# 位置情報付き写真における撮影位置の航空写真を利用した画像認識

八重樫　恵太[†1]　柳井啓司[†2]

我々は，位置情報付き写真の一般画像認識において写真の撮影位置に対応する航空写真を付加的な画像特徴量として利用する研究を行っている．これまでの研究から，写真画像から抽出した特徴量に，対応する位置の航空写真から抽出した特徴量を組み合わせて認識を行うことによって，写真からの特徴量のみ用いた認識よりも認識精度が向上することが判明している．しかしながら，従来は写真から抽出した特徴と航空写真から集出した特徴を単純に結合する認識手法を用いていたため，精度向上に航空写真の情報がどの程度寄与していたか定量的な評価が困難であった．そこで，本研究では，マルチカーネル学習 (Multiple Kernel Learning) を用いて，学習時における写真と航空写真の最適な重みの比を推定することにより，様々な認識カテゴリにおいて，画像認識における位置情報の有効性の度合いを評価する．本実験により，航空写真を利用した分類が極めて有効であるカテゴリとそうでないカテゴリが存在することが示された．

# Image Recognition for Geotagged Photos Using Corresponding Aerial Photos

Keita Yaegashi[†1] and Keiji yanai[†2]

We have been conducting study of exploiting aerial photos as additional image features for visual recognition of geotagged photos. In our previous work, since we simply concatenate feature vectors extracted from geotagged photos and ones extracted from aerial photos, it was difficult to evaluate how much geotags helped improve the recongnition performance. In this paper, we introduce multiple kernel learning (MKL) to evaluate contribution of both features for recognition by estimating the weights of geotagged photo and aerial photo. The experimental results demonstrate effectiveness of usage of aerial photos for recognition of consumer photos.

## 1. はじめに

今日では，デジタルカメラの普及により，撮影位置の情報を持つ位置情報付き写真は WWW(World Wide Web) 上に大量に存在している．大量の写真情報の普及にもかかわらず，それらを自動で整理・分類し，ユーザーの手間を省くことは未だ困難な課題であり続けている．単純な画像分類へのソリューションとして，現状ではタグ (内容を表現する複数の単語の集合) やタイトルなど，テキストベースのメタデータが主流になっている．写真を分類するための一般画像認識の基礎技術が高度化する中で，認識精度の向上を図るにあたり，写真と関連する多様な情報をいかに効率的に組み合わせるかが求められる．

我々は，位置情報付き写真の一般画像認識において写真の撮影位置に対応する航空写真を付加的な画像特徴量として利用する研究を行ってきた[6),7)]．画像認識における位置情報の有効性が高い認識カテゴリと，そうでないカテゴリを明確に区別することを目標として位置付けている．

一般に，認識対象と位置は密接な関係があり，例えば，海岸は海と陸の境目にしかなく，海や陸の真ん中には存在し得ない．しかしながら，海岸の写真の認識において位置情報を役立てるには，世界中の海岸の位置を学習データに持っておかないといけない．これには，膨大なデータを用意する必要がある．そこで，我々は，写真の撮影位置の地理的な状況を表す情報源として，航空写真に注目している．写真の画像特徴量に合わせて，航空写真から抽出した画像特徴量を認識に用いることで，写真の撮影場所の地理的なコンテキストを反映した認識が可能となると考えている．

本研究においても従来の我々の研究と同様に，Flickr から収集した位置情報付き写真と，それぞれの位置情報に対応する複数の縮尺 (レベル) の航空写真を用いる．ただし本研究においては，認識実験において画像特徴と位置情報特徴のどちらがどれだけ有効に作用したかについて，より明確に判断するべく，機械学習の段階においてマルチカーネル学習 (MKL) を用いて，写真と航空写真の重みを推定し，写真と位置情報の有効性を評価する．

本稿は以下のように構成される．まず実験の全体的手順と本研究の方針について第 2 節で触れる．後述するように，実験には画像収集，特徴抽出，機械学習の手順を踏むが，特徴抽出で適用する手法の詳細は，第 3 節で説明する．また，機械学習に用いる手法については第 4 節で述べる．実験方法と評価方法，結果を第 5 節で考察し，第 6 節で結論付ける．

## 2. 手順と方針

我々の認識実験は全体的に，**図 1** に示す要領で行われる．全体の流れとしては，Flickr より収集した位置情報付き画像から特徴抽出したものを，機械学習することで認識精度を検証するものである．

†1 電気通信大学大学院 電気通信学研究科 情報工学専攻
Department of Computer Science, The University of Electro-Communications
†2 電気通信大学 電気通信学部 情報工学科
Department of Computer Science, The University of Electro-Communications

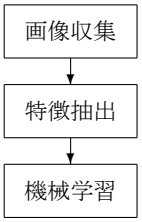


**図 1** 研究の全体的手順
Fig. 1 General procedure for our work

機械学習の段階において，画像と位置情報がどのように有効に利用されているかどうかを考察する必要がある．これにあたり従来の我々の手法[6),7)] では，位置情報と航空写真 (ないしは位置情報) を単純に結合した特徴量を学習させ，結合する特徴の種類の組み合わせによりこれを検証していた．ただし組み合わせごとの精度を比較するのみでは，画像と位置情報との有効性を柔軟に判断する上では不十分であった．

本研究では，第 4.2 節で説明する MKL-SVM を用いて，認識精度のほかに画像と航空写真の特徴の有効性を重みの推定により検証する．各手順において用いる手法の詳細については後述する．

## 3. 画像からの特徴抽出

本節では，画像から認識に必要な特徴を抽出する方法について述べる．画像の特徴を記述する手法としては，本実験では，特徴抽出のために局所特徴の一種である SIFT 特徴を用いる．また，この局所特徴を簡潔に記述するために bag of keypoints 手法を用いてデータをベクトル量子化する．画像と航空写真のいずれについても同じ特徴を抽出する．

### 3.1 SIFT 特徴

SIFT(Scale Invariant Feature Transform)[2)] とは，David Lowe によって提案された特徴点とそれに付随する特徴ベクトルの抽出法である．画像からいくつかの特徴点を抽出し，特徴点周りの局所画像パターンを勾配方向ヒストグラムから成る 128 次元特徴ベクトルで表現する．SIFT 特徴は，回転や拡大縮小に対してロバストであるので，類似画像に対しては類似する特徴点を多く保持することができる．

特徴点の抽出は，自動で行う方法 (ガウシアン差分による選定) と，手動で指定する方法 (主に格子状あるいはランダム点指定) がある．自動で行う手法は，主に特定の物体同士の対応点を検出するのに有用であるが，本実験では，特定の物体を認識することに固執しないので，10 ピクセルの格子状に点を抽出する方法を採用する．

### 3.2 Bag of visual words 表現

bag of keypoints モデル[1)] とは，画像を局所特徴の集合と捉えた手法である．局所特徴の特徴ベクトルをベクトル量子化し，visual words と呼ばれる特徴ベクトルを生成する．それらの集合をコードブックと呼び，それを記述子として画像の特徴ベクトルを生成する．これにより画像を visual words の集合 (bag) として表現する．ベクトル量子化は，ユークリッド距離を尺度とし，k-means 法を用いて行う．本研究では，処理速度や結果における精度の差異を考慮した上で，クラスタ数を 1000 に固定した．

**表 1** 認識カテゴリの定義
Table 1 Definisions of the categories

| カテゴリ | 定義の詳細 |
|---|---|
| 景色 | 遠くが見える景色で，近くに物体がないもの |
| ラーメン | 食べられる状態にあるラーメン |
| ディズニーリゾート | 位置情報などから，ディズニーリゾートで撮影されたと推測できる写真 |
| 東京タワー | 東京タワーが目立つように写っている写真 |
| 山 | 山頂を含む山の景観が目立つように写っている写真 |
| 神社 | 実際の神社の社殿や，鳥居などの建築と思われるもの |
| 鉄道 | 電車又は気動車が目立つように写っている写真 |
| 道路 | 道路が全体的に目立つように写っている写真 |
| 花 | 花がアップで撮影されている写真か，写真の殆どが花で占められているような写真 |
| 海岸 | 実際の海岸が写っている写真 |

## 4. 学習と分類

本節では，機械学習で用いる手法の詳細を説明する．認識精度の検証に当たっては，我々の従来の手法と同様にサポートベクタマシン (SVM) を用いる．

### 4.1 サポートベクタマシン

サポートベクタマシン (SVM) はニューロンのモデルとして最も単純な線形しきい素子を用いて，2 クラスのパターン識別器を構成する手法である．カーネル学習法と組み合わせると非線形の識別器になる．本実験では，カーネル関数として非線形の $\chi^2$ カーネルを用いる．

### 4.2 マルチカーネル学習

本研究では，特徴を統合して画像を認識するために，複数の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，それをサポートベクターマシン (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 $j$ を学習する．これはマルチカーネル学習 (Multiple Kernel Learning, MKL)[3)] 問題と呼ばれ，統合カーネルは以下で定式化される．

$$K_{\text{combined}}\left(\boldsymbol{x},\, \boldsymbol{x}'\right) = \sum_{j=1}^{K} \beta_j k_j\left(\boldsymbol{x},\, \boldsymbol{x}'\right)$$

$$\beta_j \geq 0,\ \sum_{j=1}^{K} \beta_j = 1 \qquad (1)$$

最近の研究では，この MKL 問題を凸面最適化問題として効果的に解く方法が提案されている[4)]．マルチカーネル学習は SVM のみを前提としたものではないが，SVM のフレームワークで解く方法が一般的で，MKL-SVM と呼ばれることもある．本研究では，SHOGUN[5)] ツールキットを用いて実装した MKL-SVM を使用して実験を行う．

## 5. 実 験 方 法

実験は，各画像から抽出した特徴量をサポートベクタマシン (SVM) で学習させ，分類結果により精度を判定することにより行う．学習と分類は正解画像と不正解画像の 2 クラスで行う．正解画像については，画像の内容に応じていくつかのカテゴリのデータセット用意し，不正解画像は正解画像に用いないデータをラ

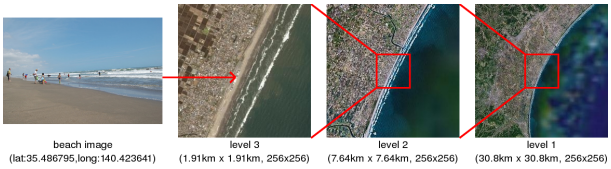


**図 2** 位置情報付き写真と航空写真の対応付け
Fig. 2 Correspondence between a geotagged photo and aerial photos.

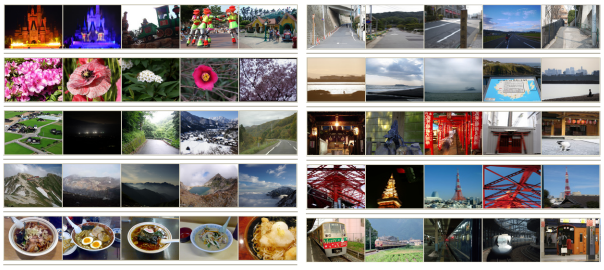

**図 3** 実験に用いる正解データセットの例．左列はそれぞれ上からディズニーリゾート，花，景色，山，ラーメンを，右列はそれぞれ上から道路，海岸，神社，東京タワー，鉄道を示す．
Fig. 3 Examples of the positive datasets.

ンダムに選択することで構成される．また，SVM のフレームワークの下で，マルチカーネル学習で種類ごとの重みを推定する (MKL-SVM) ことにより，画像と航空写真のどちらがどれだけ有用であるかを判断する．比較のため，画像のみの SVM によるで学習・分類と，従来の手法である，画像と航空写真の特徴量をベクトル結合させた特徴量の SVM によるで学習・分類も併せて行う．

### 5.1 データセット

正解データセットについては，**表 1** に示すような 10 種類のカテゴリを準備した．これらのデータは，タグなどのメタデータである程度分類したものの中から，カテゴリごとに写真の外観，位置情報ともに適切なものを手動で日本国内で撮影された 100 枚を選定した．正解画像の一例として，**図 3** を挙げることができる．

不正解データセットは，Flickr から収集した画像データの中から，正解データに用いられていないものをランダムに選定することで 100 枚準備した．

いずれのデータセットも，一般性を保証するため写真の撮影者が偏向しないように選定した．

画像は縦と横のうち長い方の画素が 500 ピクセルになるようにリサイズしてある．

### 5.2 航 空 写 真

航空写真を利用するにあたり，それぞれの画像に対応する位置情報を地図サイトで表示し，スクリーンキャプチャしたものを特徴抽出に利用する．特徴抽出の方法はデータセットの画像と同様である．写真の内容によっては，より詳細な縮尺を持つ航空写真の方が検証に際し有効であると考えられるが，撮影位置によっては，詳細な航空写真を入手できないこともある．広範なデータセットに対応するため，**図 2** に示すような 3 種類のスケールを採用する．図 2 に示すように，航空写真は撮影位置が中央にくるような正方形に加工したものを利用する．

### 5.3 評　　価

認識実験に際しては，後述のように写真の内容を予め各カテゴリに分類し，カテゴリごとに，画像と 1 種類の縮尺の航空写真をペアにしたものと，画像と全種類の航空写真を組み合わせたものを，それぞれ MKL-SVM を用いて学習・分類・重みの推定を行う．

学習と分類はクロスバリデーションで行う．すなわち，データセットを 5 分割し，1 つを分類用に，残り 4 つを学習用に充てることを，分類用に相当する各 fold について 5 通り繰り返す．

認識精度は各分類結果における SVM の出力値に対して，以下のように平均適合率を計算することで評価する．SVM による出力値に基づいて，テストデータをソートする．ソートしたデータを最初から順番に読み込み，positive のデータが出現した時点で，それまでの読み込んだデータの数 $m_i$ と，positive データの出現頻度 $i$ を記録する．ここで $p_i = \frac{i}{m_i}$ とおく．最後までテストデータを読み込んだときのすべての positive データの出現数を $n$ とすると，平均適合率 (avarage precision) は，

$$P = \frac{1}{n}\sum_{i=1}^{n} p_i \qquad (2)$$

で計算される．

### 5.4 実 験 結 果

画像と航空写真を MKL-SVM で分類した結果について示す．ここでは，画像と航空写真の 2 組について MKL-SVM で分類，重み推定を行った．結果については，5 回のクロスバリデーションの平均値を示す．

平均適合率の結果を**表 2** に示す．表 2 における Level1〜3 は，画像と，航空写真のいずれかの縮尺の 2 種類を MKL カーネルとして入力した結果であるが，括弧内に MKL ではなく画像特徴との単純なベクトル結合により SVM のみで学習させたときの結果を比較として示す．また，Multi は，画像と，航空写真のすべての縮尺の 4 種類を MKL カーネルとして入力した結果である．更に比較のため，画像のみを SVM で分類させたものをベースラインとして左端の Image に示す．この結果から，MKL を利用したものについては認識精度の上昇を確認することができ，カテゴリによっては，認識精度に改善が見られたことがわかる．「山」のようにベクトル結合の方が認識精度が高いカテゴリも存在するが，単純なベクトル結合の場合は，画像特徴と位置情報のいずれにいかなる効果があったかを明確に追究することは困難である．位置情報が単なる付加的情報ではなく，適切な重みにより分類されていることを確認するために，**表 3** に画像と全ての種類の位置情報の 4 カーネルで MKL を行った時の重みの比較を示す．更に，**表 4** に画像と位置情報の 2 カーネルで MKL を行った時の重みの比較を示す．表 4 における LV1〜3 は，画像と，航空写真のいずれかの縮尺のものを MKL カーネルとして入力した場合のものを示す．いずれも全体として，画像特徴としての位置情報を適度に利用した上で認識精度を上げたものであると考えてよい．

「ディズニーリゾート」と「東京タワー」は，位置情報の重みが比較的大きいカテゴリである．特にディズニーリゾートについては，位置情報が局所的に集中することから，航空写真がほとんど同一となり，航空

**表 2** MKL-SVM における平均適合率の計算結果．Level1〜3 は，それぞれ画像とそのレベルの航空写真を MKL カーネルとしたときの結果であるが，括弧内に MKL ではなく画像特徴とのベクトル結合による結果を比較として示す．Image は画像のみの分類結果である．Multi は，それぞれ画像とすべての航空写真を MKL カーネルとしたときの結果である．
Table 2 The results of average precisions on MKL-SVM for each concept.

| | Image | Level1 | Level2 | Level3 | Multi |
|---|---|---|---|---|---|
| ディズニー | 64.95 | 83.50 (*83.62*) | 83.61 (83.49) | **83.74** (83.57) | 83.50 |
| 花 | 75.31 | 76.00 (75.37) | 76.89 (74.78) | **77.23** (*77.50*) | 76.82 |
| 景色 | 80.60 | **83.02** (78.95) | 82.74 (78.81) | 82.62 (79.72) | 82.91 |
| 山 | 76.54 | **79.08** (*82.99*) | 78.21 (*82.65*) | 78.34 (*82.92*) | 78.95 |
| ラーメン | 79.71 | **82.49** (82.03) | 81.72 (*81.77*) | 80.85 (*80.68*) | 82.13 |
| 道路 | 77.13 | 77.82 (77.56) | 77.73 (77.24) | **78.38** (*77.88*) | 77.77 |
| 海岸 | 80.51 | 82.14 (*82.18*) | 81.38 (81.31) | 81.52 (*81.73*) | **82.21** |
| 神社 | 74.35 | 75.70 (*75.81*) | 76.39 (74.62) | **76.56** (*76.82*) | 75.98 |
| 東京タワー | 79.48 | **83.47** (83.29) | 83.44 (83.08) | 83.13 (*83.49*) | 83.28 |
| 鉄道 | 77.18 | **78.66** (78.00) | 77.26 (75.83) | 77.38 (*78.40*) | 78.26 |

**表 3** MKL による重み推定結果．位置情報付き写真と位置情報 (航空写真 (Level1〜3)) をすべて組み合わせた結果を示す．
Table 3 Weight estimation results for combination of geotagged photos and multiple aerial photos (LV1〜3) by MKL.

| | Image | Level1 | Level2 | Level3 |
|---|---|---|---|---|
| ディズニー | 0.4049 | 0.3998 | 0.1116 | 0.0837 |
| 花 | 0.8812 | 0.0315 | 0.0648 | 0.0224 |
| 景色 | 0.5235 | 0.4427 | 0.0074 | 0.0263 |
| 山 | 0.7987 | 0.1445 | 0.0000 | 0.0568 |
| ラーメン | 0.8106 | 0.1051 | 0.0796 | 0.0048 |
| 道路 | 0.9306 | 0.0000 | 0.0689 | 0.0005 |
| 海岸 | 0.6604 | 0.2632 | 0.0115 | 0.0649 |
| 神社 | 0.8802 | 0.0508 | 0.0335 | 0.0356 |
| 東京タワー | 0.4149 | 0.2266 | 0.0000 | 0.3586 |
| 鉄道 | 0.8767 | 0.1081 | 0.0092 | 0.0061 |

写真が認識に大いに貢献すると考えられる．一方，東京タワーの場合，タワーの遠方から撮影された写真もあるので，前者ほど航空写真の外観が極度に一致することはない．「山」や「海岸」，遠方の「景色」は上述のカテゴリ程ではないが，航空写真の有効性が広範囲の航空写真であるレベル 1 の場合，比較的高い．これは，レベルの低い航空写真が自然の地形 (山脈や海面，遠方にある都会や緑の風景) の特徴をよく表わしており，それらの類似した特徴がこれらのカテゴリに多く集中したためと推測される．航空写真はその画像特徴によって都市部か山間部かを区別するのが容易であるため，都市部にラーメン店が多く存在するという理由から「ラーメン」においても航空写真による位置情報の有効性が高い．「道路」や「神社」，「鉄道」や「花」は位置情報のばらつきが高いため，位置情報はあまり有効でなく，写真自体の特徴量の有効性が高くなったと考えられる．

今回実験に使用した航空写真は，地上の物体や建築物を視認できるほど詳細なものではなく，それ故に周辺に存在する都会や森林や水面などの粗い特徴量を多く学習したことになる．そのため，位置情報があまり効かなかったカテゴリについて，高いレベルほど精度向上に寄与したとは限らないことが言える．この問題を厳密に検証するにあたり，今後はより詳細な航空写真を実験に採用することを検討する．

## 6. おわりに

本研究では，位置情報付き写真の一般画像認識において写真の撮影位置に対応する航空写真を付加的な画像特徴量として利用する研究手法を改良した．撮影位置の情報は画像認識の精度向上に対していかに貢献するかを検討するにあたり，SVM による機械学習において得た認識精度と，マルチカーネル学習 (MKL) での重み推定による写真と位置情報の有効性の度合いを評価した．本実験により，航空写真を利用した分類が有効であるカテゴリが存在することが示された．

今後予定する取り組みとしては，多様なカテゴリと多種にわたる画像特徴において分類を行い，画像認識における位置情報の有効性についてより明確にする必要があることである．これを実現するにあたり，大量の画像の中から必要ないしは有効なものを効率的に収集し，データセット作成を強化することを考慮する．画像特徴や画像に関連するデータに関しても，色特徴や位置における周辺情報を活用するほか，時間帯や季節情報などの時系列に関連した情報についても認識に用いることを検討する．また，航空写真においても，より明確な特徴を持つ画像を利用する必要があるために，より詳細な航空写真を用いて実験することを視野に入れる．

**表 4** MKL による重み推定結果．位置情報付き写真と 1 種類の位置情報 (航空写真 (LV1〜3)) を組み合わせた結果を示す．
Table 4 Weight estimation results for combination of geotagged photos and aerial photos (LV1〜3) by MKL.

| | 種類 | 画像特徴 | 位置情報 | | 種類 | 画像特徴 | 位置情報 |
|---|---|---|---|---|---|---|---|
| ディズニーリゾート | LV1 | 0.4020 | 0.5980 | 道路 | LV1 | 0.9061 | 0.0939 |
| | LV2 | 0.5091 | 0.4909 | | LV2 | 0.9334 | 0.0666 |
| | LV3 | 0.3576 | 0.6424 | | LV3 | 0.9353 | 0.0647 |
| 花 | LV1 | 0.8700 | 0.1300 | 海岸 | LV1 | 0.7036 | 0.2965 |
| | LV2 | 0.9119 | 0.0881 | | LV2 | 0.9204 | 0.0796 |
| | LV3 | 0.9269 | 0.0731 | | LV3 | 0.8929 | 0.1071 |
| 景色 | LV1 | 0.7898 | 0.2102 | 神社 | LV1 | 0.8570 | 0.1430 |
| | LV2 | 0.9182 | 0.0818 | | LV2 | 0.9137 | 0.0863 |
| | LV3 | 0.8897 | 0.1103 | | LV3 | 0.9092 | 0.0908 |
| 山 | LV1 | 0.4869 | 0.5131 | 東京タワー | LV1 | 0.5156 | 0.4844 |
| | LV2 | 0.8173 | 0.1827 | | LV2 | 0.6264 | 0.3736 |
| | LV3 | 0.7951 | 0.2049 | | LV3 | 0.4601 | 0.5399 |
| ラーメン | LV1 | 0.7855 | 0.2145 | 鉄道 | LV1 | 0.8654 | 0.1346 |
| | LV2 | 0.8840 | 0.1160 | | LV2 | 0.9611 | 0.0389 |
| | LV3 | 0.9627 | 0.0373 | | LV3 | 0.9729 | 0.0271 |

## 参考文献


1) G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *Proc. of ECCV Workshop on Statistical Learning in Computer Vision*, pp. 59–74, 2004.
2) D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110, 2004.
3) G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, Vol. 5, pp. 27–72, 2004.
4) S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schaölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, Vol. 7, pp. 1531–1565, 2006.
5) Shogun - A Large Scale Machine Learning Toolbox. http://www.shogun-toolbox.org/
6) K. Yaegashi and K. Yanai. Can Geotags Help Image Recognition?. *Proc. of the Pacific-Rim Symposium on Image and Video Technology*, pp. 361–373, 2009.
7) 八重樫 恵太, 柳井 啓司: 撮影位置の情報を用いた一般画像認識の可能性の検討, 情報処理学会 CVIM 研究会, pp.15–22, CVIM163-3, (2008)