*1] Graduate School of Fundamental Science and Engineering, WASEDA University
[*2] Graduate School of Global Information and Telecommunication Studies, WASEDA University

1. まえがき

近年，電子書籍市場の拡大に伴い，電子コミックに関するより高度な検索システムを提供するために，マンガ画像からメタデータを自動抽出する技術について研究が行われている．

本稿では，マンガ作品において重要な要素である登場人物の顔領域検出について，学習サンプルの違いによる検出率の変化を検討する．マンガの登場人物の顔領域は現実の顔画像と比較して，個々の特徴変化が大きい傾向にある．そこで，顔検出器の学習に，特定の登場人物の顔画像のみを使用した場合と，複数の登場人物の顔画像を使用した場合について，検出率の比較を行った．

2. 顔領域の検出手法

本稿では，マンガ画像から顔領域を検出するための手法として，Felzenszwalb らによって提案された物体検出手法である Deformable Part Model [1]を使用した．Deformable Part Model の検出モデルは物体の大まかな形状を捉えるルートフィルタと，物体の各パーツの形状を捉える移動可能なパートフィルタの 2 種類から構成されている．具体的な検出手順は以下のようになる．

1. 複数の解像度の画像（画像ピラミッド）についてそれぞれ HOG（Histograms of Oriented Gradients）特徴量を計算し，HOG ピラミッドを求める．
2. HOG ピラミッドに対するルートフィルタと各パートフィルタの応答を計算する．ここでパートフィルタではルートフィルタの 2 倍の解像度の画像に対する応答を計算する．
3. 各パートフィルタの応答から移動コストを減算した値を最終的なパートフィルタの応答とする．
4. 全てのフィルタの応答の和を取り，評価関数を計算する．
5. 評価関数の値の大きい場所が物体として検出される．

従来の物体検出手法は物体について各パーツの位置が固定のため，物体の姿勢が変化した場合に正しく検出できないといった問題があった．本手法は各パーツの位置についてある程度可変であるため，物体の姿勢変化について頑強であるという利点を持っている．

この手法はマンガ画像に対しても有効であり，HOG 特徴量と SVM を用いた手法よりも高い精度でマンガ画像内の顔領域の検出を行えることが示されている[2].

3. 実験

マンガ 1 作品における 4 種類の登場人物 A~D の正面正立画像をそれぞれ 70 枚ずつ切り出し，1 種類の登場人物のみをポジティブサンプルとした場合と，4 種類の登場人物全てをポジティブサンプルとした場合について検出精度を調べた．ネガティブサンプルはマンガ画像から登場人物の顔領域を含まない領域を切り出した画像 800 枚を使用し，認識対象とする入力画像は A~D の正面正立画像をそれぞれ含むマンガ画像 30 枚ずつとした．

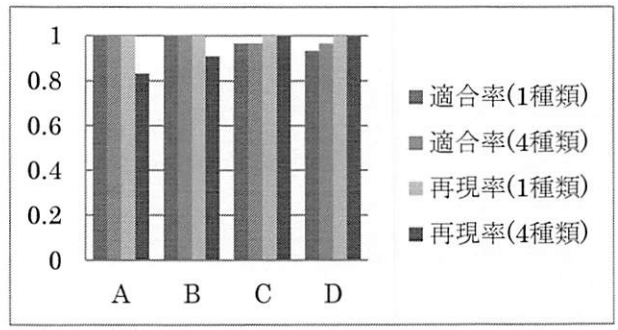


図 1 各キャラクターに対する検出結果

図 1 は登場人物 A~D についてそれぞれ該当する登場人物のみを学習した検出器と，4 種類の登場人物を学習した検出器における顔検出結果の適合率，再現率を表している．4 種類の人物を学習した検出器では，D について約 3.3%の適合率の上昇が見られた．一方，再現率は A について約 16.7%，B について約 9%の減少が見られた．この結果から，複数の登場人物を学習することで，変化が大きい顔領域を認識することが可能になったが，同時に顔以外の領域を誤検出する確率も増加したことが分かる．

4. まとめ

本稿では，マンガ画像からの顔検出に関して，複数の人物の顔画像を学習した場合の検出率の変化について検討を行なった．今後は，異なる作品間における Deformable Part Model の検出率の変化についても検討を行ないたい．

謝辞　本研究は JSPS 科研費 25330137 の助成を受けたものである．

参考文献

[1] P. Felzenszalb, D. McAllester, D. Ramanan, “A Discriminatively Trained, Multiscale, Deformable Part Model,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[2] H. Yanagisawa, D. Ishii, H. Watanabe: “Face detection for comic images with deformable part model,” The 4th International Workshop on Image Electronics and Visual Computing 2014 (IEVC2014), 4A-1, Oct. 2014.

R-CNNを用いたマンガキャラクター検出に関する一検討
A Study on Character Detection for Comic Images with R-CNN

柳澤 秀彰　　渡辺 裕
Hideaki Yanagisawa　Hiroshi Watanabe

早稲田大学大学院 基幹理工学研究科 情報通信専攻
Graduate School of Fundamental Science and Engineering,
Waseda University

Abstract: 近年の電子コミックの普及に伴い，マンガコンテンツのメタデータを利用して作品の検索や要約の作成を行なうといった新たなサービスが提案されている．しかし，現状ではマンガ画像より手動でメタデータ抽出を行なう必要があるため，効率化のためにメタデータを自動抽出する技術が必要である．本研究では，マンガ画像からの登場キャラクターを検出について検討を行なう．過去の研究において，HOG 特徴量や Deformable Part Model (DPM)などの一般物体検出に用いられる手法がマンガ画像に対しても有効であることが示されている．一方，現在の一般物体検出の分野において，畳み込みニューラルネットワーク(CNN)を用いた手法が既存手法を上回る検出精度を示していることが示されている．本稿では，CNN で画像特徴量学習し，物体検出を行なう手法である Regions with Convolutional Neural Network Features (R-CNN)のマンガ画像への適用について検討を行なった．その結果 R-CNN は DPM と同様に，高い精度でキャラクターの検出が可能であることが分かった．

1 はじめに

近年の電子書籍の普及に伴い，文字情報やオブジェクトといったマンガ内のコンテンツをコード化することによって，電子コミックに新たなサービスを提供するといった提案がなされている．しかし現状の電子コミックコンテンツの多くは単に紙媒体のマンガをスキャンして電子化したものであるため，コード化には人手でメタデータを付与する必要があるため，コスト面が問題となっている．よって電子コミックサービス実用化のために，マンガコンテンツよりメタデータを自動抽出する技術が必要となっている．

本研究では，メタデータ抽出技術の中でマンガ画像内のキャラクター位置検出について検討する．本稿では，畳み込みニューラルネットワークで学習された画像特徴量を用いて物体検出を行なう手法である Regions with Convolutional Neural Network Features (R-CNN)のマンガ画像への適用について検討した．

2 従来研究

マンガ画像よりメタデータを抽出する技術は，現在コマの識別や，キャラクター・フキダシの抽出について研究が行なわれている．

キャラクターの抽出において，マンガキャラクターは主に 2 値の線画によって表現されるため，一般物体と比較すると，認識に使用できる特徴が少ないことや，場面ごとの見た目の変化が大きいといった問題がある．過去研究では，HOG 特徴量と対象物体の各パーツの位置関係を利用した物体検出手法である Deformable Part Model (DPM)がマンガキャラクター検出に有効であることが示されている[1] [2].

一方，近年の一般物体検出分野では，ニューラルネットワークを用いた手法が高い性能を示しているが，この手法がマンガ画像に対しても有効であるかといった検討はまだ行なわれていない．

3 R-CNN

R-CNN は Girshick らによって提案された一般物体検出のアルゴリズムであり，従来の手法と比べ，物体検出の精度と速度が向上していることが示されている[3]．R-CNN による物体検出の流れは以下のようになっている．

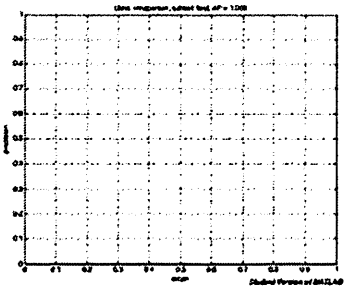
1. Selective Search によって画像内より似たような領域をセグメンテーション化し，物体の候補領域を抽出する．
2. 抽出された候補領域をそれぞれ規定の大きさにリサイズし，畳み込みニューラルネットワークに入力する．
3. 畳み込みニューラルネットワークの出力を特徴量として SVM で分類する．
4. 矩形の座標を回帰して，候補領域のズレを補正する．

R-CNN は一般物体のデータセットである PASCAL VOC 2007 および PASCAL VOC 2010 における実験において，DPM 等の従来手法を上回る精度を記録しており，D-CNN で抽出した特徴量が HOG 特徴量より高い記述能力を持つことが示唆されている．

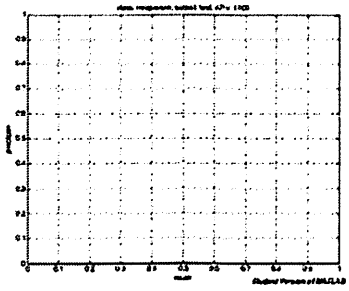
4 実験

R-CNN と既存手法である DPM について，マンガキャラクターに対する検出率の比較を行なった．マンガ作品ではキャラクターごとの特徴変化が大きいため，学習およびテスト画像に含まれるキャラクターの種類によって検出率が影響される．本実験では，3 種類のキャラクターA, B, C それぞれについて R-CNN と DPM の顔検出器を作成し，各キャラクターに対する検出率を比較した．

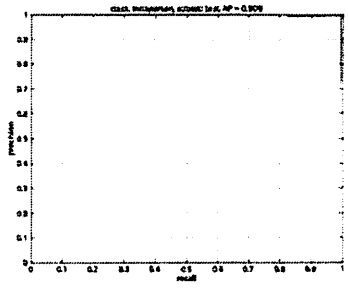
4.1 実験条件



キャラクターA

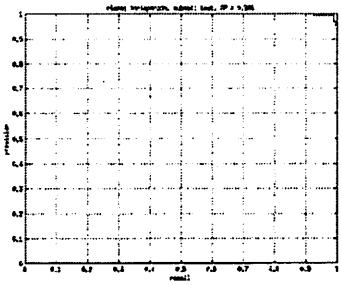


キャラクターB

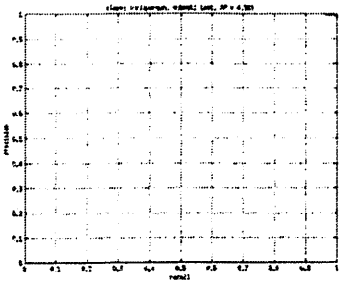


キャラクターC

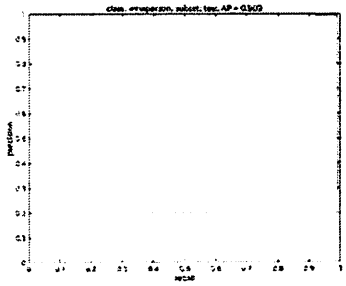
図 1: R-CNN による顔検出結果



キャラクターA



キャラクターB



キャラクターC

図 2: DPM による顔検出結果

表 1: R-CNN と DPM の比較(AP)

	A	B	C
R-CNN	1.000	1.000	0.909
DPM	0.996	0.999	0.909

R-CNN と DPM のプログラムには，Fast R-CNN [4] および voc-release 5 [5]をそれぞれ使用した．ここで Fast R-CNN は R-CNN を改良し検出精度の向上や計算量の削減などを行なった手法である．

検出器の学習では，キャラクターの顔領域を指定したマンガ画像を正例，顔以外の領域を切り出した画像を負例として，各キャラクターの顔検出器を作成した．学習に使用する画像の枚数はそれぞれ正例 200 枚，負例 1000 枚とした．

また，顔領域検出のテストに使用する画像は正例 200 枚，負例 500 枚であり，全て学習に使用したものとは異なる画像とした．

4.2 実験結果

キャラクターA, B, C について，R-CNN による顔検出結果を図 1 に，DPM による顔検出結果を図 2 に示す．図 1, 2 について，縦軸は適合率，横軸は再現率を表している．また，R-CNN と DPM による平均適合率(AP)の比較を表 1 に示す．

実験結果より，R-CNN と DPM ともに 90%以上の精度でキャラクターの検出が可能であることが示された．また，キャラクターA, B については R-CNN の検出率がわずかに DPM を上回った．このことからニューラルネットを用いた物体検出手法がマンガ画像に対しても有効であることが分かった．

5 まとめ

本稿では，R-CNN をマンガ画像に適用した際のキャラクター検出精度について検討を行なった．DPM との比較実験から R-CNN がマンガ画像に対しても有効であることを確認した．今後は両手法のより正確な検出性能を検討するために，学習枚数を変化させた場合における性能の比較を行いたい．

参考文献

[1] 石井大祐, 渡辺裕: マンガからの自動キャラクター検出に関する検討, 情報処理学会研究報告, Vol.2012-AVM-76, No.1, pp.1-5 (2012).

[2] H. Yanagisawa, D. Ishii, H. Watanabe: "Face detection for comic images with deformable part model," IEVC2014, 4A-1, (2014).

[3] R. Girshick, J. Donahue, T. Darrell, J. Malik: Rich feature hierarchies for accurate object detection and semantic segmentation in IEEE conference on Computer Vision and Pattern Recognition, (2014).

[4] R. Girshick: Fast R-CNN, arXiv preprint arXiv:1504.08083, (2015).

[5] R. B. Girshick, P. F. Felzenszalb, D. McAllester: Discriminatively Trained Deformable Part Models, Release 5, http://people.cs.uchicago.edu/~rbg/latent-release5/,

早稲田大学大学院　基幹理工学研究科 情報通信専攻　渡辺研究室
〒169-0072 東京都新宿区大久保 3-14-9 早大シルマンホール 401
Phone: 03-5286-2509
E-mail: bule-cosmo@ruri.waseda.jp

マンガキャラクター検出における学習画像枚数の影響

Effect of Training Image Samples for Comic Character Detection

柳澤 秀彰† 渡辺 裕†
Hideaki Yanagisawa† and Hiroshi Watanabe†

†早稲田大学大学院 基幹理工学研究科 情報通信専攻
†Graduate School of Fundamental Science and Engineering, Waseda University

Abstract For practical use of e-comic services using metadata, the technique extracting metadata automatically from comic images is important. In this paper, we examined the influence of the amount of training images for character detection by DPM and R-CNN. As a result, the detection rate is converged by about 140 positive samples.

1. はじめに

近年，メタデータを利用した電子コミックサービスの実用化を目的として，マンガ画像よりメタデータを自動的に抽出する研究が行なわれている．マンガキャラクター検出について，既存研究より DPM や R-CNN といった手法の有効性が示されている．しかし，これらの手法に対して学習画像枚数が与える影響は未知である．本稿では，学習枚数を変化させた場合における DPM と R-CNN の検出率について検討を行なった．

2. マンガキャラクター検出手法

マンガ画像のキャラクターは主に 2 値の線画によって表現されるため，一般物体よりも認識に使用できる特徴が少ないことや，場面ごとの変化が大きいといった問題がある．既存研究では，物体の概形を記述する HOG 特徴量と，物体のパーツの位置情報を利用した検出手法である Deformable Part Model (DPM)が高い精度を示している[1]．また，ニューラルネットワークから生成される特徴量を用いた検出手法である Regions with Convolutional Neural Network Features (R-CNN)は近年の一般物体検出において既存手法を大きく上回る精度を記録しており，マンガ画像に対しても適用できることが確認できた．

3. 実験

マンガ画像よりキャラクターの正面顔領域を抜き出した画像を正例，顔以外の領域を抜き出した画像を負例とする．キャラクター1 種類について，負例 1000 枚，正例を 20 枚ずつ増加させて DPM と R-CNN の学習を行なった．テスト画像 200 枚に対する検出結果を図 1 に示す．また，4 種類のキャラクターに対して，負例 1000 枚，各キャラクターについて正例を 20 枚ずつ増加させて学習を行なった．各キャラクターについてテスト画像を 200 枚用意し，合計 800 枚の画像に対して検出を行なった結果を図 2 に示す．結果より，正例が 20 枚の段階で検出率は 90%以上となり，キャラクター1 種類につき，約 140 枚で検出率が収束することが分かった．

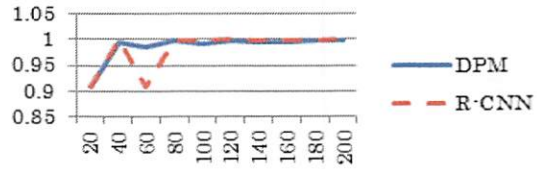


図 1: 1 種類のキャラクターに対する検出率変化(AP)

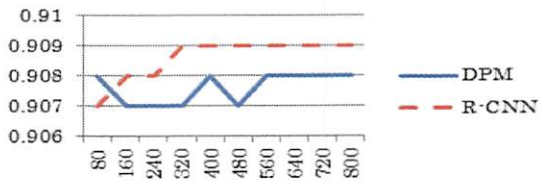


図 2: 4 種類のキャラクターに対する検出率変化(AP)

4. まとめ

DPM と R-CNN について，キャラクター1 種類と 4 種類を学習した場合において，学習枚数が及ぼす影響を検討した．その結果，キャラクター1 種につき，約 140 枚程度を学習することで十分な検出が行なえることが分かった．

5. 謝辞

本研究は JSPS 科研費 25330137 の助成を受けたものである．

文 献

[1] H. Yanagisawa, D. Ishii, H. Watanabe: "Face detection for comic images with deformable part model," IEVC2014, 4A-1, (2014).

†早稲田大学大学院 基幹理工学研究科 情報通信専攻
〒169–0072 東京都新宿区大久保 3-14-9 早大シルマンホール 401
Phone: 03-5286-2509 E-mail: bule-cosmo@ruri.waseda.jp

マンガキャラクターのマルチビュー顔検出に関する検討

Study for Multi-view Face Detection of Characters in Comic Images

柳澤秀彰　　渡辺　裕
Hideaki Yanagisawa　　Hiroshi Watanabe

早稲田大学大学院　基幹理工学研究科
Graduate School of Fundamental Science and Engineering, WASEDA University

1. まえがき

電子書籍市場の中で電子コミックは売上の大部分を占めるコンテンツである．現在の電子コミックのほとんどは単に紙媒体のものをスキャンして電子化したものだが，マンガ画像からメタデータを抽出し，タグ付けを行なうことによって，高度なマンガ検索システムや自動要約作成など様々な機能を付加できる．

既存研究より，マンガキャラクターの正面顔の検出について Deformable Part Model (DPM)や Regions with CNN feature (R-CNN)が有効であることが示されているが[1]，横顔を含んだ顔検出に関する研究は行なわれていない．本稿では，DPM を用いた手法と R-CNN を用いたマンガキャラクターのマルチビュー顔検出について検討を行なった．

2. 検出手法

(1) Deformable Part Model (DPM)

Felzenszalb らは物体の姿勢変化に頑強な物体検出手法として DPM を提案した．DPM は対象を複数のパーツから構成されるツリー構造として捉え，物体の全体および各パーツの形状とパーツ位置のずれに対するコストから評価する．

(2) Regions with CNN feature (R-CNN)

R-CNN は Girshick らによって提案された物体検出手法であり，ニューラルネットの出力を特徴量として利用する．まず，Selective Search によって約 2000 個の候補領域を抽出し，227×227 画素にリサイズした候補領域を畳み込みニューラルネットワーク(CNN)に入力する．次に，CNN の出力を SVM で分類し，候補領域の判定を行なう．

R-CNN は一般物体検出において，DPM などの既存手法を上回る性能を示している．また，Farfade らはマルチビュー顔検出において，Selective Search の代わりにスライディングウィンドウを使用することで検出率が向上することを報告している[2].

3. 実験

DPM と R-CNN について，それぞれキャラクターの正面および横顔を学習したマルチビュー顔検出器を作成し，検出率を比較した．本研究では，マンガ画像からバウンディングボックスで顔領域を指定したものを正例，顔以外の領域を切り抜いたものを負例とした．学習セットは正例を正面顔 400 枚・横顔 200 枚，負例を 1000 枚とした．テストセットは正例を正面顔 400 枚・横顔 200 枚，負例を 2000 枚とした．

DPM について，正面および横顔の左右方向に対応する 4 種類の検出器を作成し，パートフィルタ枚数を 4 枚，Non-maximal suppression (NMS)を 0.1 と設定した．また，R-CNN については正面と横顔を分類せずに学習を行ない，候補領域の抽出に Selective Search を用いた検出器とスライディングウィンドウを用いた検出器の 2 種類を作成した．NMS の値は DPM と同様に 0.1 に設定した．

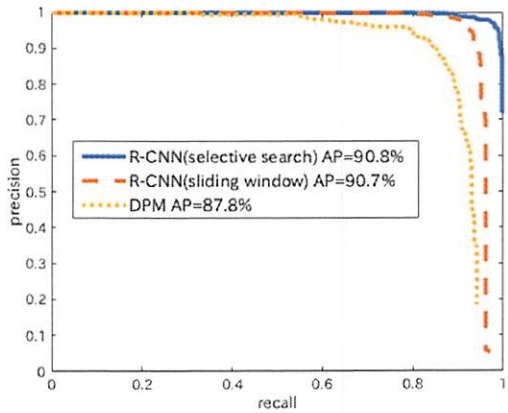


図 1: 顔検出結果の比較

実験結果を図 1 に示す．Selective Search を用いた R-CNN の Average Precision (AP)は 90.8%となり，DPM やスライディングウィンドウを用いた手法より検出率が高くなった．この理由として，

4. まとめ

本稿では DPM および R-CNN を用いたマンガキャラクターのマルチビュー顔検出について検討した．その結果，Selective Search を利用した R-CNN による検出が最もマンガ画像に適していることが分かった．

謝辞　本研究は JSPS 科研費 25330137 の助成を受けたものである．

参考文献

[1] 柳澤秀彰, 渡辺裕: “"R-CNN を用いたマンガキャラクター検出に関する一検討", 映像メディア処理シンポジウム (IMPS), (2015).

[2] S.S. Farfade, M. Saberian, L. Li: “Multi-view Face Detection Using Deep Convolutional Neural Networks”, In International Conference on Multimedia Retrieval, arXiv:1502.02766v3, Apr 2015.