深層学習による車載映像の都市名推定

山口莞爾 *1　福元和真 *1　松下侑輝 *1　川崎洋 *2　小野晋太郎 *2　池内克史 *2
鹿児島大学大学院 理工学研究科 *1
東京大学 生産技術研究所 *2

近年，ドライブレコーダーなどの車載カメラの増加と Web による画像や動画の共有サービスの一般化により，世界中の都市の風景画像や映像をインターネットから取得することが出来るようになってきた．しかし，これらの画像や映像には，GPS などの位置情報が必ずしも付加されているとは限らない．このようなシーン情報から撮影位置を同定する研究が盛んに行われているが，世界中の都市を対象とした研究例はほとんど無く，成功したと言える事例も知られていない．そこで，本論文では世界中の映像を対象として，大域的な位置推定を行うことを研究目標とする．提案手法では，Google Street View の画像を学習データとして深層学習 (Deep Learning) と Suport Vecter Machine (SVM) による認識を行う．しかし，検証実験から，カメラごとの特性の違いによる見えの変化が都市推定に影響することが分かった．また，都市の数が増えると難易度が上がると考えられる．これらの問題から，今回は主要都市のみを対象として，見えの違いを軽減した手法が，どの程度の認識が可能か検証を行った．

Global Localization of Vehicle Mounted Video using Deep Learning and Perspevtive Transformation

Kanji Yamaguchi*1 Kazuma Fukumoto*1　Yuki Matsushita*1　Hiroshi Kawasaki*2
Shintaro Ono*2　Katsushi Ikeuchi*2
Kagoshima University Graduate School of Science and Engineering*1
The University of Tokyo Institute of Industrial Science*2

Abstract In recent years, by the spread of sharing services of images and videos, we have come to be able to acquire landscape images and vehicle mounted videos of several cities in the world from the Internet. This information will be useful for scene understanding, the three-dimensional model generation of the city and the frequent update of the map, but may provide the wrong information when the data has no correct geographical information. On the other hand, many researchers try to take the information that a human being perceives out of the information about the scene and study on labeling of the scene, but there are few successful methods to identify images taken from plural cities. Therefore, we aim for the global localization of images which are taken in various cities in the world using only the visual information. To this end, we learned the information of each city from a Google street view image and constructed the city identification database, then estimated the city where an inputted image taken from. In ordert to confirm the effectiveness of proposed methods, we proved it using images acquired from 15 cities and obtained several knowledge.

Keyword: *Area Estimation, Scene Recognition, Deep Learning, Support Vector Machine*

1. はじめに

近年，ドライブレコーダーなどの車載カメラの増加と Web による画像や動画の共有サービスの一般化により，世界中の都市の風景画像や映像をインターネットから取得することが出来るようになった．しかし，これらの景観データには GPS のような位置情報が必ずしも付加されているとは限らない．そのため，このようなシーン情報から撮影位置を推定する現在研究が盛んに行われている．

そこで，我々は予め推定したい都市の画像データセットを用意し，そこらから各都市を代表する特徴量の抽出を行い，これを学習データとして識別器を構築し，都市レベルでの広域的な位置推定を行った．しかし，後述の実験から，カメラごとの特性の違いによる見えの変化が都市推定に影響することが分かった．学習・識別それぞれに用いられる画像や映像は，異なる照明環境，異なるカメラで撮影されているため，実際に利用するためには，これらのカメラ特性の違いを軽減する必要がある．また，推定したい都市の数が増えると，学習データも増加していくことから識別性能が低下すると予想される．これらに対する 1 つの解決策として，階層的な学習による特徴抽出手法が考えられる．また，本論文では，異なるカメラ特性間での幾何的・光学的な問題を克服するために，射影変換を用いた見えの違いを軽減した手法を検証した．

本論文では，上記手法の可能性を検証するため，世界中の都市から 15 都市を選び，近年注目を集めている深層学習 (Deep Learning) の手法を用いて実験および評価を行う．具体的には，一般物体認識において高い認識性能を示す Convolutional Neural Network (CNN) から画像特徴量を抽出し，信頼性の高い線形分類器である Support Vector Machine (SVM) で識別器を構築し，テストデータの都市レベルでの広域的な位置推定を行う．

2. 関連研究

位置推定の手法として，本論文は初期位置情報が全く無い状態からその大まかな撮影位置を推定する手法を対象としている [4]．

これまで提案されている大規模な位置推定の研究として，シーンを理解することで撮影位置を推定する手法が提案されている．具体的には，1 枚の画像から無数のパッチ画像を生成し，各パッチから輝度勾配と輝度強度を取り出す特徴量である Histogram of Oriented Gaussian (HOG) を抽出し利用する．

このような考えに基づく研究として，Doersch らは，世界中の都市で撮影された Google Street View 画像の撮影都市を推定する手法を提案している [1]．彼らの手法では，学習に建物正面を向いた大量のパッチ画像から各都市を代表するパターンを抽出し，都市の推定を行った．しかし，彼らの手法ではその都市らしさを抽出することが目的であったため，非学習画像において，高い認識率を示す事が出来なかった．また，Fukumoto らは撮影位置の推定手法として，HOG 特徴量とランダムフォレストによる推定手法を提案した [3]．彼らの手法では，パッチ分割した画像をランダムフォレストで学習することで，各都市を構成する代表的なパターンを識別できた．しかし，彼らの手法ではパッチに地面や空などの識別に適さない画像が含まれ，また，テスト画像に都市を代表するパターンが含まれていなかった場合に，認識率が大きく低下する．

一方，近年，画像や映像の認識や識別を行う手法として，深層学習を使った画像の撮影位置推定や特徴抽出手法が数多く提案され，顕著な識別結果を示している．深層学習は従来のニューラルネットワークの層を増やし学習効率を高めた手法であり，顔認識や一般物体認識，自然言語処理といった分野で数々の結果を更新している [8]．この深層学習を使い画像特徴量抽出を行う方法として，CNN がある．CNN は畳み込みとプーリング処理を繰り返すことで，対象物体の特徴量を効率よく抽出することが出来るため，物体認識の分野で現在盛んに用いられている．

Khosla らは 2 枚の画像が与えられたときに，どちらがより目的の場所に近いかを推定する手法を提案した [6]．また，Ordonez らは，画像から犯罪の多い地域を学習し推定する手法を提案した [7]．彼らの手法では，画像特徴量として CNN を使用し，人間の無意識な判断や知識を画像から取り出す手法を提案した．また，Zhou らは Google Street View の画像の撮影された都市を推定する手法を提案した [10]．彼らの手法も CNN を特徴量として採用し，学習と識別には SVM を用いた．加えて彼らは独自でシーンを識別するための学習モデルの提案を行った [9]．彼らは撮影画像を予め決めた 7 つのシーンカテゴリに分類し，そのカテゴリ毎に識別器を構成し認識を行った．都市は様々な情報から構成されており，海に近い街や治安の悪い街など都市によって様々である．彼らの手法では，このようなシーンの出現する差異を利用して街の特定を行った．

3. 提案手法

本論文では，市街を撮影した車載映像が与えられたときに，都市レベルで撮影位置を推定することを目的としている．提案手法では，カメラ特性による画像間の違いを軽減するために射影変換を行った．また，深層学習を用いて撮影位置の推定を行うが，そのほとんどは学習に用いていない画像とする．提案手法は学習と識別の 2 ステップで構成されている．手法の概要を以下に述べて，続いて詳細を説明する．

3.1 概要

手法の概要を図 1 を用いて説明する．

学習ステップでは，まず，CNN により都市画像の特徴を抽出する．今回は，Zhou らによって生成・公開されている，屋外の属性分類と一般オブジェクト分類用の学習済みモデル [9] を特徴抽出器として採用している．画像を与えると，複数回の畳み込み処理を経て，全結合層の第 6 層から 4096 次元の特徴ベクトルを取得する．その後，これを SVM に与えることにより都市識別器を構築する．

識別ステップでは，テスト画像を同じニューラルネットワークに与えて特徴量を抽出し，これを更に構築した SVM による都市識別機に与える．ただし，提案手法では幾何的影響を軽減するためにテスト画像に予め射影変換を施す．

本論文では，車載映像を識別のターゲットとしているため，対象とする画像の前後のフレームを用いて，識別精度の向上が可能となる．今回は，150 フレームを 1 シーケンスとし，1 シーケンス毎の識別結果を求めて，シーケンス内の画像すべての識別結果の総和を用い，最も出現数の高かったクラスをそのシーケンスの撮影された都市とした．

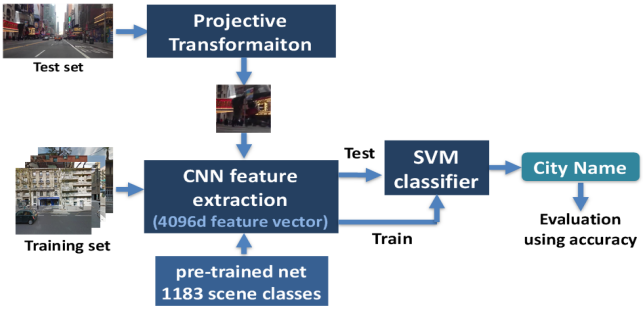

Fig.1 提案手法の概要

3.2 CNN によるシーンの学習

一般に，これまでの機械学習においては，SIFT や HOG に代表されるような特徴量をあらかじめ人手により定めて（既存のものを利用，あるいは自ら設計して）おき，それを学習に用いていた．例えば [3, 5] では，車載映像をパッチ分割して HOG 特徴量を抽出し，これをランダムフォレストに与えて学習・識別を行っている．しかし，この場合，パッチ分割によって地面や空などの特徴のない画像が学習・識別に使用され，識別率の低下を招く結果となっていた．また，人手により特徴量を設計する際は，様々なパラメータを細かく設定する必要がある．例えば HOG では，stribe の間隔，1 パッチあたりのサイズなどのパラメータがあり，これを変更することで最終的に取得される特徴量の次元が大きく異なる．加えて，適切なパラメータを選択するためには，パラメータの調整を複数回行った後，学習モデルを生成し，識別率を確認するという処理を繰り返す必要がある．加えて，これらの調整は全て人間の知識や経験に基づくものである．

これに対して，近年，上記問題を解決する手法として深層学習が注目されている．例えば深層学習の一手法である CNN を用いると，すべてのパラメータチューニングを誤差逆伝搬法によって自動的に行うことができるため，人間の知識や経験に頼らなくても学習を行うことができる．CNN は，畳み込み層やプーリング層，局所コントラスト正規化層を多層に組み合わせた後に，ニューラルネットワークで学習する手法である．畳み込み層ではフィルタと類似性のある特徴量を抽出することが可能であり，プーリング層では画像に映る物体の局所的な微動性に対して，ある程度の普遍性を確保できる．そこで，我々は CNN の中間層から抽出できる特徴量を，高い識別性能を誇る SVM で学習し，広域的な位置を推定する手法を採用した．

3.3 学習と識別に用いたデータ

深層学習においては，学習データと識別データの確保が問題となる．一つの解として，世界の主要都市を対象とするため，インターネット上で公開されている画像をデータセットとして用いることが考えられる．特に，本論分では，Google Street View 画像と YouTube の車載映像を対象としている．

今回は学習データとして Google Street View から，京都，ソウル，台湾，大阪，バンコク，リオン，ブリュッセル，ベルリン，プラハ，パリ，ヒューストン，シカゴ，フィラデルフィア，サンパウロの 15 都市から 10 万枚以上の画像をダウンロードした．

識別に用いるテストデータは上記の 15 都市の Google Street View 画像と YouTube 上の不特定多数の一般ユーザによって撮影された車載映像を用いた．YouTube 映像を識別に用いる場合は，フレーム毎に切り出してから用いる．我々は各都市から 10 本ずつ車載映像をダウンロードし，テストデータとした．各映像は，10 分から 45 分程度で構成され，画質やノイズが著しく低い映像は使用していない．

3.4 異なるカメラ特性間による画像の差異

Google Street View は世界各国でほぼ同じ仕様・規格で撮影され、また都市シーンのバリエーションが豊富なため，これを学習データとして使用するのが好適と思われる．対して，YouTube で撮影された車載映像は多くのユーザーによって投稿されているため，カメラのバリエーションが豊富になる．

しかし，後述の実験から，これらのデータセット間で学習・識別を行った場合に，識別率が大幅に低下することが分かった．これは，異なるカメラ間で撮影された画像に，幾何的・光学的な問題が発生するためと考えられる．図 2 は，異なるカメラ間における画像の違いを示している．この問題については後述の実験で詳述する．

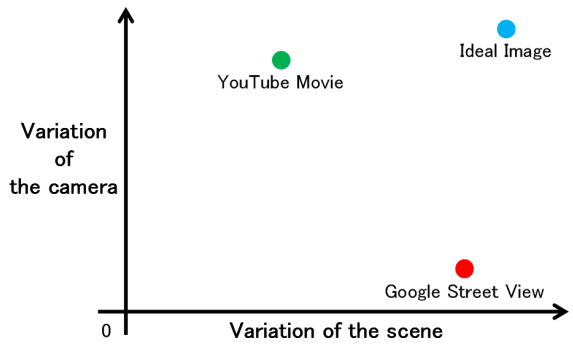

Fig.2 異なるカメラ間で撮影された画像の違い

3.5　射影変換による認識率の向上

前述のように，CNN は過去の多くの問題を解決可能のため大きく注目されているものの，万能という訳ではない．本論文においても，後述の実験のように YouTube で撮影された映像から切り出した画像と Google Street View で撮影された画像を比較する際には，同じシーンでも画像の幾何的・光学的な違いが問題となる．

そこで，識別率を低下させる要因がどこにあるのかを調査するため，我々はまず，ホワイトバランスを変化させてその影響を調べた．その結果，全体に対する色の変化はほとんど識別に影響しなかった．一方でガンマ値を変化させると識別率が大きく低下した．そこで，実験では，光学的な影響を少なくするために，なるべく日中，同じ程度の明るさで撮影された画像のみを使用するようにした．

次に，幾何的な影響を調査するため，進行方向を向いた Google Street View 画像を学習し，建物を正面方向を向いた画像をテストデータとして撮影位置の推定を行った．この場合，識別率は非常に低いものであった．そこで，本論文では，幾何的影響を軽減するため，テスト画像を建物正面画像へ変換することで識別率の向上を試みた．このように見えの違いを補正し認識を行う手法として，Deep Face がある [2]．Deep Face では顔のモデルを使うことで撮影対象者の顔方向をあたかも正面から見たようにに補正し，認識を行う手法である．しかし，市街においては場所ごとに 3 次元形状が大きく異るため，詳細なモデルを用いた補正は難しい．そこで我々は，市街における建物は，道路に沿った平面上に配置されているという単純なモデルを採用して射影変換を行った．平面位置を推定するためには，消失点を用いた．消失点は，建物に平行な直線群が無限遠で交わる点のため，水平・垂直それぞれの消失点が分かれば，射影変換行列を求めることができる．

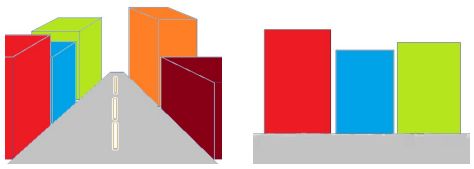

Fig.3 進行方向を建物正面の画像へと射影変換する模式図

4.　実験 1 ：進行方向の検証

4.1　Google Street View 画像間の学習・識別

まず，進行方向を向いた Google Street View 画像で学習を行い，学習に用いていない Google Street View 画像をテストに用いて評価を行った．評価結果を表 2 に示す．平均の識別結果は 86%で，概ね正しく識別が行えた．このことから，Google Street View 画像間では，正しくそれぞれの街の特徴量を抽出し，学習・識別が行えたことが分かる．

Fig.4 識別時の進行方向の Google Street View 画像例

Fig.5 学習時の進行方向の Google Street View 画像例

Table 1 Street View 学習・Street View 識別

	bangkok	kyoto	osaka	seoul	taiwan	berlin	bruxelles	lyon	paris	prague	chicago	houston	ny	philadelphia	sanpaulo
bangkok	82.4	0.9	0.7	0.3	3	0.1	0.3	0.8	1.7	0.7	0.8	0.7	1.3	1.6	4.7
kyoto	0.5	83.6	8.2	2	0.6	1	0.4	0.5	0.3	0.6	0.3	0.2	0.1	0.7	1
osaka	0.3	8.1	85.3	0.3	1.4	0	0.5	0.5	0.3	0.8	1.3	0.3	0.5	0.3	0.1
seoul	0.2	0.6	0.4	92	1	2.3	1.2	0.1	0.4	0.3	0.2	0.3	0.6	0.2	0.2
taiwan	3.8	0.7	1.5	3.5	77.2	0.4	1.3	0.4	0.6	1.7	1	0.9	3.2	2	1.8
berlin	0.4	0.8	0.4	2.6	0.6	90.4	1.4	0.1	0.6	0.7	0.2	0.6	0.6	0.6	0
bruxelles	1.6	1.5	1.1	2.8	1.9	2.6	60.6	5.9	5.7	5.3	2.1	1.2	2.7	3.3	1.7
lyon	1.7	0.8	0.6	0.8	0.2	1.9	5	68.8	9	4.5	1	0.9	0.6	1.8	2.4
paris	1.8	1.6	0.7	1.3	1.9	3.7	5.5	8.7	57.3	6.4	1.8	0.9	4.9	1.5	2
prague	1.4	0.7	0.5	1	1.2	1.1	3.7	4.3	3.7	71.1	0.6	2	4.9	1.7	2.1
chicago	1.4	1	1.7	0.2	2.8	0.2	2.5	1.9	1.2	1.4	66.5	9	4	4.6	1.6
houston	0.7	0.4	0.5	0.1	0.6	0.2	0.7	0.9	0.4	0.9	4.3	84.9	1.7	1.6	2.1
ny	1.7	0.2	0.4	0.6	3	0.5	2.6	0.9	2.9	4.4	3.8	1.6	69.9	5.5	2
philadelphia	2.1	0.3	0.9	0.8	1.9	0.6	1.5	1.3	1.3	2.4	5.8	2.8	6	68.9	3.4
sanpaulo	4.4	0.8	1.1	0.4	3.3	0	1.4	3.4	1.2	1.8	1.6	3.4	1.9	3.6	71.7

4.2　YouTube の車載映像識別

次に，この Google Street View 画像で学習した識別器に，YouTube の車載映像を与え性能評価を行った．ここでは各都市 10 本の車載映像からランダムに 500 フレーム取り出し識別に用いた．識別率は，12.53%とほとんどの都市で誤った識別が行われている．識別結果を表 2 に示す．

Table 2 Street View 学習・YouTube 識別

	bangkok	berlin	bruxelles	chicago	houston	kyoto	lyon	ny	osaka	paris	philadelphia	prague	sanpaulo	seoul	taiwan
bangkok	9	4	2	0	4	16	0	3	5	4	13	4	4	5	22
berlin	0	16	3	1	4	5	0	12	1	10	25	0	1	0	22
bruxelles	0	5	4	10	13	10	0	2	5	29	9	2	0	4	7
chicago	0	6	5	13	5	20	0	6	1	4	21	2	1	0	16
houston	0	8	4	3	25	5	0	1	4	6	21	2	1	7	13
kyoto	0	1	1	13	0	9	1	6	6	9	37	0	1	3	13
lyon	0	7	2	6	12	10	3	5	4	30	10	1	0	2	8
ny	0	2	2	16	4	3	0	20	0	6	19	1	1	2	22
osaka	2	13	2	4	3	12	1	3	3	9	16	1	4	15	12
paris	0	5	11	6	1	10	4	13	0	13	7	3	1	11	15
philadelphia	0	3	14	4	1	14	1	3	3	7	23	1	2	3	21
prague	0	9	11	5	4	8	6	5	3	17	17	5	0	2	8
sanpaulo	0	2	5	7	3	15	1	3	2	35	5	1	7	2	12
seoul	0	4	1	24	10	5	2	8	10	5	4	2	5	10	10
taiwan	0	3	6	6	0	7	6	8	14	11	3	0	2	6	28

4.3　YouTube の車載映像間の学習・識別

そこで，このような要因を検証するため，YouTube 映像の前半からランダムに抽出した画像を学習に用いて識別器を構築し，識別には映像の後半からランダムに選んだ画像を用いた実験を行った．学習には各都市 2000 枚ずつ，識別には 500 枚を用いた．結果を表 3 に

示す．平均の識別結果は 68%であった．

実験から，Google Street View 画像間での学習・識別では高い識別結果を示している．また，YouTube の車載映像間でも識別率は下がるものの，同様の結果となった．つまり，同じカメラ規格で撮られた画像間では，識別率が高くなると考えられる．対して，Google Street View 画像を学習データに用いて，YouTube の車載映像をテストデータに用いた場合には，大きく識別率が下がる．このことから，学習データとテストデータでカメラ特性が異なる画像間では識別率が大きく低下することが考えられる．

ニューヨーク　京都　パリ

Fig.6 学習時の進行方向の YouTube 画像例

ニューヨーク　京都　パリ

Fig.7 認識時の進行方向の YouTub 画像例

Table 3 YouTube 学習・YouTube 識別

	bangkok	kyoto	osaka	seoul	taiwan	berlin	bruxelles	lyon	paris	prague	chicago	houston	ny	philadelphia	sanpaulo
bangkok	79.6	1	0.8	2	2.2	2.4	0.4	1.2	0.6	2.2	2.4	2.8	1	0.2	1.2
kyoto	4	65.8	2.2	3.6	6	1	0.6	2.4	2	1.6	0.4	3.6	4	0.4	2.4
osaka	1.8	5.2	78.2	2.2	1	0	2	0.4	0.2	1	0.2	1.4	1.4	1.4	3.6
seoul	1.8	5.4	9.6	47.8	3.4	3.6	1.6	3.8	1.8	3.2	5.6	2.2	8.6	0.2	1.4
taiwan	2.6	1.2	5.4	8.4	63.4	1.2	3.6	1.2	4	3	0.8	0.8	2.8	0.6	1
berlin	2.2	0.8	0.2	4.8	0.8	57.4	8.8	5	5	7.2	0.6	3.2	1.2	2	0.8
bruxelles	2.6	1.4	4.8	5.8	3.4	1.8	65.4	1.6	4.4	1.8	1.2	1.2	2.8	0.2	1.6
lyon	0	1.8	0.6	2.2	2.4	6.4	9.8	49	3	13.4	1	5.2	2	3	0.2
paris	0.8	0.8	0.2	2.4	4.2	4.2	4	5.2	65	7.4	0	2.2	1.8	0.6	1.2
prague	1	0.6	0.2	2.4	0.6	4.4	3.8	1.6	3.2	72.8	3.6	2.8	1.4	0.8	0.8
chicago	2.2	0	0.2	2	2.2	2.2	1.2	1.4	1.2	3.6	72.4	6	3.6	0.4	1.4
houston	1.6	1	1.8	3	3.2	3.4	2.4	1.4	1	4	3.4	67.2	1.8	2	2.8
ny	2.6	1.8	1.4	3	3.2	2.6	0.8	1.8	2.8	2.6	3	3	67.2	1.2	3
philadelphia	1	2	0.4	1.4	1.2	0.4	2.2	3.4	1.8	3.4	2.4	2.4	1.6	74.6	1.8
sanpaulo	1.6	1	0.6	2.8	5.4	1.2	1.4	1	1.6	0.6	1.8	2.2	1.2	1	76.6

5.　実験２：射影変換による効果の検証

5.1　建物正面の Street View 画像間の学習・識別

次に，建物正面を向いた Google Street View 画像間で学習・識別を行った．Google Street View 画像は，ダウンロードの際に建物正面を向いた画像を予め取得できるため，この Google Street View 画像には射影変換を施していない．実験では，学習に各都市 2000 枚，認識に各都市 500 枚を使用した．結果から，識別率は 75%であった．結果を表 4 に示す．進行方向を向いた Google Street View 画像間の実験に比べて，やや認識率が劣るが，これは学習とテストに用いた画像セットの画像サイズが小さいためである．検証実験から，画像セットの解像度が減少することで認識率が下がることが分かっている．

ただし，それでも認識率は依然として高い．これは，前実験同様に，カメラ規格がほぼ統一されている画像間での学習・識別なためと考えられる．このときに CNN が出力する，畳み込み画像を図 8 に示す．建物正面を向いた画像では建物と空の境界や地面との境界が強調されること無く，画像が伝播されていることが分かる．そのため，建物の純粋なテクスチャ情報などが学習されたと考えられる．

Fig.8 入力画像と畳み込み画像

Table 4 建物正面での Street View 画像間の認識率

	bangkok	kyoto	osaka	seoul	taiwan	berlin	bruxelles	lyon	paris	prague	chicago	houston	ny	philadelphia	sanpaulo
bangkok	82.4	0.9	0.7	0.3	3	0.1	0.3	0.8	1.7	0.7	0.8	0.7	1.3	1.6	4.7
kyoto	0.5	83.6	8.2	2	0.6	1	0.4	0.5	0.3	0.6	0.3	0.2	0.1	0.7	1
osaka	0.3	8.1	85.3	0.3	1.4	0	0.5	0.5	0.3	0.8	1.3	0.3	0.5	0.3	0.1
seoul	0.2	0.6	0.4	92	1	2.3	1.2	0.1	0.4	0.3	0.2	0.3	0.6	0.2	0.2
taiwan	3.8	0.7	1.5	3.5	77.2	0.4	1.3	0.4	0.6	1.7	1	0.9	3.2	2	1.8
berlin	0.4	0.8	0.4	2.6	0.6	90.4	1.4	0.1	0.6	0.7	0.2	0.6	0.6	0.6	0
bruxelles	1.6	1.5	1.1	2.8	1.9	2.6	60.6	5.9	5.7	5.3	2.1	1.2	2.7	3.3	1.7
lyon	1.7	0.8	0.6	0.8	0.2	1.9	5	68.8	9	4.5	1	0.9	0.6	1.8	2.4
paris	1.8	1.6	0.7	1.3	1.9	3.7	5.5	8.7	57.3	6.4	1.8	0.9	4.9	1.5	2
prague	1.4	0.7	0.5	1	1.2	1.1	3.7	4.3	3.7	71.1	0.6	2	4.9	1.7	2.1
chicago	1.4	1	1.7	0.2	2.8	0.2	2.5	1.9	1.2	1.4	66.5	9	4	4.6	1.6
houston	0.7	0.4	0.5	0.1	0.6	0.2	0.7	0.9	0.4	0.9	4.3	84.9	1.7	1.6	2.1
ny	1.7	0.2	0.4	0.6	3	0.5	2.6	0.9	2.9	4.4	3.8	1.6	69.9	5.5	2
philadelphia	2.1	0.3	0.9	0.8	1.9	0.6	1.5	1.3	1.3	2.4	5.8	2.8	6	68.9	3.4
sanpaulo	4.4	0.8	1.1	0.4	3.3	0	1.4	3.4	1.2	1.8	1.6	3.4	1.9	3.6	71.7

5.2　射影変換された車載映像の識別

次に，提案手法の射影変換による効果を検証した．YouTube の車載映像から消失点を推定し，射影変換によって建物正面を向いた映像に変換した．この変換された映像の 150 フレームを 1 シーケンスとしてテストデータとする．また，前実験で建物正面を向いた Google Street View で構築された識別器をテストに用いた．結果を表 5 に示す．平均の認識率は 36%と，実験 1 の結果に比べて，ある程度認識率が向上した．結果から，極端な認識率向上ではないが，射影変換によって見えの違いを軽減できたと思われる．15 都市の識別のため，チャンスレートが 3%程度であることを考えると，高い認識率を示したとも言える．しかし，依然としてカメラ規格が統一された Google Street View 画像同士での認識結果が高い．

ニューヨーク　京都　パリ

Fig.9 学習時の建物正面の Google Street View 画像例

ニューヨーク　京都　パリ

Fig.10 射影変換前の建物正面の YouTub 画像例

ニューヨーク　京都　パリ

Fig.11 射影変換後の建物正面の YouTub 画像例

Table 5 射影変換後の YouTube 画像での認識率

	bangkok	kyoto	osaka	soul	taiwan	berlin	bruxelles	lyon	paris	prague	chicago	houston	ny	philadelphia	sanpaulo
bangkok	46.4	8.6	13.6	0	2.2	0.6	6	11.4	0	0.4	4.4	0	0	4	2.4
kyoto	0.6	38	44.6	0	0.6	0.6	3.2	7.2	0.2	0	3.4	0.4	0	0.6	0.6
osaka	0	1.8	92.6	0	0.6	0	1.6	2.2	0	0	0.6	0.2	0.2	0.2	0
soul	1.4	2.4	22.2	9.8	8.4	1.4	21.2	14	0.4	0.2	15	0.4	0	3.2	0
taiwan	2	5.4	28.8	0	37.8	0.4	8.2	8.6	0.2	0.2	5.4	0	0	2.6	0.4
berlin	0.4	1.2	7.4	0.4	1.2	50.8	12.6	14.2	0.8	0.2	8	1.2	0	1.6	0
bruxelles	0.8	3.8	17.6	0	1.8	1.8	49.8	18	0	0.2	4.2	0	0	2	0
lyon	0	1.6	12.8	0	0.8	1	12.4	67.2	0	0.6	1.8	0	0	1.6	0.2
paris	0.2	3.2	19	0.2	1.2	2.2	16	45.6	3.6	0.8	4.2	0	0.2	3	0.6
prague	0.8	4.8	22.4	0.2	1.4	1.4	14.2	28.2	0.2	18.4	3.2	0	0	4.6	0.2
chicago	0.2	2.4	31.6	0	1.4	0.4	8.6	9	0	0.2	41	0.4	0	4.6	0.2
houston	0	3	31.8	0.2	0.6	0.6	10.2	16.2	0	0	17.6	13.4	0	5.8	0.6
ny	1	2.4	27.6	0.2	3	0.4	13.2	9.4	0.8	1.4	20.2	0.2	0.6	19	0.6
philadelphia	0.6	3.6	11.6	0	0.8	0.6	7.2	9.2	0.2	0.2	11.2	0.4	0	53.8	0.6
sanpaulo	2	1.28	29.6	0	2.2	0.2	11.6	15.6	0	0	5.2	0.4	0	6.6	13.8

6. 結論

我々は初期位置情報が付加されていない映像や画像を，深層学習によってシーンの情報から広域的な位置推定を試みた．本実験から，画像のカメラ情報による影響が識別率を大きく左右していることが分かった．提案手法では進行方向を向いた車載映像を建物の正面を向いたように射影変換することで，カメラ固有の影響を軽減し，建物のテクスチャ情報を識別に利用することを提案した．

実験の結果，射影変換を施す前と後を比べて識別率が 12.53%から 36%に向上した．また，昨年我々が提案した HOG 特徴量とランダムフォレストによる 15 都市推定での認識率は平均して 15.34%であったのに対して，同じ条件下で，今回提案した手法では識別率が向上した．これは，深層学習の物体認識カテゴリにおける既存手法に対する優位性と同時に，射影変換によるカメラ特性の異なる画像間での見えの違いを軽減できたためと思われる．しかし，依然としてカメラ規格が似通った画像同士での識別が高い結果となった．

今後の課題として，テクスチャ情報の共起性やオブジェクトの出現率など，統計的な分析によって更なる認識率向上が図れると思われる．また，光学的な補正も必須と考えられる．

参 考 文 献

1) C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *ACM SIGGRAPH*, 31(4), 2012.
2) H. Fan, Z. Cao, Y. Jiang, Q. Yin, and C. Doudou. Learning deep face representation. *CoRR*, abs/1403.2802, 2014.
3) K. Fukumoto, H. Kawasaki, S. Ono, H. Koyasu, and K. Ikeuchi. On-vehicle video localization technique based on video search using real data on the web. *International Journal of ITS Research*, 13(2), 2014.
4) J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *CVPR*, 2008.
5) Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *CoRR*, abs/1408.5093, 2014.
6) A. Khosla, B. An, J. J. Lim, and A. Torralba. Looking beyond the visible scene. In *CVPR*, 2014.
7) V. Ordonez and T. L. Berg. Learning high-level judgments of urban perception. In *ECCV*, 2014.
8) O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge, 2014.
9) B. Zhou, À. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495, 2014.
10) B. Zhou, L. Liu, A. Oliva, and A. Torralba. Recognizing city identity via attribute analysis of geo-tagged images. In *ECCV*, pages 519–534, 2014.