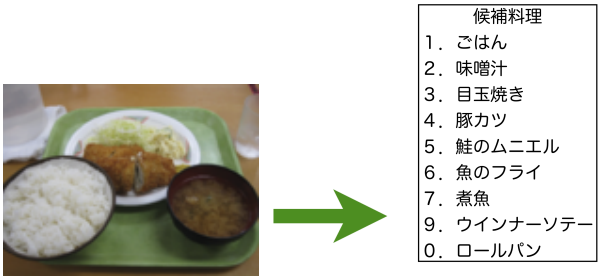
料理間の共起関係を考慮した食事画像認識

松田裕司† 柳井啓司†

† 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: †matsuda-y@mm.cs.uec.ac.jp, ††yanai@cs.uec.ac.jp

あらまし 本研究では，料理間の共起を用いて複数の料理を含む画像の認識精度の改善を行う．提案手法では，食事画像データベースや Web 上のテキストから料理間の共起確率を求め，Manifold Ranking を用いて，SVM による評価値の再ランキングを行った．複数品目を含む画像に対して，100 種類の料理について分類を行い性能の評価を行ったところ，10 個の候補を表示したときに，データベースから共起確率を求めた場合に，従来手法と比べて 7.4 ポイント向上し，63.33%の分類率，Web から共起確率を求めた場合に，従来手法と比べ 0.35 ポイント向上し，56.27% を達成し，複数品目を含む画像に対して共起関係を用いることが有効であることが示された．

キーワード 食事画像認識, 共起



(a) 入力画像 (b) 結果表示

図 1 本研究で構築した認識エンジンは入力画像から料理の候補推定し出力する．

1. はじめに

近年，携帯電話やスマートフォン等の情報端末を利用して食事記録をとるサービスが普及しつつある．食事情報を記録することで，食生活について意識したり，栄養分の評価を行うことができる．食事情報を記録する際の一般的な方法として，ユーザーがテキストを入力し，サービスに登録してある食べ物を検索する方法や，サービスに登録してある食べ物から階層的なリンクを用いて選択する方法が挙げられる．それらは，摂取した食品毎に登録をする必要があり，複数品目の料理を毎食記録するのは特に手間が大きい．そこで，より手軽に，より短時間で食事の記録をとる方法が望まれている．

本研究では，食事内容を少ない手間で記録するために，画像認識技術を用いて，画像中に含まれると推測される料理名の候補を表示する認識エンジンを構築した (**図** 1)．

2. 関連研究

食事画像の認識に関する関連研究として，FoodLog(注1)では，画像から得られる画像特徴を用いて，栄養を直接推定している．この方法は，どのような種類の料理でも認識対象にすることもできるが，認識結果が本当に正しいかどうかは，知識のないユーザーには理解しづらい．それに対して，本研究では，複数品目を含むような画像にも対応し，料理の種類を認識してユーザーの記録のサポートを行い，その後，栄養を計算するというアプローチを最終目標としている．

Yang ら [1] は，野菜やパンや肉などの材料の位置関係を特徴ベクトルとする事で，米国でよく食べられている 61 種類のファーストフードの分類に取り組み，28.2%の精度で分類する事ができた．また，Zong ら [2] も同様のファーストフードデータセット [3] に対して，SIFT 特徴点検出と Local Binary Pattern 記述子を用いた分類で，ベンチマークよりも良い分類精度を出した．我々の研究では，これらの米国の食生活に基づくデータセットとは異なり，日本でよく食べられている物を中心にデータセットを構築している．

我々は以前から食事画像認識について研究をしている [4]．そこでは，複数の料理が写っている画像や対象の料理が大きく写っていないものも考慮するために，ESS, 円検出, 領域分割を用いて料理領域の推定を行った．85 種類の料理を対象とした実験の結果，10 個の候補を表示するとき，一品が写ってる画像において 71.6%，複数

(注1)：http://www.foodlog.jp

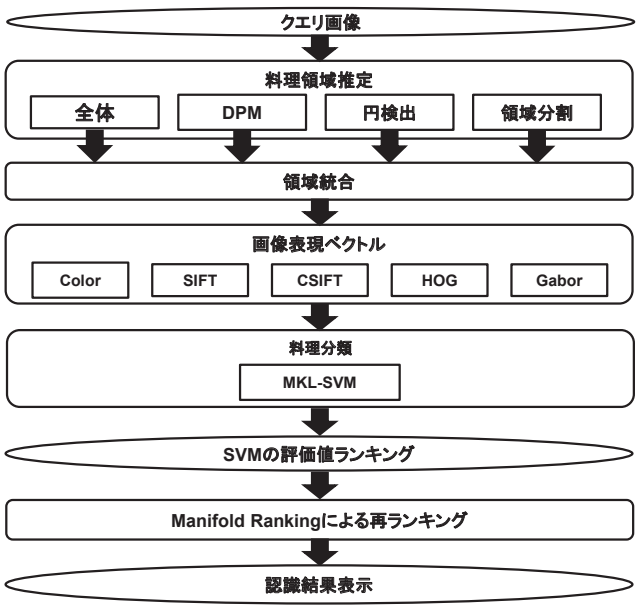


図 2　認識の流れ

品目が写っている画像において，60.2% の分類率を達成した．本研究では，料理領域推定の精度が良くなかった ESS に代わり，スライディングウィンドウ方式の物体検出手法である Deformable Part Model [5] を採用している．

以前の研究では各料理は独立して認識を行っていた．しかし，複数のオブジェクトを含む画像認識ではオブジェクト間の関係が重要な手がかりとして利用されている．先行研究として，データセットから得られた共起行列と Web 検索エンジンを利用して得られた共起行列を基に CRF を用いて認識結果の改善を行ったもの [6] や共起関係と空間関係を木構造を用いてモデル化し，オブジェクト検出の精度改善を行ったもの [7] がある．本研究では，データセットおよび検索エンジンから得られた料理間の共起確率を基に Manifold Ranking を用いて複数品目を含む画像において認識精度の改善を行う．

3. 認 識 手 法

本研究で作成した料理画像認識エンジンの処理の概要を**図** 2 に示す．認識対象の画像が与えられたら，従来手法に基づいて，画像中の料理領域を推定し，各候補領域に対して SVM による分類を行う．次に料理の共起確率を利用し Manifold Ranking で SVM の評価値の再ランキングを行う．最終的な分類結果は，再ランキング後のスコア上位 N 個を表示する．

3.1 従 来 手 法

認識対象の画像が与えられてから SVM による評価値を得るまでの手法についての詳細を記述する．

(a) 入力画像　(b) 輪郭抽出　(c) 円検出

図 3　円検出の流れ

3.1.1 候補領域検出

複数の料理が写っている画像では，特徴ベクトルを作成する前処理として，候補領域の推定を行うことで，認識精度が向上する [4]．本研究では，画像全体に加えて，Deformable Part Model，円検出，領域分割を用いて料理の位置推定を行っている．

a） Deformable Part Model

Felzenszwalb らは，全体を表す global root filter と複数のパーツモデルの 2 層でオブジェクトを定義する Deformable Part Model(DPM) [5] を提案した．モデルの学習には，各パーツの相対的な位置関係も含まれており，検出の際には，パーツの HOG 特徴量によるスコアと位置関係の変形によるスコアで評価される．

本研究では，100 種類の各料理モデルを作成し，それぞれについて検出を行っている．モデルの学習および検出には [8] を使用した。

b） 円　検　出

画像から円形の輪郭を抽出する事で，皿の領域を検出し，それを候補領域とする．

まず，入力画像をグレースケール画像に変換し，Canny Edge Detector により，輪郭を抽出する．抽出された輪郭に対して，Hough 変換による円検出を行うことで，画像から円形の輪郭を抽出する (**図** 3)．

なお，予備実験では楕円検出も試みたが，楕円検出では楕円領域が多く抽出され過ぎる場合がしばしばあったため，今回は円検出のみを用いることとした．

c） 領 域 分 割

領域分割とは，似た色を持つ領域に画像を分割する事である．本研究では，領域分割アルゴリズムとして JSEG [9] (注2) を用いた．JSEG では，色空間の量子化を行い，カラークラスマップを作成することで，空間分割を行う．JSEG では，パラメータとして分割後の領域数を設定することができる．本研究では，画像をおよそ 10 個の領域に分割し，候補領域とした．

また，領域分割によって得られた領域の 2 つを結合した時の円形度が，結合された 2 つの領域より大きくなる場合，結合した領域も料理の候補領域とする (**図** 4)．円

(注2)：http://vision.ece.ucsb.edu/segmentation/jseg/software/

(a) 入力画像　(b) 領域分割の結果　(c) 領域の結合

図 4　領域分割での候補領域検出

形度とは，領域がどの程度円に近いかを示す指標である．円形度は領域の面積を S，領域の周囲長を L とした場合，$(4\pi S)/L^2$ で求められ，この値は最大 1 となり，大きいほど円形に近い．

d）候補領域の選定

それぞれの手法で検出した候補領域は以後の処理で同等に扱うために，検出領域を含む bounding box とした．このとき，各手法で検出した候補領域に対して，領域の形を調べ，明らかに間違っている候補領域を除去する事で，分類にかかる計算コストを削減し，かつ，ノイズとなる評価値も削減する．本研究では，検出された候補領域の短辺が 60 ピクセル以下の物は小さすぎる領域として候補領域から除外する．さらに，学習画像から各種類の料理の縦横比の平均と標準偏差を計算しておき，縦横比の値が平均値を中心として標準偏差の ±2 倍以内の範囲から外れている，縦横比が極端なものを候補領域から除外する．

3.1.2　候補領域に対する種類分類

a）特　徴　量

本研究では，候補領域を分類するために，色特徴，SIFT, CSIFT, HOG, Gabor の 5 つの画像特徴を用いた．

色特徴，SIFT, CSIFT は 10pixel ごとの Dense Sampling で抽出を行い，3×3 の各領域で 1000 次元の Bag of Features 表現にして，合計 9000 次元の特徴ベクトルとした．HOG は，領域を 8×8 セルに分割し，1 ブロック 3×3 で合計 36 ブロックとし，全体で 2916 次元の特徴ベクトルとした．Gabor は，領域を 8×8 セルに分割し，それぞれで 6 方向，4 周期のガボール変換カーネルによって特徴を抽出することで，1536 次元の特徴ベクトルとした．

b）分　類　器

本研究では，分類器として Support Vector Machine(SVM) を用いて，各料理クラスに対する評価値を求める．その際，複数の画像特徴を利用しているため，Multiple Kernel Learning(MKL) により複数のカーネルを線形結合して，最終的な統合カーネルを得る．ここでの料理ごとの評価値は，多数の候補領域から得られた評価値のうち最も高いものをその料理の評価値としてみなしている．

3.2　共起関係を用いた再ランキング

一般に食事には組み合わせが存在し，“ごはん” と” みそ汁” や” ハンバーガー” と” フライドポテト” は一緒に食べられることが多いが，“寿司” と” ハンバーガー” などは一緒に食べられることはほとんどないと考えられる．

独立に求められた各料理の評価値に対して，このような料理間の共起の統計情報を用いることで，複数の料理が存在する場合の認識精度の向上をはかる．

3.2.1　Manifold Ranking

本研究では，共起情報を反映させる手法として，Manifold Ranking [10] を用いる．Manifold Ranking とは，Google の Page Rank を一般化したもので，類似するデータが多いものほど上位に来るようにスコアを再ランキングする手法である．画像検索においては，He ら [11] が，検索結果に対する適合性フィードバックに Manifold Ranking を用いた手法を提案している．

最初の q 個をクエリとするデータ $X = \{x_1, \cdots, x_q, x_{q+1}, \cdots, x_n\}$ が与えられ，X についてのランキングスコア $f = \{f_1, \cdots, f_n\}$，x_i がクエリであるときに $y_i = 1$，それ以外では $y_i = 0$ とするベクトル $y = \{y_1, \cdots, y_n\}^T$ を定義する．このとき，Manifold Ranking のアルゴリズムは，以下のようになる．

Manifold Ranking のアルゴリズム

（1）　$x_i.x_j$ の距離からなる類似行列 W を作成する．

（2）　D_{ii} が W の i 行目の総和となる対角行列を用いて，$S = D^{-1/2}WD^{-1/2}$ のように対称正規化を行う．

（3）　$f(t+1) = \alpha S f(t) + (1-\alpha)y$ の計算を収束するまで行う．ここで，α は $[0,1)$ をとるパラメーターである．

（4）　収束した結果 f^* の値に基づいてランキングする．

また，Zhou らは [10] で，反復の結果が $f^* = (1-\alpha)(I - \alpha S)^{-1}y)$ に収束することを示した．

本研究では，画像の類似度行列の代わりに料理の共起行列を用いて，分類器による評価値の再ランキングを行う．

つまり，提案手法における最終的な評価値は次のようにして得られる．

$$r^* = (I - \alpha S)^{-1} r \tag{1}$$

ここで，r^* は再ランキング後の評価値，r は初期の評価値，S は共起確率行列である．

SVM の出力値 v_i に基づく初期の評価値 r は，次のようにして得られる．

$$\boldsymbol{r}_i = \frac{(1+\exp(-\boldsymbol{v}_i))^{-1}}{\sum_j (1+\exp(-\boldsymbol{v}_j))^{-1}}. \tag{2}$$

3.2.2 共起確率

本研究では，共起確率をデータベースから得る方法と Normalized Google Distance を利用し Web から得る方法の 2 通りで求めた．

a） データベース

ひとつは，これまで収集してきた料理画像データベース中の画像から共起頻度を求める方法である．共起確率行列は**式** (3) のようにして求める．

$$S_{i,j} = \frac{c_{i,j}}{\sum_{k \in F \wedge k \neq j} c_{k,j}}, \tag{3}$$

ここで，$c_{i,j}$ は料理画像データベース中で料理 i と料理 j が同時に出現した回数を表しており，F は認識対象の料理カテゴリー全体の集合を表す．

b） Normalized Google Distance

もうひとつは，Web から料理同士の共起確率を得る方法である．これには Normalized Google Distance(NGD) [12] という検索エンジンのヒット件数を用いた単語間の類似性尺度を利用する．

NGD は，x, y を単語とし，$f(x)$ を単語 x のヒット件数，$f(x,y)$ を単語 x と y のアンド検索時のヒット件数，M を検索エンジンの保持する総 Web ページ数とすると，**式** (4) で求められる．

$$NGD(x,y) = \frac{\max(\log f(x), \log f(y)) - \log f(x,y)}{\log M - \min(\log f(x), \log f(y))} \tag{4}$$

ただし，NGD は類似性の尺度であるため類似する単語同士の場合に値が小さくなるため，**式** (5) のようにして共起行列に変換する．

$$S'_{i,j} = \frac{\exp(-NGD(food_i, food_j)}{\sum_{k \in F \wedge k \neq j} \exp(-NGD(food_k, food_j))} \tag{5}$$

4. 実　　験

4.1 データセット

実験には我々が構築している食事画像データセットを用いる．このデータセットには，100 種類の料理がそれぞれ 100 枚以上含まれており，複数品目の料理が写っている画像は 900 枚含まれている．今回の実験で認識対象としている 100 種類の料理を**図** 5 に示す．その中から学習用に単品画像を 8500 枚，複数品目画像を 400 枚，評価用に複数品目画像を 500 枚用いた．評価用の 500 枚の画像には合計 1178 品の料理が含まれている．

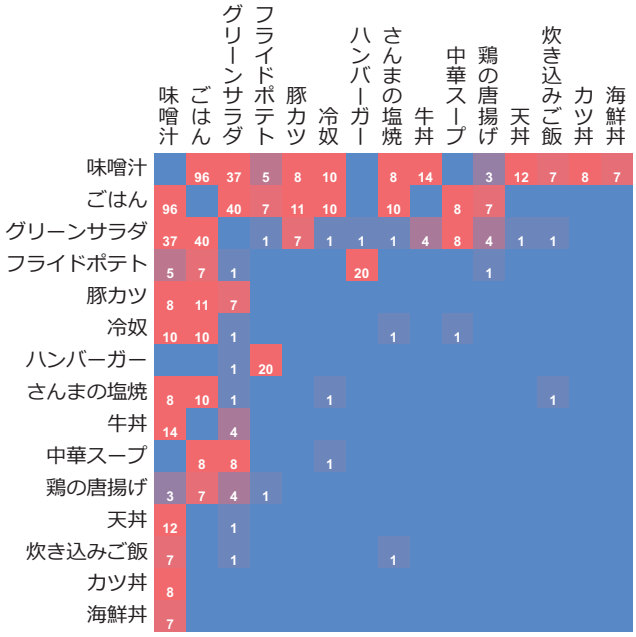


図 6　データベースから得られた共起頻度行列 (一部)

学習用の複数品目を含む画像 400 枚から得られた共起頻度の行列を**図** 6 に示す．この図は，赤色が濃いほど共起する頻度が高いことを示している．大部分は青色になっており，共起しない組み合わせであるが，ごはんや丼ものなどとみそ汁やサラダの組み合わせがあるほか，ハンバーガーとフライドポテトなどの組み合わせも存在する．

4.2 評価方法

分類結果の評価に用いる基準として，以下で定義する分類率を用いる．

$$\text{分類率} = \frac{\text{第 N 候補までに挙げられた正しい料理の数}}{\text{認識対象の料理数}}$$

ただし，画像中に対象としている 100 種類以外の料理含まれている場合，それは料理数に数えないものとする．

4.3 実験結果

本実験では，共起を利用しない場合をベースラインとして，データベースから得られた共起確率を利用する場合，Web から得られた共起確率を利用する場合との比較を行った．

提案手法では，Manifold Ranking を用いる際に，パラメーター α を指定する必要がある．今回の実験では，0 から 0.3 まで 0.05 刻みで評価を行い，最も分類率が高くなった $\alpha = 1.0$ を採用した．α を変化させたときの分類

図 5　100 種類の料理のサンプル

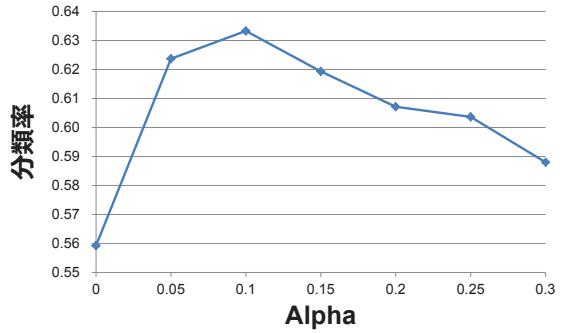


図 7　データベースから求めた共起確率を利用した場合の α

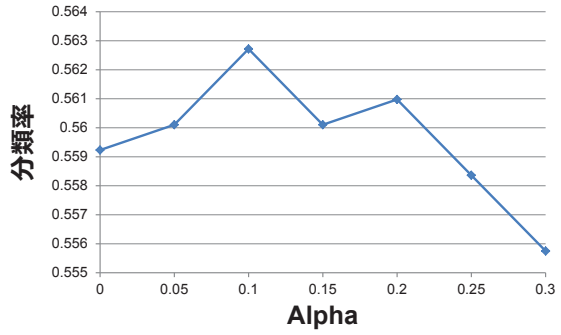


図 8　Web から求めた共起確率を利用した場合の α

率の変動は**図** 7 および**図** 8 に示す.

データベースから得られた共起確率を利用した場合と Web から共起確率を推定した場合それぞれについて，出力候補料理数を変化させたときの分類率の推移を**図** 9, **図** 10 に示す.

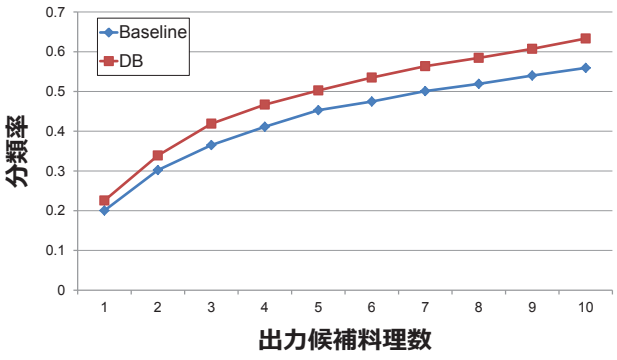


図 9　DB から求めた共起確率を利用した場合

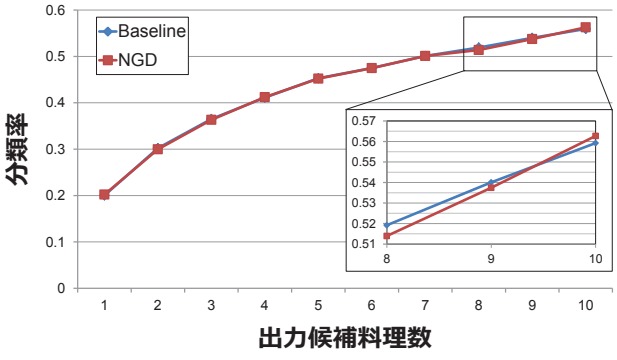


図 10　Web から求めた共起確率を利用した場合

データベースから得られた共起確率を利用した場合には，候補料理を 10 個まで出力したとき，共起を利用しない場合と比べ 7.4 ポイント向上し 63.33%の分類率を達成した．また，Web から得られた共起確率を利用した場合

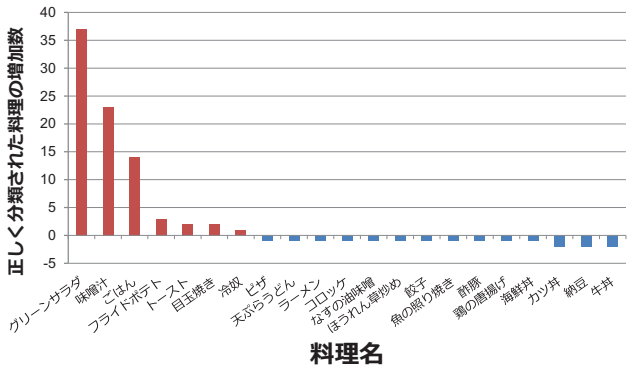


図 11　10 個の料理を出力した際の料理別の正しく認識された数の変化

には，同様に候補料理を 10 個まで出力したとき，共起を利用しない場合と比べ 0.35 ポイント向上し 56.27%の分類率を達成した．

データベースから得られた共起確率を利用することで，共起を利用しない場合よりも良い認識精度が得られ，提案手法の有効性が示された．しかし，Web から得られた共起確率を利用した場合には，ほとんど改善が無かった．これは，Web 上に存在するテキストが同時に食べられる料理という意味での共起関係を表していないことが要因であると考えられる．

4.4　料理ごとの認識結果

データベース中の共起頻度を用いたときの結果について，10 個の候補を出力した場合に正しく認識された数に変化があったものを料理別に**図** 11 にまとめた．今回の実験では，多くの料理と共起する頻度の高かった，グリーンサラダ，味噌汁，ごはんの 3 品目に関して特に認識精度の向上があった．一方で，画像中に前述の 3 品目が含まれない場合についても同様に上位に再ランキングされ，共起頻度の高くない他の料理の順位が下がってしまい，それにより 10 位以下となってしまうケースもいくつかみられた．しかし，“ごはん”，“みそ汁”，“グリーンサラダ” は，画像中に表れる回数も多いため，全体としては分類率が向上するという結果になっている．

5.　ま　と　め

本研究では，Manifold Ranking を用いて，料理間の共起関係を反映させることで，複数品目画像における認識精度の向上をはかった．

10 個の料理を表示するとき，データベースからの共起確率を用いた場合に，従来手法と比べ 7.4 ポイント向上し，63.33%の分類率，Web からの共起確率を用いた場合に，従来手法と比べ 0.35 ポイント向上し，56.27%の分類率を達成し，共起関係を用いることが有効であることを示した．

今後は，データセットの拡充を行いより一般的な料理間の共起を反映させることや，検索エンジンを用いた手法はあまり有効では無かったため，他の共起を表すソースを見つけることが課題となる．

また，提案手法は複数品目画像に対しては有効であることが示されたが，単品画像に対しては，データベースから得られた共起確率を用いて Manifold Ranking を適応したところ，分類率が 68.14%と 3.09 ポイント下がる結果となった．そのため，単品が写っている画像であるか，複数品が写っている画像であるかを識別する必要性もある．

文　　献

[1] S. Yang, M. Chen, D. Pomerleau and R. Sukthankar: “Food recognition using statistics of pairwise local features”, Proc. of IEEE Computer Vision and Pattern Recognition (2010).

[2] Z. Zong, D. Nguyen, P. Ogunbona and W. Li: “On the combination of local texture and global structure for food classification”, 2010 IEEE International Symposium on MultimediaIEEE, pp. 204–211 (2010).

[3] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar and J. Yang: “Pfid: Pittsburgh fast-food image dataset”, IEEE International Conference on Image ProcessingIEEE, pp. 289–292 (2009).

[4] 甫足, 松田, 柳井：“候補領域推定による複数品目に対応した食事画像認識”, 画像の認識・理解シンポジウム (MIRU), pp. 1–7 (2011).

[5] P. Felzenszwalb, R. Girshick, D. McAllester and D. Ramanan: “Object detection with discriminatively trained part-based models”, IEEE Transactions on Pattern Analysis and Machine Intelligence, **32**, 9, pp. 1627–1645 (2010).

[6] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora and S. Belongie: “Objects in context”, Proc. of IEEE International Conference on Computer Vision (2007).

[7] M. J. Choi, J. J. Lim, A. Torralba and A. S. Willsky: “Exploiting hierarchical context on a large database of object categories”, Proc. of IEEE Computer Vision and Pattern Recognition (2010).

[8] P. F. Felzenszwalb, R. B. Girshick and D. McAllester: “Discriminatively trained deformable part models, release 4”.

[9] Y. Deng and B. S. Manjunath: “Unsupervised segmentation of color-texture regions in images and video”, IEEE Transactions on Pattern Analysis and Machine Intelligence, **23**, 8, pp. 800–810 (2001).

[10] D. Zhou, J. Weston, A. Gretton, O. Bousquet and B. Schölkopf: “Ranking on data manifolds”, Advances in Neural Information Processing Systems, **16**, pp. 169–176 (2004).

[11] J. He, M. Li, H. Zhang, H. Tong and C. Zhang: “Manifold-ranking based image retrieval”, Proc. of ACM International Conference Multimedia, pp. 9–16 (2004).

[12] R. Cilibrasi and P. vitanyi: “The google similarity distance”, Knowledge and Data Engineering, IEEE Transactions on, **19**, 3, pp. 370–383 (2007).