社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Deformable Part Modelを用いた料理の位置検出

松田裕司†　　柳井啓司†

† 電気通信大学 情報理工学研究科 総合情報学専攻
E-mail: †matsuda-y@mm.cs.uec.ac.jp, ††yanai@cs.uec.ac.jp

あらまし　本研究では以前の研究で構築した画像中に含まれると推測される料理の候補を表示する認識エンジンの改良を行う．以前の研究では，高速スライディングウィンドウ探索や領域分割，円検出を用いて，画像中の料理の候補領域を推定し，その部分の分類を行うことで，画像中に複数の料理がある場合に対応した．本研究では，候補領域の推定に Deformable Part Model を用いることで，領域推定の精度改善を行う．実験で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，従来手法と比べ 3.6 ポイント向上し，56.0%の分類率を達成し，料理の候補領域推定に Deformable Part Model を用いることが有効であることが示された．さらに，料理同士の共起確率を用いることで，同様に 10 個の候補を表示したときに，8.4 ポイント向上し，64.4%の分類率を達成し，共起確率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，Multiple Kernel Learning

Food region detection using Deformable Part Model

Yuji MATSUDA† and Keiji YANAI†

† Department of Informatics, The University of Electro-Communications, Tokyo, Japan
E-mail: †matsuda-y@mm.cs.uec.ac.jp, ††yanai@cs.uec.ac.jp

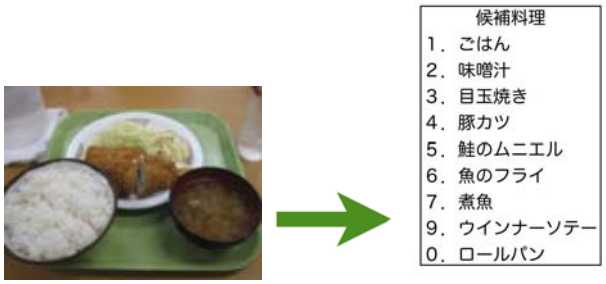
Abstract　In this paper, we improve our food recognition system with Deformable Part Model.

Key words　Food image recognition, Multiple Kernel Learning

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).



(a) 入力画像　　(b) 結果表示

図 1　本研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．

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

(注1)：http://www.foodlog.jp

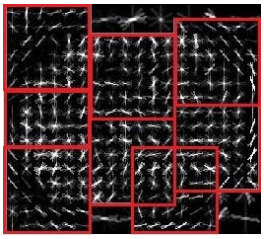


図 2　DPM のモデル例

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], [5]. [4] では，画像中に一品の料理が大きく写ってるものを対象とし，9 種類の視覚的特徴を Multiple Kernel Learning(MKL) を用いた方法で効率良く特徴を統合することで，50 種類の料理について 61.34%の割合で正しく分類する事が出来た．また, [5] 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うことで，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 69.4%，複数品目が写っている画像において，52.4%の分類率を達成した．本研究では，Deformable Part Model を用いて，料理領域の推定の精度向上に取り組む．

3.　料理候補領域検出

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 [5]．ここでは，本研究で用いる候補領域検出手法をまとめる．

3.1　Deformable Part Model

Felzenszwalb らは，物体のモデルをパーツの集合として表現し，それぞれのパーツの妥当性およびそれらの相対位置関係で評価を行う Deformable Part Model(DPM) [6] を提案した．

DPM で学習した料理のモデルの例を図 2 に示す．本実験では，DPM のモデルの学習および検出に [7] を使用した。

3.2　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．

(a) 入力画像　(b) 輪郭抽出　(c) 円検出

図 3　円検出の流れ

3.3　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 [8] (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．

(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

3.4　候補領域の選定

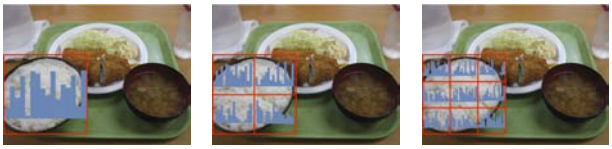
それぞれの手法で検出した候補領域は以後の処理で同等に扱うために，検出領域を含む bounding box とした．このとき，各手法で検出した候補領域に対して，領域の形を調べ，明らかに間違っている候補領域を除去する事で，分類にかかる計算コストを削減し，かつ，ノイズとなる評価値も削減する．本研究では，検出された候補領域の短辺が 60 ピクセル以下の物は小さすぎる領域として候補領域から除外する．さらに，学習画像から各種類の料理の縦横比の平均と標準偏差を計算しておき，縦横比の値が平均値を中心として標準偏差の ±2 倍以内の範囲から外れている，縦横比が極端なものを候補領域から除外する．

4.　候補領域に対する種類分類

4.1　画像特徴ベクトルの生成

料理画像を視覚的特徴から認識するためには，色特徴や SIFT 特徴を単純に利用するだけでは良い結果は得られない [1], [4]. そこで，本研究では以前の研究 [5] と同じく，複数の視覚的特徴を効果的に統合して利用する．本節では本研究で利用した画像特徴について記述する．

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



(a) level 1　(b) level 2　(c) level 3

図 5　各ピラミッドレベルで Bag-of-features での特徴ベクトルの作成

4.1.1　Dense sampling による局所特徴

局所特徴とは，画像中の小領域の特徴ベクトルである．本研究では，dense sampling によって，10 ピクセル毎に，半径 8 ピクセルと 16 ピクセルの二つのサイズに対して局所領域を選択した．各局所領域に対して，カラーヒストグラム，局所画像パターンである SIFT，照明変化に頑強な色空間に対して SIFT を得る CSIFT [9] の 3 種類の特徴を抽出した．

4.1.2　空間ピラミッド表現

抽出された局所特徴は空間ピラミッド表現 [10] を用いて，各領域の画像特徴ベクトルとした．空間ピラミッド表現では，認識対象とする領域を階層的にグリッドで分割し，それぞれのグリッドに対して，Bag-of-features 表現を用いて，局所特徴の出現頻度の特徴ベクトルを作成する事で，局所特徴の空間情報も考慮した特徴ベクトルを得ることができる．ピラミッドレベル l では，画像を $l \times l$ のグリッドに分割する (**図** 5)．本研究では，Bag-of-features 表現では，各局所特徴は 1000 種類の代表パターンのいちばん近い物に割り当て，ピラミッドレベル 3 を使用した．ピラミッドレベル 3 では 9000 次元の画像特徴ベクトルが得られる．

4.1.3　Histograms of Oriented Gradients

一般物体認識の為の gradient-base の視覚的特徴として，Dalal ら [11] は Histograms of Oriented Gradients（HOG）を提案した．HOG は SIFT と同様に，輝度の勾配方向をヒストグラム化した特徴量である．SIFT は特徴点の周りに対して特徴量を記述するのに対し，HOG は一定領域に対する特徴量の記述を行う．そのため，物体の大まかな形状を表現することが可能である．

本研究では与えられた領域を 8×8 セルに分割し，1 ブロックを 3×3 とし，$6 \times 6 = 36$ ブロックを取る．よって，与えられた領域全体で 2916 次元のヒストグラムを得た．

4.1.4　ガボール特徴

ガボール特徴は，画像から局所的な濃淡情報の周期と方向を表した特徴量である．カーネルの形を固定し，それを周期を変えて伸び縮みさせたり，回転させて方向を変えたりして，様々な周期や方向のカーネルフィルタカーネルを作成する．周期的濃淡変化を抽出する．解像度 m，方向 n のガボールフィルタは次式で表される．

$$g_{m,n}(x,y) = \frac{k_m^2}{\sigma^2}\ \exp\left\{-\frac{k_m^2(x^2+y^2)}{2\sigma^2}\right\} \times\left[\exp\{jk_m(x\cos\theta_n + y\sin\theta_n)\} - \exp\left(-\frac{\sigma^2}{2}\right)\right] \tag{1}$$

ここで，式 1 の k_m および θ_n は，以下のように表される．

$$\begin{aligned} k_m &= a^m \quad (0 \leqq m \leqq S-1) \\ \theta_n &= \tfrac{n\pi}{K} \quad (0 \leqq n \leqq K-1) \end{aligned} \tag{2}$$

K は方向の数，S は解像度の数，a は拡大率を表す．式 1 で表されるフィルタを用いて，それぞれに対応した空間周期の特徴を抽出 (パターン強度を数値化) する．ガボールフィルタは，特定の向きのエッジと特定の幅のエッジを抽出する．最後に，各フィルタ毎に強度の平均を求め，それをヒストグラムとする．本研究では与えられた領域を 8×8 セルに分割し，それぞれ 6 方向，4 周期のガボール変換カーネルについて特徴を抽出することで 1536 次元の特徴ベクトルを得た．

4.2　分　類　器

対象の領域から特徴ベクトルの抽出をした後，事前に学習された特徴ベクトルと比較して，どの種類の料理クラスに属するかを決定する．

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計算する．

a）　カーネル関数

線形識別器は 2 クラスが線形分離可能であるときには高い認識率を期待できるが，非線形で複雑な問題に対してはその限りではない．そこで，非線形な写像 Φ で写像される先での内積 $(\Phi(x) \cdot \Phi(x^{'}))$ は，元の空間で定義されるカーネル関数 $K(\mathbf{x_1}, \mathbf{x_2})$ の値と一致するものとする．本実験で用いるカーネル関数として，線形識別関数の線形カーネルと χ^2 距離に基づく χ^2RBF カーネルについての定義は以下のようになる．

$$K(\mathbf{x},\mathbf{y}) = \exp\Big(-\gamma \sum_i \frac{\|x_i - y_i\|^2}{x_i + y_i}\Big)$$

ただし，$\gamma > 0$ となる実数であり，パラメータとして与える必要がある．

Zhang らは [12] は，χ^2RBF カーネルのパラメータ γ に，全ての学習画像のベクトルの組み合わせの χ^2 距離 $\sum_i \frac{||x_i - y_i||^2}{x_i + y_i}$ の平均の逆数を設定することによって，良い結果を報告している．本研究においても，γ のパラメータは同様の方法で設定する．

b）　Multiple Kernel Learning による特徴統合

本研究では，より高精度に料理を認識するために，複数の画像特徴量のカーネルを線形結合することにより統合カーネルを作成し，それを SVM に適用して特徴統合による画像認識を実現する．

最適なカーネル (カーネルを重みつきで線形結合したカーネル) のサブカーネルに対する重み β_j を学習する．統合カーネルは以下の式のように表される．

$$K_{\text{combined}}(\mathbf{x},\mathbf{x}') = \sum_{j=1}^{K} \beta_j k_j(\mathbf{x},\mathbf{x}') \quad \text{with}\ \ \beta_j \geqq 0,\ \ \sum_{j=1}^{K} \beta_j = 1. \tag{3}$$

各サブカーネルをそれぞれの特徴と対応させることによって，

MKL は特徴選択や特徴統合に用いることができる．

本研究では，以前の研究 [4], [5] と同様 [13] に MKL の学習を行う．

4.2.1 候補の出力

入力画像に対する最終的な認識結果は，全ての候補領域から得られた出力値のうち各料理ごとに最大となるものをそれぞれの料理の評価値と見なし、評価値の上位 N 個の料理を候補として出力する．

4.3 共起関係の利用

複数品目が写っている画像において，同時に現れる料理同士は無関係であることは少なく，そこには関連があるものと考えられる．そこで，独立に求められた各料理の評価値を共起確率を用いることでより上位に正しい結果が現れるように修正を行う．

He ら [14] は，画像検索において manifold ranking [15] を用いることで検索結果の改善を行った．本研究では，画像の類似度に基づく遷移行列の代わりに料理の共起確率行列を用いて結果の改善を行う．manifold ranking によるランキングスコアの計算は式 4 で表される．

$$f^* = (I - \alpha S)^{-1} y \tag{4}$$

ここで，$f*$ は再ランキング後の評価値，y は元の評価値，S は共起確率行列，α は [0, 1) の値を取るパラメータである．

5. 実　　験

本研究では，実験のために独自に食事画像データセットを構築した．データセットには,「食事バランスガイド」を基にして，我々が普段食べるであろう物を中心にした 100 種類の料理が，それぞれ 100 枚以上含まれている．

本実験では，このデータセットから，8632 枚（8854 オブジェクトを含む）の画像を学習に利用し，複数品目が写った画像 500 枚（1178 オブジェクトを含む）を対象に分類を行った．

5.1 評 価 方 法

分類結果の評価に用いる基準として，分類率を用いた．本実験で使用する分類率を以下の式で定義する．

$$\text{分類率} = \frac{\text{第 N 候補までに挙げられた正しい料理の数}}{\text{分類されるべき全ての料理の数}}$$

5.2 比 較 手 法

本研究で構築した認識エンジンの精度を評価するためにいくつかの手法との比較を行った．以前の研究 [5] では，候補領域推定に画像全体，ESS，円検出，領域分割を用いている．さらに単純な手法との比較として，本研究で行った各候補領域の検出を単独で用いた物と精度を比較する．本研究では，各候補領域検出方法に対して検出されたすべての領域を使用する．また，正しく候補領域が与えられた時の精度についても調べた．

5.3 実 験 結 果

複数品目が写っている画像について出力候補料理数を変えたときの分類率を図 7 に示した．この図では, [5] の手法、それぞれ各候補領域検出手法を単体で利用した場合，その 3 つの候補領域検出手法を全て利用する提案手法，正解データの bounding box を利用して領域が既知だった場合の分類率を示した．

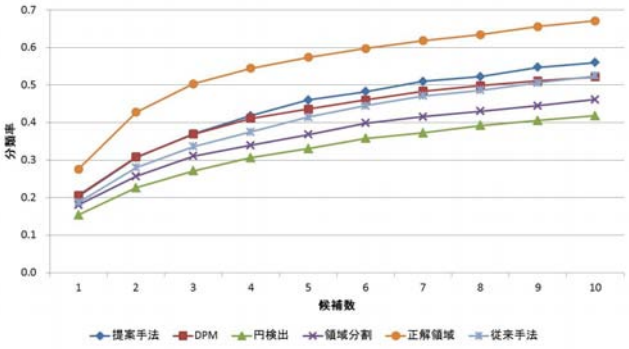


図 7　複数品を含む画像での第 N 候補までの分類率

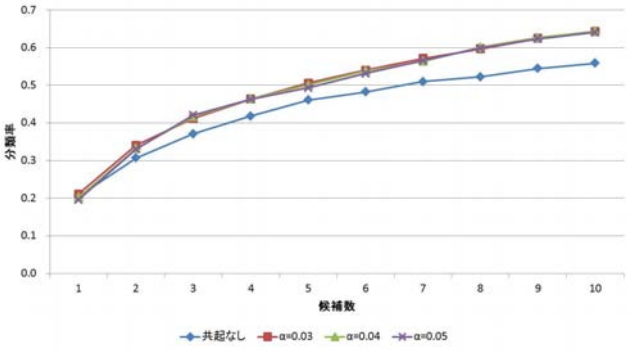


図 8　共起関係を利用した場合での第 N 候補までの分類率

第 10 候補までの料理名を出力した場合，複数品を含む画像での分類率においては，提案手法は従来手法から 3.6 ポイント向上し，56.0%の分類率を達成した．

なお，単品画像に対しても適応した結果，従来手法が 69.5%に対して提案手法は 68.9%と分類率が僅かながら低下する結果となった．本研究では主に複数品目を対象としているため，単品画像が含まれるデータセットに対する精度向上は今後の課題とする．

5.4 共起関係の利用

本研究ではさらに複数品目が写っている料理画像を対象として，料理間の共起確率を利用することで精度向上を行った．本実験ではデータセット中の複数品目画像が少ないため，共起確率の集計には学習用セット，テスト用セットの区別無く，複数品目が写っている画像全てを対象とした．

共起確率を使用しない場合と共起確率を使用した場合の分類率を図 8 に示した．式 4 の α は，0.03 から 0.05 まで変化させ比較している．

先ほどと同様に第 10 候補までの料理名を出力した場合，共起関係を用いた場合では，$\alpha = 0.04$ のときに，共起関係を用いない場合から 8.4 ポイント向上し，64.4%の分類率を達成した．

6. まとめと今後の課題

本研究では，食事画像中の料理領域の推定に Deformable Part Model を用いて食事画像認識エンジンの精度向上をはかっ

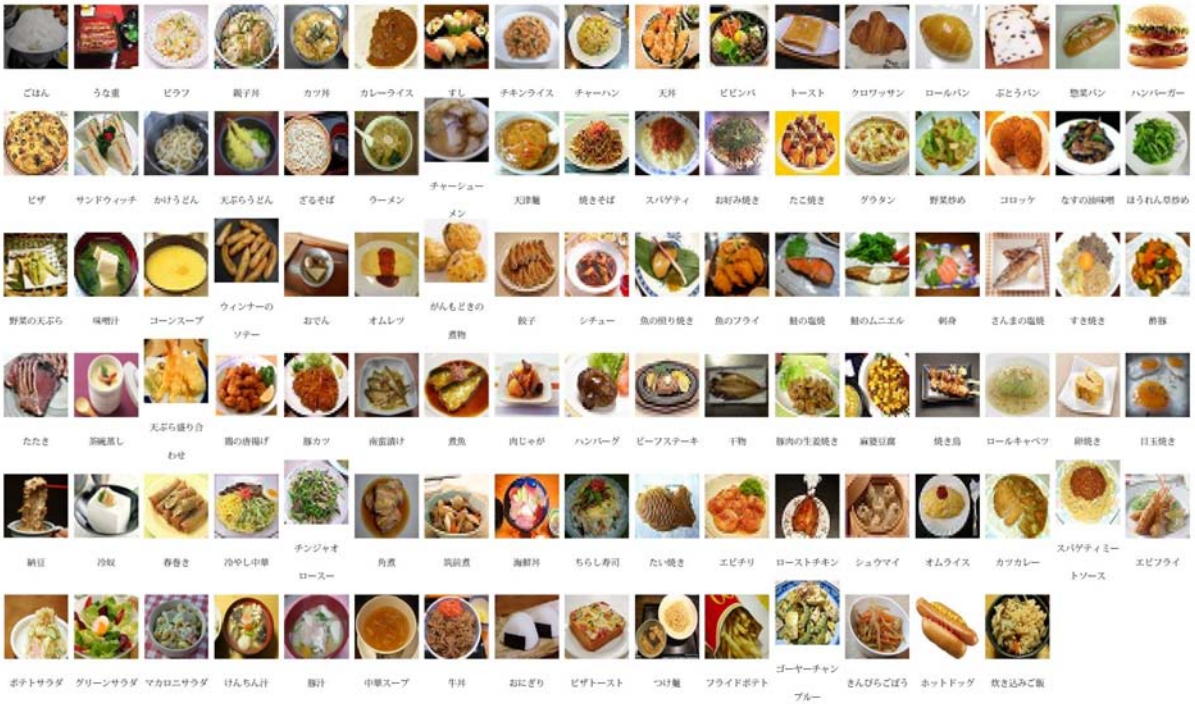


図 6　100 種類の料理のサンプル

た．10 個の料理候補を表示するとき，複数品目を含む画像において，従来手法と比べ 3.6 ポイント向上し，56.0%の分類率を達成し，提案手法が有効であることを示した．

また，料理同士の共起確率を用いることで，同様に 10 個の料理候補を表示するとき，8.4 ポイント向上し，64.4%の分類率を達成し，共起確率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料理同士の共起関係をより正確に反映させることや，CRF を用いるなど異なる手法との比較を行う必要がある．また，今回は単品画像は考慮していないため，実際の料理画像に対して共起関係を用いるためには，単品が写っている画像であるか複数品目が写っている画像であるかということを識別する必要性もある．

さらに，現在の認識エンジンは非常に多くの処理を行うため，GPGPU や効率の良いアルゴリズムを使用することで，処理の高速化を目指すことも重要な課題となっている．

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, pp.2249–2256, 2010.

[2] Z. Zong, D.T. 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,” Image Processing (ICIP), 2009 16th IEEE International Conference onIEEE, pp.289–292 2009.

[4] 上東太一，甫足創，柳井啓司，“Multiple kernel learning による 50 種類の食事画像の認識，” 電子情報通信学会論文誌 D, vol.J93-D，no.8，pp.1397–1406，2010.

[5] 甫足　創，松田裕司，柳井啓司，“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.

[6] 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.

[7] P.F. Felzenszwalb, R.B. Girshick, and D. McAllester, “Discriminatively trained deformable part models, release 4”. http://www.cs.brown.edu/ pff/latent-release4/

[8] Y. Deng and B.S. Manjunath, “Unsupervised segmentation of color-texture regions in images and video,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.23, no.8, pp.800–810, 2001.

[9] A. Abdel-Hakim and A.A. Farag, “Csift: A sift descriptor with color invariant characteristics,” Proc. of IEEE Computer Vision and Pattern Recognition, vol.2IEEE, pp.1978–1983 2006.

[10] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” Proc. of IEEE Computer Vision and Pattern Recognition, pp.2169–2178, 2006.

[11] N. Dalal, B. Triggs, I. Rhone-Alps, and F. Montbonnot, “Histograms of oriented gradients for human detection,” Proc. of IEEE Computer Vision and Pattern Recognition, pp.886–893, 2005.

[12] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, “Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study,” International Journal of Computer Vision, vol.73, no.2, pp.213–238, 2007.

[13] M. Varma and D. Ray, “Learning the discriminative power-invariance trade-off,” Proc. of IEEE International Conference on Computer Vision, pp.1150–1157, 2007.

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

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