既存カテゴリの活用とクラウドソーシングによる食事画像データセットの自動拡張

河野 憲之[1,a)]　柳井 啓司[1,b)]

1. はじめに

我々は，以前より食事画像認識システムを構築してきた [1]．そのための食事画像データセットは手動で構築してきた．本研究では，より実用的な認識システムの実現に向けて，データセットの自動拡張を目的としている．食事は，カテゴリ間の視覚的類似度が高く，fine-grained なカテゴリ群である．そのため，データセットを拡張する際に，既存の食事画像データセットの視覚的情報を活用することで，効率よく未知カテゴリの画像収集を行う．そして，Amazon Mechanical Turk (AMT) によるクラウドソーシングにより，画像に注釈付けを行うことでデータセットの自動拡張を行う．図 1 に，提案フレームワークを示した．この図に示したように，入力は食事カテゴリを表すキーワードで，最終的な出力は，図の左下にあるようなバウンディングボックス（BB）付きの食事画像セットである．

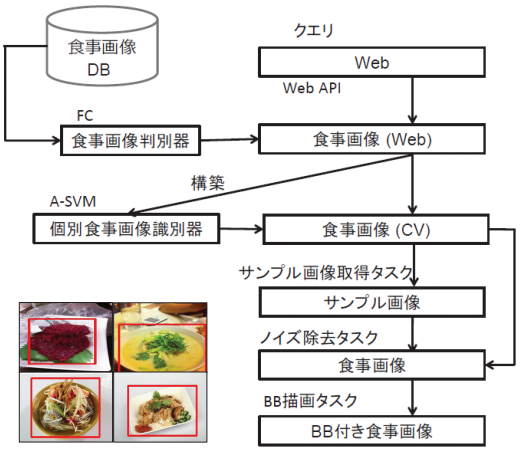

図 1　提案フレームワーク．

2. 画像判別

2.1 食事画像判別器

キーワードにより Web から収集しただけの画像には，ノイズが多く含まれる．そのため，収集した画像がキーワードで指定した食事画像であるかを判別する．

まず，既存の 100 種類の食事画像データセットを用い，100 種類の判別器を構築し，評価を行い混同行列を得る．画像特徴量は [1] で使用された特徴量をパラメタを変更して使用した．次に，混同行列を用い，いくつかのグループに分割する．他のカテゴリと混同が多い条件として，正しく分類できた数の x%（x=40）以上あるカテゴリに誤分類されていた場合，それらのカテゴリ同士を結合する．結合されるカテゴリが存在しない場合は，x を半減させる．これらを再帰的に行い，13 の食事グループが作成された．そして，正例はそのグループに属する食事画像，負例は手動で収集した画像として，13 の判別器を構築した．最終的な評価値は，構築した 13 の判別器の最大反応値とした．この食事画像判別器は，SVM で判別が難しい食事グループを結合し，その食事グループに属するか，食事であるかを評価している．未知カテゴリにおいても，同じグループに属するかを評価する．この評価値を食事らしさとして扱い，評価値の高い，より食事画像らしい画像を使用する．

2.2 個別食事画像判別器

食事画像判別器だけでは，食事画像以外はほぼ除去できるものの，一部に対象の食事とは異なる種類の食事画像が含まれている場合がある．そのため，対象の食事画像であるかを判別する．

未知カテゴリについて，画像を収集後，その画像群を用い，未知カテゴリに特化した判別器を構築する．食事らしさのスコア上位の画像群を，画像類似度によってランキングをし，その上位の画像群を疑似正例の学習画像として，個別食事画像判別器を構築する．このとき，既存食事データセットの視覚的特徴を活用するために，Adaptive SVM（A-SVM）[3] により転移学習を行った．未知カテゴリのソースドメインは，食事画像の判定に最も貢献した食事グループとした．ターゲットドメインの負例は，食事らしさのスコアが閾値以下上位の画像群とした．

3. クラウドソーシング

本研究では，食事画像認識のための学習データ構築を目的としているため，精度 100%が不可欠である．そこで，不

[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
[a)] kawano-y@mm.inf.uec.ac.jp
[b)] yanai@cs.uec.ac.jp

完全な自動画像判別のあとに，クラウドソーシングを用いて，さらにノイズ画像の除去を行い，最終的に BB 付きの食事画像セットを構築する．

一般にクラウドソーシングでは，同時に複数の種類の作業を依頼すると作業精度が落ちるため，単一種類の作業に全体の作業を分割し，クラウドソーシングに依頼するのが望ましいと言われている [2]．この経験則に従い，クラウドソーシングの作業を 3 ステップに分割することとする．

3.1 サンプル画像の取得

インド人の割合が大きい AMT のワーカーは対象の食事について無知であることが考えられる．そのため，ノイズ画像除去タスク，BB 描画タスクで提示する小数の高品質なサンプル画像を取得する．

個別食事判別器により食事らしさのスコア上位の画像群を提示し，一般的な対象の食事画像を選択させた．ワーカーには，画像検索サイトのリンクを設置し学習させた．実験では，1HIT を 0.06 ドルとし，各 HIT を 5 人のワーカーに依頼した．

3.2 ノイズ画像の除去

画像認識により，除去できなかったノイズ画像を人の力を借りることで除去する．ワーカーの仕事量を減らすために，ノイズ画像を除去する専用のタスクを設計した．

個別食事判別器により食事らしさのスコア上位の画像群からランダムに提示し，対象の食事画像であるかそうでないかをチェックさせた．複数のワーカーからの結果の結合は多数決とした．実験では，1HIT を 0.03 ドルとし，各 HIT を 5 人のワーカーに依頼した．

3.3 バウンディングボックスの描画

ノイズ画像除去タスクにより，ノイズでないと判定された画像群に対し，BB を付与する．

対象の食事画像群をランダムに提示し，対象の食事画像であれば，BB を描画させた．また，このタスクにおいても，ノイズ画像であれば除去させた．複数のワーカーからの結果の結合は，複数のワーカーが描画した BB が閾値以内に存在していることとした．実験では，1HIT を 0.05 ドルとし，各 HIT を 4 人のワーカーに依頼した．

4. 実験

本節では，100 種類の新たな食事に対して，画像を収集し，AMT により BB を付与し，データセットの拡張を行う．画像は，Web API[*1][*2][*3]を用いて収集した．

画像のノイズ除去方法と拡張された部分のデータセットにおける食事画像集合の適合率，コスト（賃金）と回収率の評価を行う．回収率とは，AMT に用いた画像枚数に対して，ワーカーが対象の食事画像であるとして回収された画像枚数の割合と定義する．ノイズ除去方法には，食事画像判別器（FC），個別食事画像判別器（A-SVM），ノイズ除去タスク（NR Task）の組み合わせを変化させた．

[*1] http://flickr.com
[*2] https://twitter.com
[*3] http://bing.com

表 1 各方法における，回収率と 100 枚の BB 付き食事画像を得るために要した平均コスト （ドル）

	ノイズ除去		BB 描画		総計
	回収率	コスト	回収率	コスト	コスト
FC	-	-	64.2	3.11	3.11
FC + A-SVM	-	-	74.7	2.68	2.68
FC + A-SVM + NR Task	80.9	0.74	86.7	2.31	3.16

表 2 3 つの方法によるデータセットにおける食事画像集合の適合率

	適合率	gain
FC	91.10	-
FC + A-SVM	94.19	+3.09
FC + A-SVM + NR Task	97.83	+3.64

表 1 に，各方法における回収率と 100 枚の BB 付き食事画像を得るために要した平均コスト（賃金）を示した．食事画像判別器に加え，個別食事画像判別器により再評価をすることで，AMT に用いる画像の適合率が上昇したため，回収率も向上し，14%低コストであった．さらに，ノイズ除去タスクを加えた場合，BB 描画タスクでの回収率は上昇した．だが，ノイズ除去タスクのコストを考慮する必要があるため，結果的にはコストが少し多くかかった．次に，表 2 に，各方法によって拡張された部分のデータセットにおける食事画像集合の適合率を示した．個別食事画像判別器を追加，ノイズ除去タスクを追加するとそれぞれ適合率が上昇した．これは，AMT の性能が完璧でないため，AMT によって誤った画像がデータセットに追加されることを防ぐことができたためである．また，ノイズ除去タスクを追加することで，ワーカーの仕事が分散されたため，質が向上したためである．

5. まとめ

本研究では，2 段階の食事画像判別器によるノイズ画像の除去，および，クラウドソーシングにより，食事画像データセットの自動拡張を行った．今後も拡張を続け，1024 種類の食事画像データセットを構築したいと考えている．

参考文献

[1] Kawano, Y. and Yanai, K.: FoodCam: A Real-time Food Recognition System on a Smartphone, *Multimedia Tools and Applications* (2014). (in press).
[2] Noronha, J., Hysen, E., Zhang, H. and Gajos, K. Z.: Platemate: crowdsourcing nutritional analysis from food photographs, *Proc. of ACM Symposium on User Interface Software and Technology* (2011).
[3] Yang, J. and Yan, R.and Hauptmann, A. G.: Cross-domain video concept detection using adaptive svms, *Proc. of ACM International Conference Multimedia* (2007).