Vol.2014-CVIM-192 No.33
2014/5/16

都市の景観特徴の学習による初期位置情報の全くない車載カメラ映像からの撮影地域推定手法

福元 和真[1,a)]　川崎 洋[1]　小野 晋太郎[3]　子安 大士[2]　池内 克史[3]

概要：近年，ドライブレコーダーの普及による車載カメラの増加と，Web による動画共有サービスの一般化により，多くの車載カメラ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．これらの映像は安全運転や自動運転における学習への応用や，都市の 3 次元モデル生成への応用が期待できるが，様々な都市の映像が入り混じっていると学習結果の精度低下や誤ったモデル生成を引き起こす可能性がある．一方で，映像情報の中身を理解し，ラベリングする研究が盛んに行われているが，車載映像にこれを適用して撮影位置の同定に成功した事例はあまり知られていない．そこで，本論文では車載映像を対象として，大域的な位置推定を行うことを目標とする．提案手法では，予めストリートビューから各都市を代表する特徴的なパターンを抽出し SVM で学習させることで，撮影位置不明の車載映像を都市のスケールで推定する．手法の有効性を確認するため，3 都市の車載映像を用いて実験を行った．

1. はじめに

近年，ドライブレコーダーの普及や動画共有などに対する関心の高まりから，Web 上の動画サイトに数多くの車載カメラ映像がアップロードされており，多くの映像を取得することが出来る．これらの映像は安全運転や自動運転における学習の応用や，都市の 3 次元モデル生成への応用が期待できる．この時，様々な都市の映像が入り混じっていると学習結果の精度低下や誤ったモデル生成を引き起こす可能性があるため，どの国のどの都市で撮影されたかという大域的な位置推定が必要となる．しかし，これらの映像には GPS のような位置情報が付加されていることは稀であり，撮影位置の推定には視覚情報が有効な鍵となる．視覚情報による撮影位置推定を行う際，過去に撮影された画像を学習し構築した辞書を用いて撮影位置を推定する手法がある．しかし，学習・検索それぞれに用いられる画像や映像は，異なる照明環境，異なるカメラ視点で撮影されているため，実際に利用するためには，様々な撮影環境の画像や映像を用いて学習することが必要である．

そこで本研究では，予めストリートビュー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辞書を生成し，車載映像の撮影位置を都市レベルで推定することを目指す．手法の有効性を確認するため，3 都市の車載映像で識別実験を行った．

[1] 鹿児島大学
[2] 埼玉大学
[3] 東京大学
[a)] sc109061@ibe.kagoshima-u.ac.jp

2. 関連研究

これまで，視覚情報に基づいた撮影位置同定の手法は数多く報告されている．Efros らや Xiao らは，1 枚の画像から高次元の様々な特徴量を抽出し，マルチカーネル学習を使い撮影位置の特定を行っている [1]，[2]．しかしこの場合，計算コストの問題やメモリの問題が発生する．さらに，これらの手法では認識率を下げ得る多くのノイズが含まれている．このような問題を解決するため，各シーンから他のシーンでは現れない象徴的な特徴量を抽出することで，小さな計算コスト，メモリで撮影位置の推定が可能と考えられる．

Efros らは，世界中の都市で撮影されたストリートビュー画像の撮影都市を推定する手法を提案している．彼らは，ストリートビュー画像をパッチ画像に分割し，各パッチ画像から HOG 特徴量を抽出し，撮影位置の特定を行う手法を提案している [3]．彼らの手法では，学習に Google Street View から建物に対して正面を向いた画像をランダムに取得している．この時彼らは，各都市で得られた大量のパッチに対してクラスタリングを行い，各都市を象徴するパターンを抽出した．この各都市を象徴するパターンを SVM で学習させ，撮影位置の特定を行った．しかし彼らの手法では，クエリデータとして学習同様に Google Street View から建物の正面を向いた画像しか用いていない．また，11 都市で認識テストを行ったが，ほとんどの都市で高

い認識率を示す事が出来なかったと報告している．

視覚特徴量としては，SIFT や SURF のような局所的な領域から特徴量を抽出する方法がある [4]， [5]．これらの特徴量とフローベクトルを使いダイナミックなシーンの分類を行う手法が提案されている．しかし，これらの画像特徴量は我々が対象とするような不特定の撮影者によって撮影された映像においては照明変動やカメラの視点が異なるため適していない．

一方で，ビデオを使った撮影位置特定の手法も数多く報告されている．我々は，ビデオから時空間特徴量を抽出することで映像のローカライズを行う手法を提案した [6]．この手法では，映像を Temporal Height Image (THI) という建物の高さ情報を使った時系列画像に変換し，THI から局所特徴量を抽出することで，車載映像の撮影位置の同定を行った．しかし，この手法では数キロオーダーの撮影位置に関する初期情報が必要となり，グローバルな撮影位置推定を行うことには適していない．今回提案する手法では，このような初期情報を必要としないグローバルな撮影位置推定手法を実現する．

3. 提案手法

本研究では，Web の動画投稿サイトにアップロードされている撮影場所が不明な車載映像がクエリとして与えられた際に，グローバルな撮影位置を特定する手法を提案する．この時，照明変動やカメラの視点の問題があるが，このような問題に対してロバストなマッチングを実現する．提案手法は学習と検索の 2 ステップで構成されている．手法の概要を図 1 に示す．

学習では，Google Street View [7] から全方位画像を取得し，建物の正面を向いて撮影したように変換する．本手法では，世界中の場所を推定対象としているが，全世界すべてのストリートビュー画像を学習することは現実的ではない．そのため，本手法では世界の各都市を構成する「代表的なパターン」の抽出を行い学習を行う．そして，この代表的なパターンの出現頻度によって撮影された都市の判別を行う．この代表的なパターンとは，NY の Fire escapes や京都の木目の外壁のように他の国の都市では出現頻度が低い特徴的なパターンを指す．このようなパターンを見つけるため，本手法では画像を 80x80pixel のパッチ画像に分割し，クラスタリングを行う．このクラスタリングには，予めパッチ画像から抽出した，Histogram of Gradient (HOG) という局所領域における輝度の勾配強度情報を使用する．(図 2) HOG は色情報の影響を受けないため，車載映像のようにカメラ毎に照明環境が異なる場合でも照明変動に影響を受けない頑健な特徴量の抽出が可能となる．そして得られた代表的なパッチを使い機械学習を行う．今回は機械学習に Support Vector Machine (SVM) を用いた．また，今回の学習では one vs rest を採用しており，注目する都市

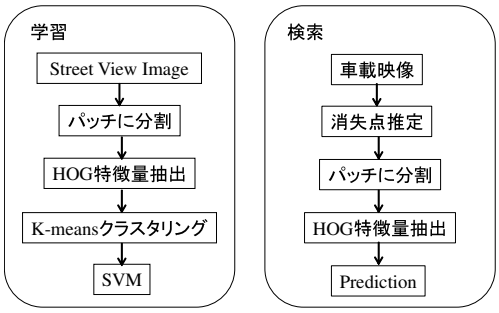


図 1 手法概要

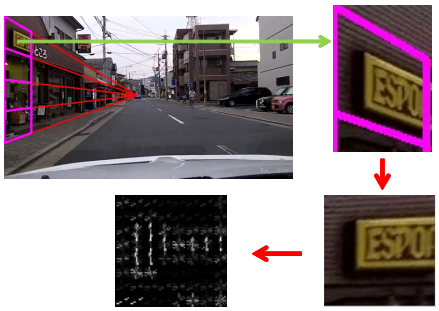


図 2 HOG 特徴量の抽出

かそれ以外かの 2 クラスの判定を行う．

これに対して検索では，動画投稿サイトにアップロードされている撮影位置不明の車載映像を使用する．学習と同様に各フレームをパッチに分割した後特徴量を抽出し，予め学習させた SVM から検索を行う．この時，1 フレームから複数のパッチ画像が生成されるが，クラスタリングを行っていないため，ノイズとなり得るデータも含んでいる．そのため，撮影位置の判定に分類された枚数の合計値を用いると誤認識を引き起こす可能性がある．このような問題を回避するため，提案手法では回帰と，そのとき得られる超平面からの距離に重みを付けた値を判定に用いた．またこの時，車載映像はカメラ毎に視点が異なるため，切り出されるパッチが異なるという問題が発生する．そのため本手法では，各フレームで消失点を推定し，パッチが建物の正面を向くように変換し検索を行った．

4. 都市の景観の学習

本手法は Web などにアップされた，付随情報のない車載映像の撮影位置の特定を目指す．このような映像は，撮影したドライバーごとに撮影環境が異なるため，光学的・幾何的な問題が発生する．本手法では，これらの問題に影響を受けにくいロバストな学習方法を提案する．

4.1 学習パッチ画像の生成方法

本手法では，Google Street View から各都市の全方位画像をダウンロードし，進行方向の両横の画像を切り出し使用する．それぞれの画像を 80x80 のパッチ画像をに分割し，各都市で特徴的に出現するパターンの抽出を行う．この

ようなパッチ画像を使った画像の検索方法として Saurabh Singh らが提案した手法がある [8]．彼らは，ミドルレベルの領域において，オブジェクト毎に代表的なパッチを見つける事で画像検索を実現した．我々はこの手法を都市の認識に拡張する．具体的には，都市の景観を撮影した画像においても，オブジェクトのようにあるカテゴリーに共通して出現する「代表的なパッチ」が存在する．我々はこのようなパッチを都市ごとに見つけることで高精度な学習・検索の実現を目指す．

4.2 照明変化に頑健な特徴量抽出

Web 上の車載映像は映像毎に撮影した日時，時間帯が異なるため照明条件が異なる．例えば，早朝に撮影された映像は朝日の影響を受け，曇りの日に撮影された映像はトーンが下がる．この時，SIFT や SURF といった特徴量を用いた場合，学習に使用した Google Street View 画像との間に色の違いが生じるため，頑健なマッチングが行えない．本手法ではこのような照明条件が異なる場合でもロバストなマッチングを実現するため，Higtogram of Orientated Gaussian (HOG) 特徴量を用いた [9]．HOG 特徴量は局所領域の輝度勾配に依存しているため，照明条件が異なる場合にでも頑健な特徴量の抽出が可能である．

4.3 クラスタリングによる代表的な特徴量の抽出

クラスタリングには k-means アルゴリズムを使用した [10]．クラスタリングすることで，他の都市では出現しない代表的なパッチを見つけることができる．本手法では，クラスタリングを複数回行うことで，より代表的なパッチの発見に努めた．クラスタリングには，識別したい都市のパッチ画像 25000 枚を Positive 画像に，残りの 2 都市のパッチ画像 50000 枚を Negative 画像に設定した．このとき，クラスタリング後のクラスタ内に Negative 画像が一定の割合以上存在した場合，このクラスタ内に含まれる Positive 画像にはその都市のみを示す表現能力が低いとみなし排除する．また，数枚でクラスタを形成している画像も同様に排除した．そして，このクラスタリング処理を複数回行い，最終的に残った都市の表現能力が高いパッチ画像のみを SVM に使用した．今回の実験では 5 回のクラスタリング後に，京都 5929 枚，NY5367 枚，パリ 4732 枚のパッチ画像を取得した．本手法では one vs rest を採用しており，それぞれの都市で SVM を生成した．

4.4 Support Vector Machine による学習

提案手法における学習には，Support Vector Machine (SVM) を使用した．SVM は高次元空間における 2 クラス分類手法として知られており，効率的な 2 クラス分類が可能としているため，これまでローカライゼーションに関する研究でも数多く使用されている．本手法においては，推定したい都市とそれ以外の都市の 2 クラス分類を実現した．今回の実験においては，クラスタリングで取得した 3 都市のパッチ画像の中から，検索したい都市のパッチ画像を Positive データ，それ以外の 2 都市の画像を Negative データとした．また，SVM の学習におけるカーネルには RBF カーネルを使用した．

5. 車載映像の検索手法

5.1 車載映像の幾何補正

車載映像は一般に広角を撮影するために，図 3 の (a) のように歪んでいる映像が多い．この場合，画像の左右の隅に歪みが生じ，得られるエッジも歪みの影響を受け，結果として検索結果の低下をもたらす．本手法では，このような歪による影響を抑えるため，歪みパラメータを自動推定し，補正した画像を用いる．映像一つにつき推定するパラメータは 1 つだけのため，歪みパラメータを変化させた際に検出される垂直方向のエッジの総数が最も多い値を最適なパラメータとして，全探索により推定する．提案手法により歪みを除去した画像を図 3 の (b) に示す．

(a) 歪みを含む画像　　(b) 歪みを除去した画像

図 3 レンズ歪を含む画像

また，Web にアップされた映像は，カメラがどの方向を向いて撮影されたか不明である．しかし，建物は通常道路

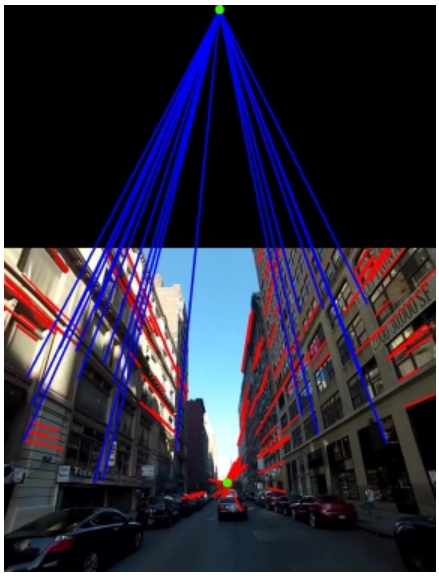


図 4 消失点の推定

に平行に建っていることが多いため，シーン中の消失点を見つけることで，校正することができる．提案手法では，建物のエッジを抽出し消失点の推定を行った．道路に対して平行に建っている建物の道路側の面には，建物に垂直なエッジと平行なエッジが多数含まれる．本手法では，垂直に伸びるエッジと，平行なエッジそれぞれを撮影画像上で求めて，2 つの消失点を推定する．このとき，歪みやエッジの検出ミスにより消失点が 1 点に定まるとは限らないため，複数の候補点の中から最も投票数が多かった点を消失点として採用した．図 4 に例を示す．こうして得られた 2 つの消失点を用いることで，車載カメラの回転方向が分かるため，これを用いて画像を地面に水平なカメラで撮影した画像に射影変換できる．この画像から検索に用いるパッチ画像を切り出すことができる．

こうして切り出した検索に用いるパッチのサイズと，学習に用いるパッチのサイズが一致していないと，検索がうまくいかない．そこで，実際にダウンロードした映像を調べてみたところ，車載カメラの画角はほとんどが広角であり，あまり差が大きくないことが分かった．また，スケールの変化について実験的に影響を調査したところ，学習で用いるパッチ自体にスケールのゆらぎがあるため，多少の変化では大きな影響がないことが確認できた．そこで，今回は 1 種類のスケールのみで検索を行った．将来的には，手法をより一般化するため，何種類かの大きさのパッチサイズを用意することが考えられる．

5.2 車載映像の SVM による識別

提案手法では映像から切り出された全てのフレームを検索および評価に用いる．まず，前節に述べた手法でそれぞれのフレームを水平カメラで撮影した画像に変換し，さらにその画像から，建物を正面から見た画像に射影変換し，パッチの切り出しを行う．このとき，建物位置が画像上のどこにあるのかを検出している訳ではないため，認識には画像全体からパッチ画像を切り出すしかなく，そのため空や道路のアスファルト面のように建物以外の領域もパッチとして切りだされ，誤認識を引き起こす．この対策のため本手法では，グラフカットを使用し，空と路面領域を一様にラベリングすることで除外し，これらの領域を含まない部分からのみパッチの切り出しを行った．各パッチの識別結果を，各フレームごとで統合することで，フレーム毎の識別を行った．各パッチの識別は，学習フェーズで作成した SVN 識別器を用いて行った．

本手法では入力に用いる車載映像から取得したパッチ画像は，識別前の段階で取捨選択することなく用いた．そのため、パッチの識別結果を統合する際，単純な投票により行おうとすると，どの都市でも共通して現れるような特徴を持ったパッチが多く存在すると，誤認識を引き起こす可能性がある．そこで本手法では，これらの影響を考慮し，検索に用いたパッチに重み付けて投票を行った．SVM 識別器による識別では，入力データが 2 つのクラスのどちらに含まれるかだけではなく，超平面からの距離も得られる．その距離が大きいほどクラスに属する可能性が高いと考えられるため，距離に比例した重みが与えられるようにして統合を行い，フレームごとの識別を行った．さらに全てのフレームごとの識別結果を用いて投票により，映像の識別を行った．

6. 実験

実験では，3 都市 (京都，NY，パリ) の認識を行った．学習は各都市から 10000 枚ずつ Google Street View [7] の全方位画像を取得し，建物の正面方向を向いた画像を取得した．そして，それぞれの画像を 80x80 のパッチ画像に分割した．この時，大量にパッチ画像が生成されるが，今回は都市ごとにランダムな 25000 枚のパッチ画像をサンプリングした．得られた合計 75000 枚のパッチ画像をクラスタリングに使用した．図 13 に学習に用いたクラスタリング後のパッチ画像の一部を示す．検索には，一般ユーザが実際に YouTube [11] にアップロードした車載映像を使用した．これらの映像は，任意の環境で撮影されており，カメラの歪みや車速，フレームレートは不明である．今回の実験において，撮影位置不明な映像として，「京都」,「ニューヨーク」,「パリ」の 3 ヶ所で撮影された映像を 1 本ずつ用いた．それぞれの映像の再生時間は 10〜30 分だが，今回はその中の連続した 150 フレームを使用した．画像サイズは 640x360pixel のものを用い，パッチは 80x80pixel とした．この時，得られたパッチ画像から HOG 特徴量を抽出するが，提案手法では 10x10pixel を 1 セルとし，さらに，1 セルを 1 ブロックとして定義し，輝度勾配は 31 方向に量子化した．また，このブロックの中から，L*a*b 色空間における a と b の値を足し，合計で 2112 次元のベクトルを抽出した．また，学習画像に使用したパッチ画像に交差検定法を用いて学習器の識別性能を計算したところ，3 都市すべての学習器において 90%以上を示した．

6.1 京都

実験結果を図 5 に示す．図 5 は，京都学習器に京都の映像を入力として与えた場合の結果である．図中の赤いグラフが京都と認識されたときの重みを示している．これより，1 シーケンス全体で京都らしい映像と識別されていることが分かる．このとき，どのようなパッチ画像が京都として識別されたのかを図 6 に示す．図 6 で，赤い部分は京都と識別されたパッチを，青い領域はそれ以外の場所で撮影されたパッチを示している．また，赤の色が濃くなるほど，京都と判別された重みが重いパッチを示し，青が濃い部分はそれ以外の都市で撮影された重みが大きいことを表す．左右の建物の領域において赤色が顕著になっていることか

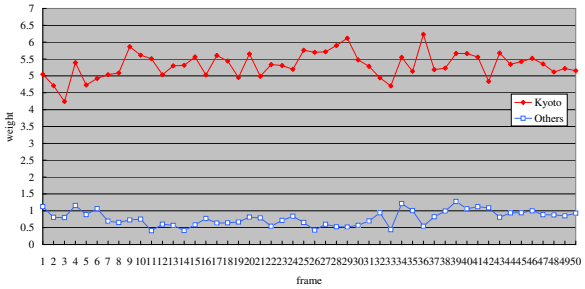


図 5 京都の映像

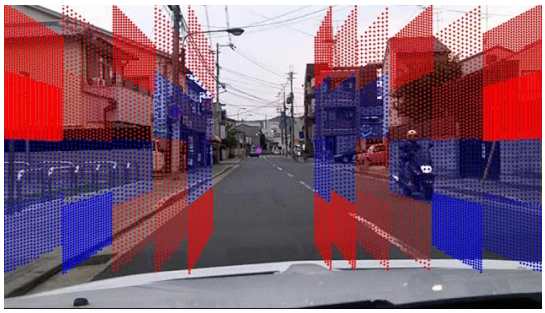
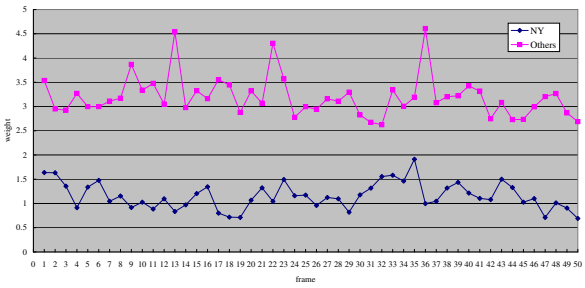
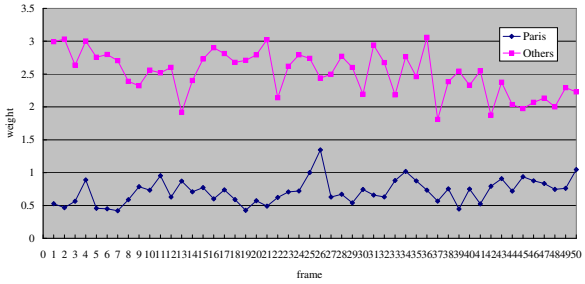


図 6 京都の映像



(a) ニューヨーク識別器



(b) パリ識別器

図 7 京都の映像を京都以外の識別器にかけた結果

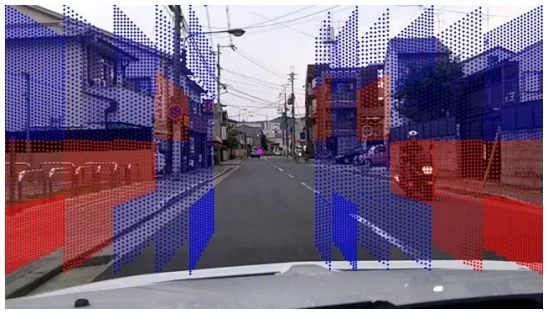


図 8 京都の映像をニューヨークの識別木にかけた結果

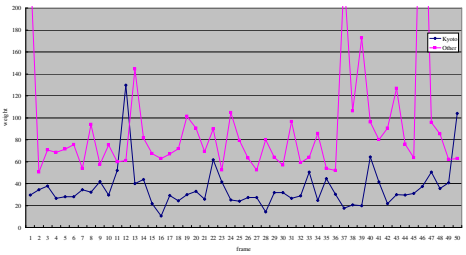


図 9 ニューヨークの車載映像を京都識別器にかけた結果

ら，このような街並みが京都らしさを表していると考えられる．しかし，石垣や歩道，一部の空の領域においては，誤認識を引き起こしていることが確認できることから，あいまいな領域や様々な場所に出現するものは誤認識を引き起こす原因となると判断出来る．次に，図 7 の (a) と (b) では，識別にニューヨークとパリの学習器を用いた．この場合，それ以外の場所で撮影されたと判断されれば正解となる．図 7 より，京都の映像をその他の学習器に入れた場合も正しく京都以外であると識別出来ている．また，図 8 は，ニューヨーク識別器における識別結果を示す．赤い領域はニューヨークで撮影されたと判断したパッチを示し，青い領域はその他の場所で撮影されたと判断したパッチを示す．図 8 では図 6 とは対象的に，歩道や石垣においてはニューヨークで撮影されたと誤認識している．しかし，建物の領域においてはニューヨーク以外の場所で撮影されたと判断していることが分かる．

6.2 ニューヨーク

次に，京都以外の都市の映像をそれぞれの識別器に与えたときに，京都もしくは，それ以外の都市で撮影されたのかの識別を行う実験を行った．まず，ニューヨークの映像を識別に用いた．図 9 に，ニューヨーク学習器にニューヨークの映像を与えた場合の結果を示す．図 9 において，ほとんどのフレームで赤色のグラフが高い値を示している．つまり，この映像が京都以外の都市で撮影されたことを示していることが分かる．また，図 10 にこのときのパッチの分類結果の一部を示す．木の領域においては，誤認識を引き起こしているが，建物の壁面においては，京都以外で撮影されたことを示しており，正しく評価できていることが分かる．

6.3 パリ

最後に，京都識別器にパリで撮影された映像を与えた．図 11 に示すとおり，この映像においても正しく識別出来ていることが確認できる．また，このときのパッチの識別結果を図 12 に示す．図 12 で，赤色の領域は京都で撮影されたと判断した箇所を示し，青色の領域は，京都以外の都

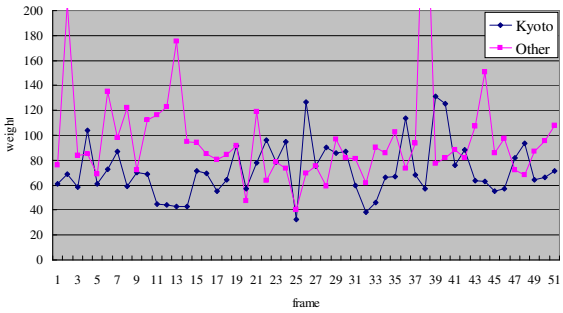


図 11 パリの車載映像を京都識別器にかけた結果

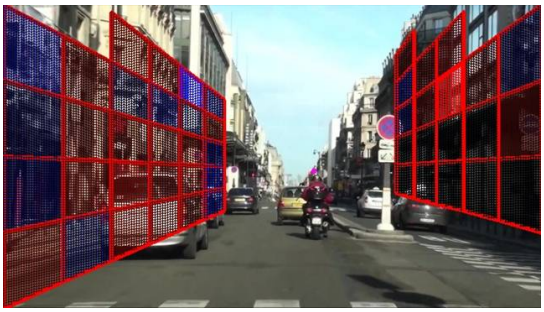


図 12 パリの車載映像

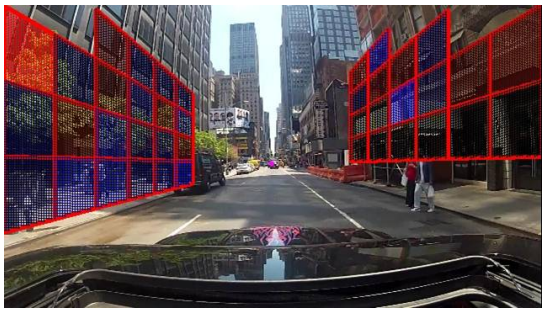


図 10 ニューヨークの車載映像

市で撮影されたと判断したことを示す. 図 12 において，それ以外で撮影されたと識別されたパッチに大きな重みが与えられている．他車の領域などでは一部誤認識を引き起こしているが，建物の領域においては京都以外の場所で撮影されたと識別される．

6.4 考察

今回の実験では入力映像において生成される全てのパッチ画像を認識に使用した．しかし，この中には空や路面，壁面など，どこで撮影されたのかを容易に判別出来ないものも含まれており，このような誤認識引き起こすパッチ画像の影響で推定の値を下げたと考えられる．また，提案手法では映像から切り出されたパッチを建物の正面を向いて撮影されたように変換した．しかしこの場合，交差点や空き地のような建物の無い場所においては始めから建物の側面が写ってしまう．さらに，標識やその他の車など進行方向に対して平行に写っているオブジェクトも同様に間違った特徴量の抽出を行ってしまう．その場合，異なったパッチの変換を引き起こしてしまい，結果として，学習には無かった視点からのパッチ画像を生成してしまい，これも誤認識を引き起こす原因となることが考えれる．さらに，消失点の推定に失敗したフレームにおいては，異なった無限遠を推定してしまい，誤ったパッチの生成をするため，同様に誤認識を引き起こすことが確認出来た．

7. まとめ

提案手法では撮影位置不明車載映像のグローバルな位置

推定を示した．実験において 3 都市での認識を行った．今後は，徐々に都市を増やし，マルチクラスでの識別を行う予定である．また，本手法のように各都市における代表的なパターンを見つけることが出来れば，自動運転の支援などに応用が可能と考えられる．

(a) 京都

(b) ニューヨーク

(c) パリ

図 13 データベースに使用したパッチ画像の一部

参考文献

[1] K. A. Ehinger A. Oliva J. Xiao, J. Hays and A. Torralba, “Sun database: Large-scale scene recognition from abbey to zoo,” in *Proceedings of 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR2010)*, 2010.

[2] Evangelos Kalogerakis, Olga Vesselova, James Hays, Alexei A. Efros, and Aaron Hertzmann, “Image sequence geolocation with human travel priors,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV ’09)*, 2009.

[3] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros, “What makes paris look like paris?,” *ACM Transactions on Graphics (SIGGRAPH)*, vol. 31, no. 4, 2012.

[4] David G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, “Speeded-up robust features (surf),” *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, June 2008.

[6] 福元和真, 川崎洋, 小野晋太郎, 子安大士, and 池内克史, “自車位置推定のための複数車載カメラ映像の効率的な時空間マッチング手法,” in **第 *11* 回 *ITS* シンポジウム *2012***, 2012.

[7] ,” Google Street View : http://maps.google.com/help/maps/streetview/.

[8] Saurabh Singh, Abhinav Gupta, and Alexei A. Efros, “Unsupervised discovery of mid-level discriminative patches,” in *European Conference on Computer Vision*, 2012.

[9] Navneet Dalal and Bill Triggs, “Histograms of oriented gradients for human detection,” in *International Conference on Computer Vision & Pattern Recognition*, Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, Eds., INRIA Rhône-Alpes, ZIRST-655, av. de l’Europe, Montbonnot-38334, June 2005, vol. 2, pp. 886–893.

[10] J. A. Hartigan and M. A. Wong, “A K-means clustering algorithm,” *Applied Statistics*, vol. 28, pp. 100–108, 1979.

[11] ,” YouTube : http://www.youtube.com/.