クラウドソーシングによる食事画像認識モデルの自動構築

大澤 翔吾† 柳井啓司‡

†電気通信大学電気通信学部 情報工学科 〒182-8585 東京都調布市調布ヶ丘 1-5-1

‡電気通信大学 大学院情報理工学研究科 総合情報学専攻 〒182-8585 東京都調布市調布ヶ丘 1-5-1

E-mail: †oosawa-s@mm.inf.uec.ac.jp, ‡yanai@cs.uec.ac.jp

あらまし 食事画像認識において認識カテゴリを増やすためには大量の bounding box 付き画像が必要である．画像に bounding box を付与する作業には非常に時間がかかるため, 近年ではこの作業をクラウドソーシングすることが多い．しかし，bounding box の描画作業を大量にクラウドソーシングするには多額の費用がかかるため，bounding box 付与作業を依頼する画像を厳選し，コストパフォーマンスを向上させる手法が必要である．本研究では，クラウドソーシングを利用した学習データ構築と，認識モデル学習を交互に行うことによって，より効率的に食事画像認識のための学習データセットを構築する手法を提案する．

キーワード 画像認識，クラウドソーシング，認識モデル自動構築

1. はじめに

近年，クラウドソーシングを機械学習に応用する研究が増加している．クラウドソーシングとは，インターネットを通じて不特定多数の人に仕事を依頼することである[1]．機械学習分野の研究では，学習データ，特に教師あり学習に用いる教師データ収集のためにクラウドソーシングを利用することが多い．画像認識を例に挙げると，与えられた画像とカテゴリに対し，各画像がどのカテゴリに属するかを判断する仕事を依頼し, 認識モデルの学習に必要な学習データを収集する．従来は, こうした作業を研究者が自ら行なっていたが，クラウドソーシングの利用により，認識モデルの自動構築が可能となった．

認識モデルの自動構築手法の一つとして，学習データ収集と認識モデル構築を交互に行う手法（ループ学習）が挙げられる．ループ学習には，学習の際にアノテーション結果を得るサンプルを選択できるため，モデルに有益なサンプルを選ぶことが出来れば，クラウドソーシング費用を削減できるというメリットがある．

本研究では，クラウドソーシングを用いたループ学習で認識モデルを自動構築する際，アノテーション作業を依頼する画像の選出戦略を 3 つ考案し，これらの戦略が収集されたデータセットと構築されたモデルに及ぼす影響を考察する．

2. 関連研究

2.1. クラウドソーシングと物体検出

Vijayanarasimhan らの研究[1]ではクラウドソーシングを用いて物体検出器の学習に必要なデータを収集している．この研究の提案手法の概要を図 1 に示す．この手法では，以下に示す手順で処理が進む．

1. 少数の学習画像で線形 SVM の学習を行う．
2. Flickr からキーワード検索で画像を収集する．
3. 収集した画像の中から最も SVM の超平面に近いサンプルを Hyperplane-Hashing[2]によって選ぶ．
4. 選んだ画像に bounding box を描画する作業をクラウドソーシングする．
   作業は計 10 人の作業者に依頼し, 作業結果を mean shift クラスタリングして得られたクラスタ中心を作業者のコンセンサスとして採用する．
5. 以上の作業で得られた bounding box 付き画像を含めて線形 SVM の再学習を行う．
6. 2.～5.の処理を繰り返す．

この手法の画像選出戦略は，初期状態の線形 SVM の出力結果を過大評価している．線形 SVM の超平面に近いサンプルの教師信号を得るという戦略は，SVM にとって曖昧な線形 SVM の学習が進んだ後には有効である．しかし，少数の学習データしかもたない線形 SVM の場合, その超平面から遠いサンプルに対する出力結果が信頼できない場合もあるため，超平面から遠いサンプルのアノテーション結果も得る必要がある．

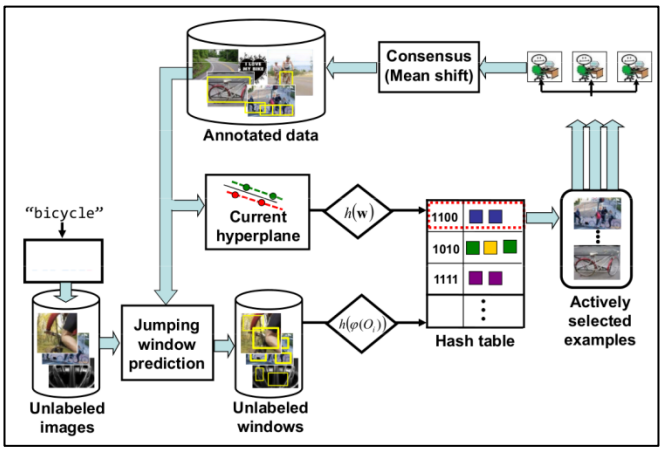

図 1. Vijayanarasimhan らの手法の概要

2.2. 作業者の分析

クラウドソーシング費用抑えるための手法として，ループ学習のほかに，スパマーを検出する手法や，提出されたデータの品質を定量化する手法などが提案されている．

Raykar らの研究[3]では，多値分類タスクにおける作業者の「スパマーらしさ」を定量化する手法を提案している．この手法では，スパマーが対象に付与するラベルの分布は，対象の真のラベルに依存しないという観点からスパマーらしさを計算している．また，この手法用いて，2 値分類タスクにおける悪意のある作業者を検出できることも示されている．悪意のある作業者とは，わざと正解とは逆のラベルを付与する作業者のことである．悪意のある作業者のラベルを逆にすることで正解のラベルが得られるため，より多くの高品質なアノテーションデータが手に入る．

Hsueh らの研究[4]では，作業者が付与したラベルの分散などの情報をもとに，仕事の難しさや作業者の能力を定量化し，それらを用いてアノテーション品質を計算している．さらに，この手法によりアノテーション品質が高いとされたデータのみを用いた方が，そうでないデータを用いるより認識モデルの認識精度が向上することも示されている．

これらの研究においては，クラウドソーシングする作業は多値分類やランキングなどの簡単な作業である．bounding box の描画といった比較的複雑な作業に以上のような簡単な作業を対象にした手法を適用することは難しい．

3. 提案手法の概要

提案手法の処理順序を以下に示す．

1. 認識モデルの初期学習（図 **2**）
   (a) キーワード検索を用いて画像を収集する．キーワードには，認識させたい対象のカテゴリ名を指定する．
   (b) 収集した画像に bounding box を付与する作業をクラウドソーシングする．
   (c) クラウドソーシングした仕事の作業結果を用いて認識モデルの学習をを行う．
2. ループ学習（図 **3**）
   (a) 認識モデルの初期学習の際と同様の手順で画像を収集する．
   (b) 収集した画像のうち，現在の認識モデルを用いたスクリーニングを行い，bounding box 付与作業を依頼する画像を選出する．
   (c) 選び出した画像に bounding box を付与する作業をクラウドソーシングする．
   (d) クラウドソーシングした仕事の作業結果を用いて認識モデルを更新する．

なお，本研究では，画像の収集に Flickr API を，クラウドソーシングに Amazon Mechanical Turk（MTurk）を，認識モデルに Deformable Part Model[5]（以下，DPM と記す）を用いる．

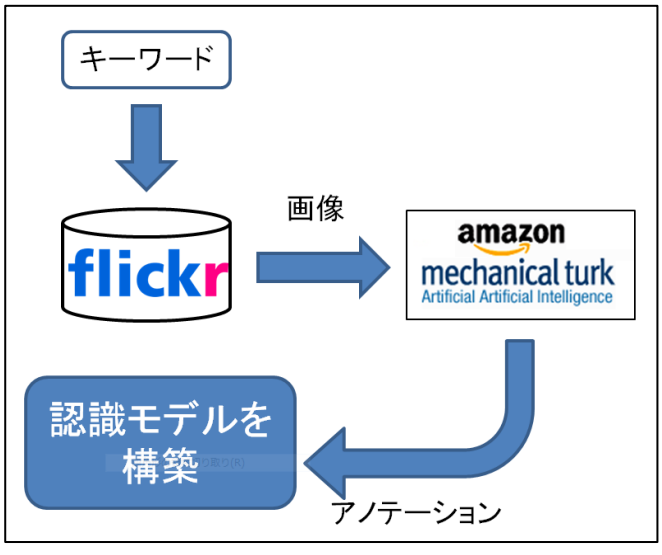

図 2. 提案手法の概要（認識モデルの初期学習）

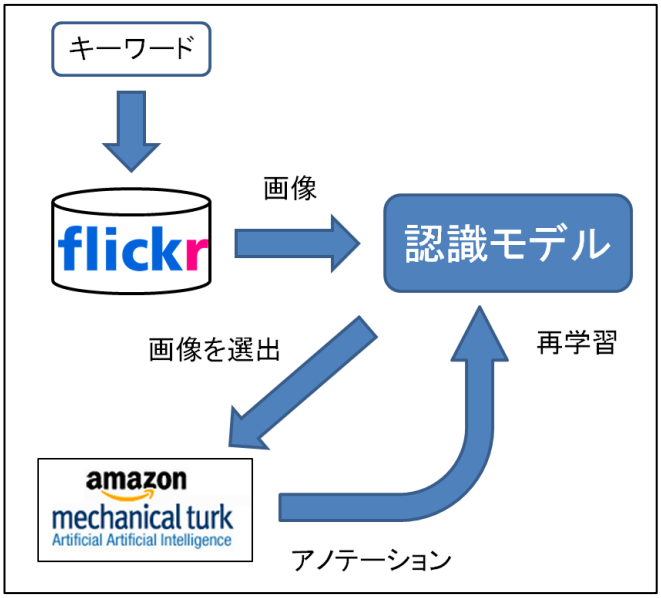

図 3. 提案手法の概要（ループ学習）

4. 提案手法の詳細

4.1. 認識モデルの初期学習

4.1.1. ポジティブ画像候補の収集

認識モデルを学習させるためのポジティブ画像候補を Flickr API のキーワード検索を用いて収集する．

4.1.2. Bounding box 描画作業の依頼

収集した画像に bounding box を付与する作業を Amazon Mechanical Turk にクラウドソーシングする．作業者には，対象のカテゴリ名と画像が与えられる．画像内に対象が写っている場合は，その周りを囲う bounding box を描画する．描画操作はマウスのドラッグ＆ドロップで行う．そうでない場合は，「対象なし」としてチェックする．対象なしとしてチェックする操作は画像を右クリックすることで行う．作業ページの例を図 4 に示す．

4.1.3. コンセンサスの生成

クラウドソーシングした作業が完了した後，以下の手順でコンセンサスを生成する．まず，作業結果のうち，bounding box の面積が一定以下のものを除外する．作業者の中には画像に対して全く操作をせずに作業結果を提出したり，でたらめな bounding box を描画したりする者が存在する．この処理はそうした作業者を取り除くためのものである．次に，残った作業結果に対し，「対象なし」かどうかの多数決をとる「対象なし」が過半数を上回った場合，その画像に対するコンセンサスを「対象なし」とする．そうでない場合，作業者が描いた bounding box の座標の平均をその画像に対するコンセンサスとする．

図 4. MTurk の作業ページの例

4.1.4. ネガティブ画像の収集

Flickr API のキーワード検索を用いてネガティブ画像を収集する．キーワードには，学習を行う認識モデルのカテゴリ名以外の単語を指定する．

4.1.5. Deformable Part Model の学習

以上の処理で得た bounding box 付きポジティブ画像とネガティブ画像を用いて DPM の学習を行う．この際，bounding box 付与作業を依頼した画像のうち，コンセンサスが「対象なし」となっているものをネガティブ画像に追加する．

DPM は，HOG 特徴量を用いた線形 SVM ベースの物体検出器である．DPM の学習には，bounding box 情報付きのポジティブ画像と，bounding box 情報なしのネガティブ画像が必要である．学習済みの DPM に画像を入力すると，物体の bounding box の位置とその評価値が計算される．評価値とは，入力画像の「ポジティブらしさ」を定量化したもので，評価値が大きいほどポジティブらしさの信頼性は高い．DPM は学習データから計算される閾値を持っており，入力画像の評価値が閾値より大きい場合，その入力画像をポジティブとして出力する．そうでない場合，ネガティブとして出力する．

4.2. ループ学習

DPM の初期学習を行った後，DPM のループ学習を行う．ループ学習では，認識モデル初期学習の際と同様の手順でポジティブ画像とネガティブ画像を追加するが，ポジティブ画像候補の扱いが初期学習の際と異なる．ループ学習の場合，Flickr から収集したポジティブ画像候補をスクリーニングし，候補の半分のみに bounding box を付与する．スクリーニング戦略には以下の 3 つを採用し，各戦略が DPM に与える影響を考察する．

- ランダムに選出（Random）
  この戦略はベースラインとしての役割を果たす．
- 評価値が DPM の閾値に近い順に選出（Near）
  この戦略は Vijayanarasimhan らの研究[1]で採用されている．現在の DPM において，評価値が非常に良いものや悪いものに対する出力結果は信頼できるので，モデルにとって曖昧なサンプルの教師データを得るべきであるという仮定に基づいている．
- 評価値が DPM の閾値から遠い順に選出（Far）
  この戦略は，初期化された DPM の出力結果はあまり信頼できないので，評価値がモデルの閾値から遠いサンプルの教師データを得るべきであるという仮定に基づいている．評価値が閾値から遠い画像に対する出力結果が間違っていた場合，その情報を認識モデルに反映させれば，認識モデルは大きく変更される．

Near 戦略と Far 戦略の模式図を図 5 に示す．Near 戦略は，評価値が閾値付近の画像のアノテーションを依頼し，Far 戦略は，評価値が閾値から遠い画像のアノテーションを依頼する．

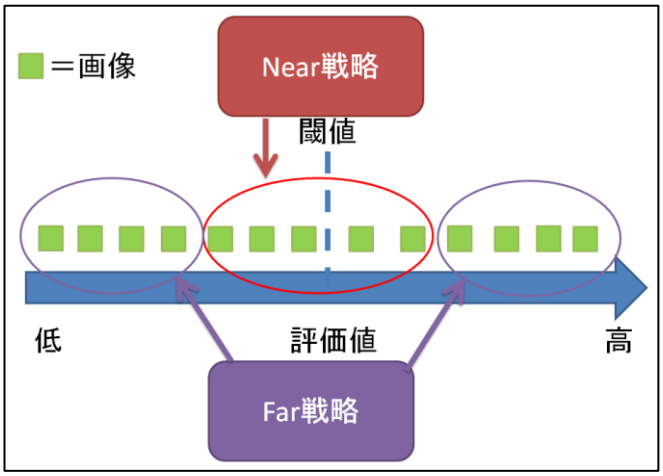

図 5. Near 戦略と Far 戦略の模式図

5. 実験

提案手法のループ学習における画像選出戦略が認識モデルに与える影響を確認する実験を行う．本実験では，画像選出戦略として 3 章で示した 3 つの戦略を採用し，各ループにおいて，DPM によって選出された画像のうちポジティブ画像が占める割合と，テスト画像に対する DPM の F 値を求めた．本実験で検出対象とするカテゴリは「牛丼」「お好み焼き」「ラーメン」「たい焼き」「肉じゃが」の 5 つである．

5.1. データセット

5.1.1. 学習画像

本実験では，ポジティブ画像候補を収集する場合，検出対象のカテゴリ名をキーワードに指定し，各ループにおいて 100 枚収集する．ポジティブ画像候補 100 枚から選出された 50 枚のうち，ポジティブ画像であるという作業者のコンセンサスを得られたものがポジティブ画像に追加され，そうでないものはネガティブ画像に追加される．加えて，検出対象以外の 4 つのカテゴリ名をキーワードに指定し，各ループにおいて，カテゴリ毎に 100 枚，計 400 枚を収集し，ネガティブ画像に追加する．DPM の初期学習を行う際には，作業を依頼する画像を選出するのに必要な DPM がないため，収集したポジティブ画像候補 100 枚すべてに対してアノテーション作業を依頼する．

5.1.2. テスト画像

本実験で用いる DPM の精度評価用テスト画像は Google 画像検索から手動で収集する．カテゴリ毎に 50 枚，計 250 枚の画像を収集する．各カテゴリ 50 枚のうち，半分が単品画像，もう半分が複数品画像である．単品画像とは，指定したカテゴリの食品以外に何も写っていない画像のことを指す．複数品画像とは，指定したカテゴリの食品以外にも食品が写っている画像のことを指す．

5.1.3. MTurk へのアノテーション作業依頼

本実験では，10 枚の画像にアノテーションを付加する作業を依頼する．報酬は 1 作業あたり$0.05 とし，カテゴリ毎に計 500 枚の画像のアノテーションを依頼した．1 つの作業を 5 人の作業者に依頼し，4 章で示した方法でコンセンサスを生成する．本実験での MTurk 利用条件の概要を表 1 に示す．

5.2. 実験結果:DPM の選出画像評価

DPM によって選出された画像のうち，ポジティブ画像が占める割合を図 6 に示す．どのカテゴリにおいて

表 1. MTurk 利用条件の概要

項目	総数	1 カテゴリあたり
作業数	250	50
画像枚数	2500	500
報酬	$62.5	$12.5

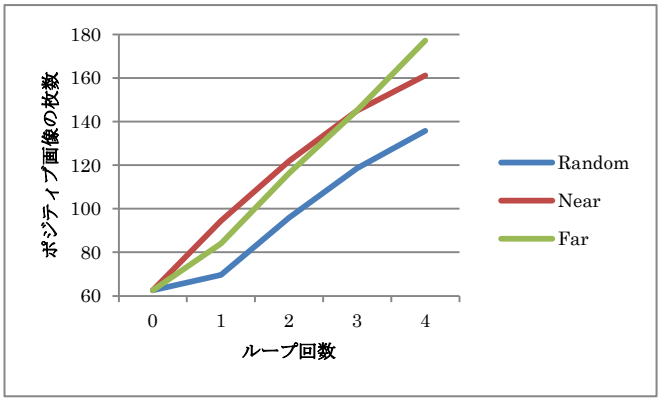

図 6. 収集したポジティブ画像の枚数

も，Random 戦略は最も割合が低く，Near 戦略はループを重ねると割合が減少する傾向にあるが，Far 戦略に減少傾向は見られない．これは，初期状態の DPM は学習画像が少ないため，評価値が閾値に近いポジティブ画像が存在するが，学習が進むにつれてポジティブ画像の多くは評価値が高くなり，閾値から遠ざかるからだと考えられる．ゆえに，Near 戦略では，評価値が閾値から遠ざかったポジティブ画像を選出できない一方，Far 戦略では評価値が高いものと低いものを選出するため，学習が進んでも評価値が高くなったポジティブ画像を選出できる．よって，より多くのポジティブ画像を収集することが望ましいという観点からは，Far 戦略が最も優れた戦略であると言える．

5.3. 実験結果:DPM の F 値

5.3.1. 評価方法

本実験では，各テスト画像に対する DPM の出力を以下の手順で評価する.評価した結果に対して適合率，再現率，F 値を以下に示す式によって計算し，評価を行う．

- DPM が bounding box を出力した場合
  DPM が出力した bounding box と正解データの bounding box が 50%以上オーバーラップしていたら true positive とする．そうでなければ false positive とする．
- DPM が bounding box を出力しなかった場合
  テスト画像の正解データがポジティブならば false negative とする.そうでなければ true negative とする．

$$(\text{適合率}) = \frac{tp}{tp + fp}$$

$$(\text{再現率}) = \frac{tp}{tp + fn}$$

$$(F\text{値}) = \frac{2 \times (\text{適合率}) \times (\text{再現率})}{(\text{適合率}) + (\text{再現率})}$$

ここで，tp，fp，fn，はそれぞれ，true positive の枚数，false positive の枚数，false negative の枚数である．

5.3.2. 評価結果

各ループにおける DPM の F 値を図 7 に示す．ループ学習においては，ループを重ねると F 値は上昇することが期待されるが，どの戦略においても，F 値は横ばい傾向にあることが示されている．これは，アノテーションデータに含まれるノイズが原因で，ループを重ねると，再現率は上昇するが適合率は減少するからである．アノテーションデータに含まれるノイズについては次章で述べる．

6. 考察

6.1. アノテーションデータに含まれるノイズ

前章で述べた実験では，ループを重ねると F 値は上昇することが期待されるが，どの戦略を用いても，ループを重ねると適合率が低下し，再現率は上昇するため，F 値は上昇せず横ばいの傾向にあることが示された．この原因として，アノテーションデータに含まれるノイズが挙げられる．

たい焼きを例に挙げると，図 8 の下段に示すような画像がポジティブ画像に含まれてしまうことが確認された．こうした画像は，図 8 の上段に示すような通常のたい焼き画像とは極端に写り方が異なり，学習画像としては不適切である．

一方で，通常のものとは極端に写り方が異なる画像が収集されなかったカテゴリも適合率が高いわけではない．それには，画像のバリエーションが大きく影響していると考えられる．例えばカテゴリ「牛丼」の場合，図 9 に示すように，様々な具が乗った牛丼の画像が収集される．一方，カテゴリ「お好み焼き」の場合，図 10 に示すように,写り方が牛丼などに比べて均一である．写り方が均一である場合，そうでない場合と比べて認識モデルの性能は良くなることが期待される．本実験では，どのループ・戦略でも一貫して牛丼よりもお好み焼きの方が良い性能を示している．

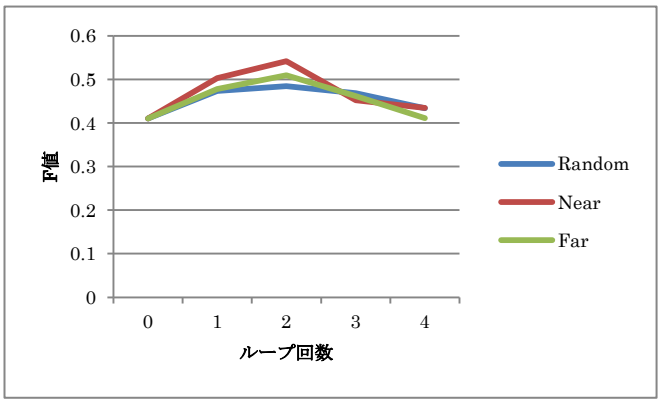

図 7. DPM の F 値

アノテーションデータにノイズが含まれるのは，アノテーション作業を依頼した作業者の作業精度が十分でないからである．しかし，本研究で依頼した作業に取り組んだ作業者は 9 割以上がインド人で，残りは少数のアメリカ人やフィリピン人などであり，日本人はほとんどいない．そのため，作業者は日本食を実際に目にしたことはほとんどなく，画像検索や Wikipedia などで情報を収集した後，作業に取り組んでいるものがほとんどであると思われる．こうした作業者に対して，非常に高いアノテーション精度を期待することは難しい．

6.2. ノイズ除去にまつわる問題

前述した 3 種類のノイズは認識モデルに悪影響を与えるため，検出して取り除く必要がある．クラウドソーシングから得たアノテーションデータにノイズ除去を施す手法が提案されているが，どれも多値分類といったタスクを対象にしており[3,4]，bounding box の描画といった複雑なタスクに適用することは困難である．また，複雑なタスクに適用できたとしても，ノイズ除去によって手に入るアノテーションデータの数が少なくなるため，より多くの作業を依頼しなければならず，クラウドソーシング費用が増加してしまう．そのため，ノイジーなアノテーションデータの品質を向上させる手法が必要である．

アノテーションデータの品質向上手法の一つとして，作業者が取り組んだ作業数に注目する方法が挙げられる．前章の実験で依頼した全 250 タスクにおける作業者が取り組んだ作業数の頻度分布を図 11 に示す．実験で依頼した作業に取り組んだ全 80 人のうち，50 人以上は 10 タスク未満しか取り組んでいない．取り組んだ作業数が 10 タスク未満である作業者は，図 12 に示すように，ほとんどが 1 回しかタスクに取り組んでいない．1 回しかタスクに取り組んでいない作業者には，タスクの意図を理解出来なかった者や，まともに作業

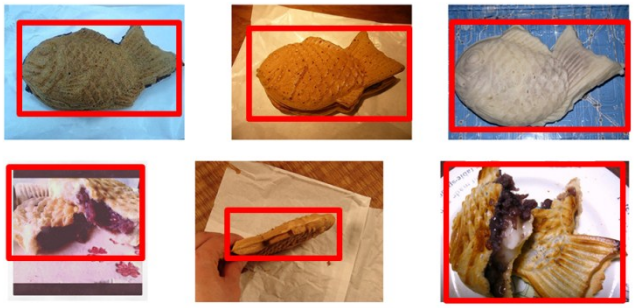

図 8. 通常の画像(上段)と移り方が異なる画像(下段)

図 9. 収集した画像の例（牛丼）

図 10. 収集した画像の例（お好み焼き）

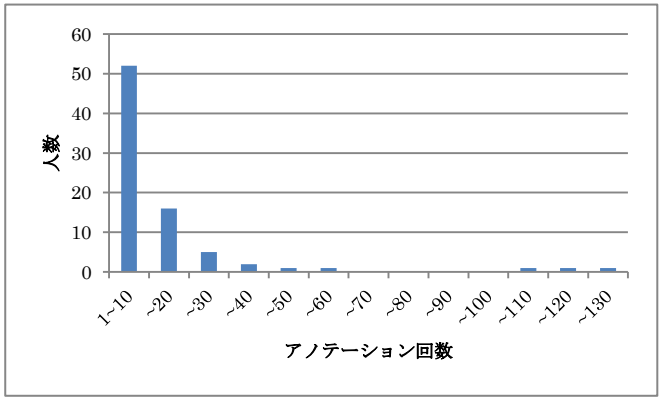

図 11. アノテーション回数の分布

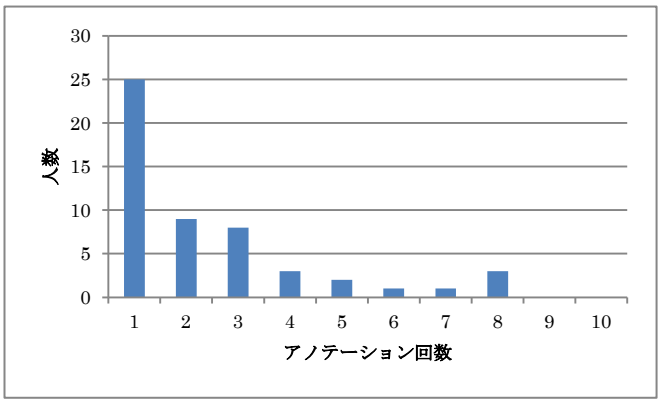

図 12. アノテーション回数の分布（10 回以下）

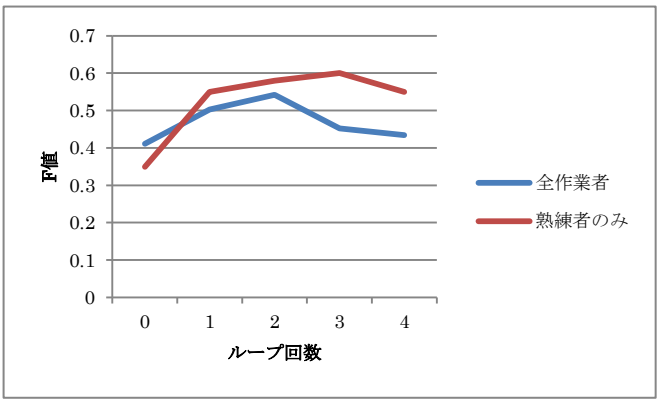

図 13. 全作業者と熟練者のみの比較（Near 戦略）

せずに結果を提出する者が多く含まれる．一方で，タスクに 100 回以上取り組んでいる作業者もわずかだが存在する．こうした作業者は「プロの作業者」だと思われる．Amazon Mechanical Turk には，タスクに取り組むことで得られる報酬で生計を立てているものがいる．こうした作業者は「プロ」と呼ばれ，作業のクオリティが高いことが知られている．

アノテーション回数が DPM の精度に及ぼす影響を考察するため，アノテーション回数が 10 回以上の作業者（熟練者）のみを用いて前章と同様の実験を行った．その結果を図 13 に示す．すべての作業者を用いた場合と比べて全体的に F 値は上昇しているが，横ばい傾向にあることは変わらない．作業者の結果をスクリーニングした場合，アノテーション精度はスクリーニングしない場合に比べて改善したが，依然ノイズが含まれており，ループを重ねると適合率が減少し再現率が増加するという同様の現象が確認された．

本実験でアノテーション作業を依頼した 2500 枚の画像に対するアノテーション精度を図 14 に示す．図 14 の「日本人」の結果は，同じデータに対する bounding box 付与作業を日本人の大学生に依頼した結果である．全作業者の精度よりも熟練者のみの精度の方が高いが，熟練者のみの場合でも依然としてアノテーションデータにノイズが含まれている．日本人のみがアノテーションした場合と比べるとその差は顕著であり，全作業者の場合と日本人のみとの差は 10%以上ある．

本研究でアノテーション作業を依頼した 5 カテゴリは文化依存性の高いカテゴリであり，更に依頼した画像には学習に適さないものも多く含まれている．このような難易度の高い作業は，日常的に作業を行なっている作業者にとっても難しいことを図 14 は示している．よって，この作業に高いアノテーション精度を求めるには，作業者の負担を減らすように作業を依頼したり，収集した画像からアノテーションしやすい画像

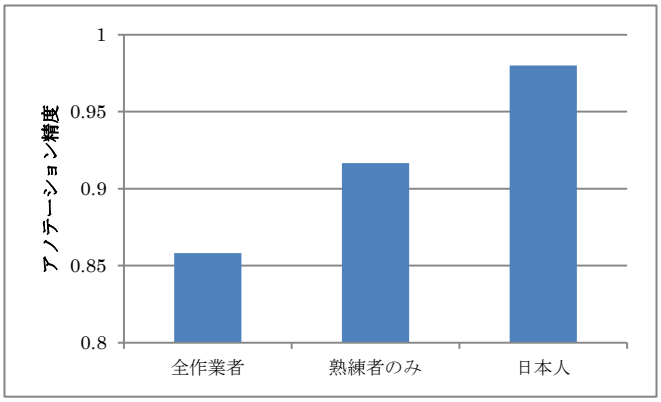

図 14. 作業者のアノテーション精度

を選出したりするなどの工夫が必要であることが分かる．これらの手法の考案は今後の課題としたい．

7. 結論

本研究では，ループ学習を用いた物体検出器の自動生成の際，学習画像に追加する画像を選択する 3 つの戦略（Random,Near, Far）を比較した．その結果，どの戦略でもモデルの性能に有意差は見られなかったが，Far 戦略を用いた場合，他の戦略に比べてポジティブ画像が多く集まることを確認した．本論文で行った実験では，ループを追うごとに DPM の性能が向上していくことが期待されていたが，必ずしもそうならないことを確認した．これは，アノテーションデータにノイズが多く含まれることが原因だと思われる．

熟練した作業者の作業結果のみを用いてもノイズは除去しきれなかったので，仕事の依頼の仕方や画像のスクリーニング手法の考案が今後の課題である．

参 考 文 献

[1] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *Proc. of IEEE Computer Vision and Pattern Recognition*, pp.1449–1456, 2011.

[2] P. Jain, S. Vijayanarasimhan, and K. Grauman. Hashing hyperplane queries to near points with applications to large-scale active learning. *Advances in Neural Information Processing Systems (NIPS)*, Vol. 23, pp. 928–936, 2010.

[3] V. C. Raykar and S. Yu. Ranking annotators for crowdsourced labeling tasks.*Advances in Neural Information Processing Systems (NIPS)*, pp. 1809–1817, 2011.

[4] P. Y. Hsueh, P. Melville, and V. Sindhwani. Data quality from crowdsourcing: A study of annotation selection criteria. In *Proc. of NAACL HLT Workshop on Active Learning for Natural Language Processing,* pp. 27–35, 2009.

[5] P.F. Felzenszwalb, R. B. Girshick, D. McAllester, and 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.