Vol.2015-CVIM-196 No.2
2015/3/6

畳込みニューラルネットワークを用いたシーン変化の検出

櫻田 健[1,a)]　岡谷 貴之[1,b)]

概要：本稿はグリッド特徴を利用して各時刻一枚ずつの画像ペアからシーンの変化を検出する手法を提案する．自動のシーン変化の検出は都市管理や，災害の復旧，復興，減災において有用である．本研究の背景として，車載カメラの画像を用いて津波被災地の街並みの変化を可視化する目的がある．既存のシーン変化の検出手法は，街並みの 3 次元モデルや，異なる時刻のデータ間におけるピクセルレベルの位置合わせを必要とした．そのため，正確な 3 次元モデルが得られない本研究のようなケースには適用することが困難である．さらに，広域の 3 次元モデルの復元や，ピクセルレベルでの位置合わせは膨大な計算量を必要とする．これらの問題を解決するために，本研究では畳込みニューラルネットワークとスーパーピクセルセグメンテーションを統合した新しいシーン変化の検出方法を提案する．提案手法はシーンの 3 次元モデルやピクセルレベルでの位置合わせを必要とせず，計算量を大幅に削減できるため，津波被災地全体のシーン変化の推定を可能とする．

1. 序論

本研究では車載カメラで撮影した各時刻 1 枚ずつの画像ペアからシーンの変化を検出する問題を考える．自動でシーンの変化を検出することにより，例えば，（事前に同じ場所を撮影していれば）車を走らせるだけで災害による被害を迅速に把握することができる．さらに，その後も定期的に撮影を行うことで，復旧および復興の進捗を把握することができる．

後者の目的のために我々は，2011 年の東北地方太平洋沖地震の津波被災地を，震災の約 1 ヶ月後から定期的に撮影している．撮影地域の全長は約 400km で，これまでに撮影した画像データは約 40TB である．図 1 に 3 ヶ月おきに同じ場所を撮影した画像の例を示す．

車載カメラの画像を用いてシーンの時間変化を検出するためにはいくつかの問題点がある．主な要因として各時刻間のカメラ視点や照明条件，撮影条件，雲などの空模様，道路の汚れによる地面の見えの違いが挙げられる．例えば，計測車両は毎回道路の違う場所を走行し，シャッターを切るタイミングも異なるため，各時刻の画像間で撮影視点が大きく異なることがある．また屋外で撮影するため天候や時間帯によって照明条件も大きく異る．これらの問題を解決するロバストな新しい時間変化の検出方法を開発する必要がある．

[1] 東北大学大学院情報科学研究科
宮城県仙台市青葉区荒巻字青葉 6-6-01
[a)] sakurada@vision.is.tohoku.ac.jp
[b)] okatani@vision.is.tohoku.ac.jp

図 1　同一シーンを定期的に撮影した画像

シーン変化を検出する既存手法はシーンの 3 次元モデルや，異時刻間のピクセルレベルの位置合わせを必要とする．しかし，推定対象領域が広域の場合には，シーンの 3 次復元や正確な位置合わせの計算量が膨大となる．

そこで，本研究では畳込みニューラルネットワーク (convolutional neural network, 以下 CNN) の特徴量とスーパーピクセルセグメンテーションを統合することで，シーンの 3 次元モデルとピクセルレベルの位置合わせを必要としない新しいシーン変化の検出方法を提案する．ピクセルレベルの位置合わせをせず変化を検出するために，提案手法で

は入力画像をグリッド状に分割し，各時刻の画像のグリッド特徴を比較することで変化を検出する．さらに，より詳細な変化を検出するために，各グリッドの比較結果をスーパーピクセルに投影する．

提案手法では畳込みニューラルネットワーク (CNN) [1] のプーリング層の特徴をグリッド特徴として利用する．一般的に，物体認識の手法では画像を各クラスに分類するために全結合層の情報を利用するが，提案手法では画像空間の位置情報を有するプーリング層の特徴を利用する．

最近の研究では，CNN の上の層ほどより抽象的で，画像空間中の広い領域の情報を有することが明らかにされている [2], [3]．そして，上の層ほど解像度は低下する．つまり，上のプーリング層は抽象度の高いオブジェクトを，逆に，下のプーリング層はエッジやテクスチャなどの低いレベルの画像特徴を認識するのに有効であると考えられる．本研究ではシーン変化におけるこの各プーリング層の特性も評価する．

2. 関連研究

画像を用いた都市のモデリングの研究が盛んになるにつれ，シーンの変化を検出する研究も活発に行われてきた [4], [5], [6], [7], [8], [9]．これまでにも 2 次元的変化，3 次元的変化を検出する手法がそれぞれ提案されている．標準的な 2 次元的変化の検出方法では，事前に得られたトレーニング画像からシーンのアピアランスモデルを学習し，クエリ画像中に大きな変化が起こったか否かを推定する [10], [11] ．また，多くの 3 次元的シーン変化の検出方法でも，定常状態のシーンモデルを構築し新たに撮影された画像と比較するといった方法を採用している．

地上を撮影した上空視点画像を対象とした研究 [10] では，20 枚から 40 枚の画像を用いてボクセルベースのシーンのアピアランスモデルを学習する．その手法を Crispell らは記憶容量を最小化するように改良した [12]．Ibrahim と David は画像空間中の線分の出現あるいは消失を推定することでシーン変化を検出する手法を提案した [13]．これらの手法は十分な枚数の画像から対象シーンのアピアランスモデルを構築している．しかし，本研究では車に搭載したカメラで走行しながらシーンを撮影するため十分な枚数の画像を得られない．つまり，これらの手法は航空・衛星画像に適した手法であり，車載カメラの画像を対象とする本研究には適していない．

また，Schindler らは何十年にも渡って都市を撮影した大量の画像から，それぞれの建物がいつ建てらたかなどの時間変化を推定する手法を提案した [14]．この手法では異なる時刻の画像間で十分な画像特徴の対応が取れることを仮定しており，全ての時刻の画像をまとめて 3 次元復元している．そのため，シーンが大きく変化している状況には適用することが難しい．さらに，大量の画像を利用しているにも関わらず，シーン変化を画像特徴の疎な点群でのみ表現されている．

アプリケーションの観点から Schindler の研究に類似したものとして, Matzen らの研究が挙げられる [15]．Matzen らはシーンの 3 次元的変化はないと仮定し，2 次点的変化のみに焦点を当てた手法を提案した．インターネットから街並みを撮影した画像を収集し，その画像セットから建物などの模様がいつどのように変化したかを推定する．さらに，全画像のタイムスタンプの整合性を最適化することで，入力された画像のタイムスタンプの誤差を修正することができる．この研究は街並みの 3 次元構造が変化していないことを仮定しているため，Schindler らの研究と同様に，複数の時刻の画像をまとめて 3 次元復元している．ゆえに，シーンの密な 3 次元構造が得られず，さらに，3 次元構造が大きく変化するシーンを対象とした本研究に適用することは困難である．

さらに，一方の時刻についてシーンの 3 次元モデルを他のセンサーあるいは方法から取得してシーン変化を検出する方法が提案されている．Huertas らの手法では，建物の 3 次元モデルが得られていると仮定し，上空視点画像から抽出した輪郭線を，3 次元モデルの地面への投影結果と比較して変化を検出する [16]．最近の Taneja らの研究も同様である [17], [18]．Taneja らの手法は車載カメラで撮影した画像から時間変化を検出するため，アプリケーションの観点から我々の研究と近い．しかし，Taneja らの研究は大都市の 3 次元モデルを低コストで更新することを目的としており，対象シーンの密な 3 次元モデルが得られていることを仮定している．

提案手法は上記全てと異なる問題を対象としている．既存研究と異なり，シーンの 3 次元モデルは必要とせず，さらに，画像間のピクセルレベルの位置合わせも必要としない．そのため，車載カメラで撮影した画像から比較的高速に時間変化を検出することが可能となる．さらに，津波被災地のような異なる時刻の画像間で十分な画像特徴の対応が得られないような環境でもシーンの変化を検出することができる．

3. グリッド特徴を利用したシーン変化の検出

本稿ではグリッドベースのシーン変化の推定方法を提案する．提案手法は以下の 3 つの情報を利用する．

(i) グリッド特徴
(ii) スーパーピクセルセグメンテーション
(iii) Geometric Context (空，地面)

提案手法のフローチャートを図 2 に示す．本研究ではカメラを搭載した車両で走行しながら街並みを撮影するため撮影視点は毎回異なる．提案手法では撮影視点の違いの影響を小さくするために，シーン変化をグリッドの特徴量を用いて検出する．グリッド特徴量を用いて大まかなシーン変

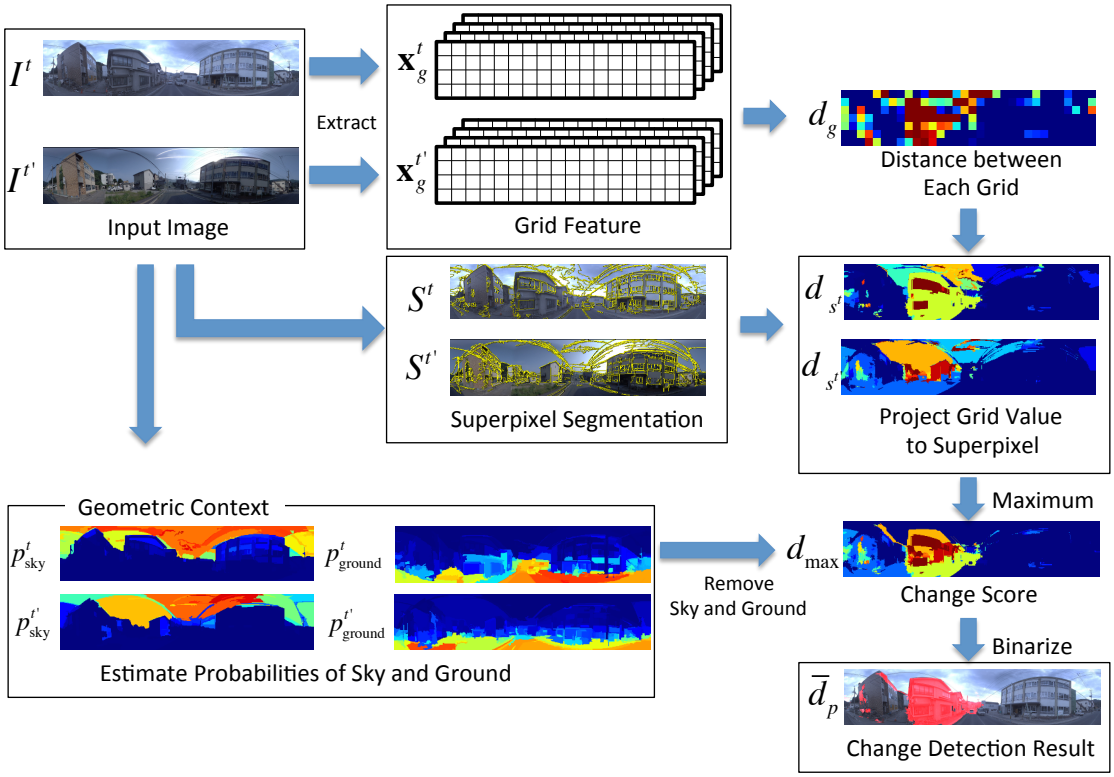


図 2 グリッド特徴を利用したシーン変化検出のフローチャート

化を推定し，その後，スーパーピクセルセグメンテーションとを利用してより正確に時間変化を推定する．

本研究ではオブジェクト（例，建物，車）をシーン変化の推定対象とし，照明条件や撮影条件の違いによる低レベルの見た目の変化は推定対象に含まない．このような抽象度の高い変化を検出するために提案手法では CNN のプーリング層をグリッド特徴として利用する．

さらに，空模様（例，雲）や路面状況（例，道路上のゴミ）の違いを推定対象から除外するために，提案手法では Geometric Context [19] を利用する．Geometric Context は幾何学的な特徴のみを利用するため，照明条件や撮影条件に影響されにくいセグメンテーション手法である．

(i) グリッド特徴

まず，入力画像 I^t をグリッド上に分割し，それぞれのグリッド $(g = 1, ..., N_g)$ から特徴量 ${\mathbf{x}_g}^t$ を抽出する．提案手法は CNN を特徴抽出器として利用する．評価実験では CNN を用いた用いた結果とベースラインの SIFT[20], [21], [22] とローカルパッチを利用した結果を比較した．その結果，CNN のプーリング層の特徴がシーン変化の検出に対して有効であることを確認した．グリッド特徴の詳細については 4 節で説明する．正規化した各時刻のグリッドの特徴量 ${\mathbf{x}_g}^t$ $(|{\mathbf{x}_g}^t| = 1)$ を用いて，異なる時刻間の各グリッドの非類似度 d_g を式 (1) により計算する．

$$d_g = |{\mathbf{x}_g}^t - {\mathbf{x}_g}^{t'}|. \tag{1}$$

そして，d_g を入力画像 I, I' に投影し，各ピクセルにおける非類似度 $d_p(p = 1, ..., N_p)$ を求める $(N_p$: ピセル数).

(ii) スーパーピクセルセグメンテーション

スーパーピクセルセグメンテーション d_g はグリッド特徴量から推定した被類似度であり解像度が荒いため，スーパーピクセルセグメンテーションを用いてより正確に変化した領域を推定する．入力画像 I^t に対してスーパーピクセルセグメンテーションを行い，それぞれのスーパーピクセルの集合を S^t とする．各スーパーピクセルの非類似度 $s^t \in S^t$ はスーパーピクセル s^t 内のピクセルの全平均とする．

$$d_{s^t} = \frac{1}{|s^t|} \sum_{p \in s^t} d_p. \tag{2}$$

さらに，$d_{s^t}, d_{s^{t'}}$ の最大値を $d_{\max}$ とする．

$$d_{\max} = \max(d_{s^t}, d_{s^{t'}}) \tag{3}$$

(iii) Geometric Context

また，本研究では空と地面は時間変化の推定対象から除外するために Geometric Context を利用した．入力画像 I^t において Geometric Context から得られる各ピクセルの空と地面の確率を $({p_{\text{sky}}}^t, {p_{\text{ground}}}^t)$ とすると，空と地面を除外した非類似度は式 (4) のようになる．

$$\overline{d_p} = \begin{cases} 0 & (({p_{\text{sky}}}^t > a) \wedge ({p_{\text{sky}}}^{t'} > a)) \\ & \vee (({p_{\text{ground}}}^t > b) \wedge ({p_{\text{ground}}}^{t'} > b))) \\ d_{\max} & (\text{otherwise}) \end{cases} \tag{4}$$

$a = t_{\text{sky}}$ と $b = t_{\text{ground}}$ はそれぞれ $0 \leq t_{\text{sky}}, t_{\text{ground}} \leq 1$ の定数とする．

4. グリッド特徴の選択

提案手法ではシーン変化を検出するために画像のグリッド特徴を利用する．典型的なグリッド特徴として SIFT やローカルパッチが多くのタスクに利用されてきた．これまでの物体認識のタスクでは，Dense SIFT の Bag-of-Visual Words (BOW) や Fisher Vector を利用した手法が高い認識精度を実現している [20], [21], [22], [23], [24], [25].

近年，畳込みニューラルネットワーク (CNN) を利用した手法がその認識精度を更新している．特徴抽出器を人工的に設計している SIFT とは対照的に，CNN の特徴抽出器は大量の画像から Network のパラメーターを学習することで得られる．CNN の構造は人間が物体を認識する仕組みと類似していることが最近の研究で明らかになっている [26]. つまり, シーンの変化などのより抽象的なレベルの分類に有効な可能性が高い. ただし, CNN の仕組みには未解明な部分も多い. そのため, 本研究ではシーンの変化が CNN のどのレイヤーに対応しているかも調査する.

シーン変化の検出に CNN を適用するために CNN のプーリングレイヤーをグリッド特徴として利用する．本研究では畳み込み後に空間解像度が保たれるようにパディングを施した CNN モデルを用いる (例， [27])．各グリッドの特徴量はスケールを 1 に正規化する．また，本研究で利用する CNN の特徴量は Rectified Linear Units (ReLUs) [1] により全て非負の値となる．そのため，異なる時刻間の各グリッドの非類似度は $d_i \in \mathbf{d}_g$ は $0 \leq d_i \leq \sqrt{2}$ の範囲の値を取る．

次節で Dense-SIFT [21], [22] とローカルパッチをベースラインとして CNN の特徴を評価する．本研究の Dense-SIFT では，グリッドサイズをベースとしてマルチスケールで特徴量を計算し，それぞれのスケールの特徴量を連結することで各グリッドの特徴量とする．また，ローカルパッチではグレースケール画像を利用する．

5. 評価実験

本節で提案した Change Detection の有効性を確認するために，Panoramic Change Detection データセットを用意して精度評価を行った．具体的には，まず，津波被災地で異なる時期に撮影した全方位パノラマ画像を用意する．そして，一枚の全方位パノラマ画像に対し異なる時期の画像セットから 3 次元空間的に最近傍の画像を一枚選ぶ．そして，この 2 枚の画像のみを用いて時間変化を検出する．

全方位カメラと GPS を搭載した車両で走行しながら約 2m 間隔で撮影したため，同一シーンを撮影した画像間でもその撮影視点は毎回異なる．さらに，天候の違いにより各時期の照明条件も大きく異る．このような視点の違いや照明条件の違いにより，本データセットの時間変化は人の目で見ても瞬時に判別するのは容易ではない．変化検出の ground-truth は 1 ペアあたり平均 15 分も要して手動で作成した．1 つの街だけで数千〜数万の画像ペアがあるため，街全体のシーンの変化を推定するためには自動でシーン変化を検出する方法が必要不可欠である．

5.1 Panoramic Change Detection データセット

シーン変化検出の精度評価用データセットとして，複数の都市のデータセットから各時刻 1 枚ずつのパノラマ画像を 20 ペア用意した．図 3 にデータセットの例を示す．本実験では精度評価のため，ペアとなる異時刻の最近傍パノラマ画像を手動で選択した．シーン変化検出のグランドトゥルースも手動で作成した．

本研究では，2 次元および 3 次元両方のシーン変化を検出対象とする．ただし，照明などの撮影条件の違いによる変化，および空と地面は検出対象から除外する．例えば，建物や車，瓦礫などが出現または消失した場合は検出対象とするが，空や地面の模様，鏡面反射などによる見えの変化は無変化として扱う．

5.2 パラメーター設定

提案手法では次のパラメーターを設定する：(1) シーン変化の推定結果を二値化するグリッド特徴間の距離の閾値 t_{dist}，(2) 空と地面を検出するための Geometric Context の確率の閾値 (t_{sky}, t_{ground}), (3) スパーピクセルセグメンテーションのパラメーター.

(1) グリッド特徴間の距離の閾値 t_{dist} は Panoramic Change Detection データセットを用いた評価で最も高い F_1 スコアを達成した値とする．表 1 に全特徴の閾値を示す．CNN のプーリング層と Dense-SIFT をグリッド特徴として利用する場合は，特徴ベクトルの全ての要素は非負の値を取るため，正規化された各グリッドの特徴間の距離 $d_i \in \mathbf{d}_g$ は $0 \leq d_i \leq \sqrt{2}$ の範囲の値を取る．グレースケールのローカルパッチの場合，d_i は $0 \leq d_i \leq 2$ の範囲の値を取る．CNN のプーリング 3，4，5 層とグレースケールのローカルパッチの閾値は，それぞれの値域のほぼ中央値である．

(2) 空と地面を検出するための閾値は全ての実験で同一とする ($t_{\text{sky}} = 0.2$, $t_{\text{ground}} = 0.8$).

(3) 本実験ではスーパーピクセルセグメンテーションの手法として Felsenszwalb の efficient graph based image segmentation[28] を用いる．スーパーピクセルセグメンテーションのパラメーター（ガウシアンカーネルのスケールと直径，各コンポーネントの最小サイズ）は全ての実験で同一とする．

CNN のモデルとして，物体認識のタスクで state of the art を達成している VGG モデルを採用する [27] ．VGG モデルは 19 の重み層を持ち，提案手法ではそのプーリング層を利用する．さらに，VGG モデルは 3×3 の畳み込みに

入力画像 (クエリ)　入力画像 (データベース)　グランドトゥルース

図 3　Panoramic Change Detection データセットの一例.

対し 1 ピクセルのパディングを行うため，畳み込み後も空間解像度が変わらない．つまり，入力画像と各層の特徴の空間的整合性が保たれる．

本実験では CNN の特徴を Dense-SIFT[20], [21], [22] とグレースケールのローカルパッチと比較する．評価のために，Dense-SIFT とローカルパッチのグリッドサイズはプーリング 5 層のグリッドサイズと等しく設定した．本実験の Dense-SIFT はグリッドごとに複数のスケールで特徴量を抽出し，それらを一列の特徴ベクトルとして連結する．つまり，Dense-SIFT の特徴の次元数は $128 \times 4 = 512$ となる．グレースケールのローカルパッチは 16×16 にリサイズして次元数を 256 とした．

5.3　考察

表 1 に Panoramic Change Detection データセットに対する各特徴の F_1 スコアを示す．プーリング 4，5 層の特徴が最も高い F_1 スコアを達成した．上位のプーリング層の方が下位のプーリング層よりも高い検出精度を達成している．プーリング 1 層の F_1 スコアはベースラインの Dense-SIFT やグレースケールのローカルパッチとほぼ同じスコアとなった．この実験結果から CNN の特徴がシーン変化の検出に有効であることを確認した．

各グリッドの特徴間の距離とシーン変化の推定結果を図 4 に示す．上から入力画像ペア，シーン変化のグランドトゥルース，各グリッドの特徴の距離を示している．図 4 から，上位層のプーリング層は物体のような抽象度の高いシーンの違いを認識できることが分かる． 一方，下位層のプーリング層はエッジのようなローレベルな画像特徴の違いを検出している．例えば，図 4 のプーリング 3 層の結果では，照明条件や視点の違いにより左の建物周辺で大きな誤検出が発生しているが，右の建物周辺の小さな変化は正しく検出できている．

Panoramic Change Detection データセットを用いてシーン変化を検出した例を図 5 に示す．上から順に，入力画像ペア，グランドトゥルース，最終検出結果，スーパーピクセルセグメンテーションの結果，プーリング 5 層の特徴を用いた場合の各グリッドの特徴間の距離，各グリッドの特徴間の距離をスーパーピクセルに投影した結果，Geometric Context を用いて推定した空と地面の確率を示している．

提案手法は車や瓦礫，取り壊された建物などのシーン変化を正しく検出できている．いくつかのケースでは Geometric Context が電線や電柱などの影響で空と地面を正しく推定できなかったり，地面と丈の低い物体（例，瓦礫，車）を区別できていない．このようなセグメンテーションによる誤差を除けば，提案手法は物体レベルのシーン変化を正しく認識できている．これらの結果から，上位のプーリング層はシーンの違いを識別する目的に有効であり，またスーパーピクセルセグメンテーションはプーリング層の解像度の低さを補えることが確認できた．

6.　結言

本稿は各時刻 1 枚ずつの画像ペアからシーンの変化を検出する統一的なフレームワークを提案した．提案手法はシーンの 3 次元モデルやピセルレベルの画像の位置合わせを必要としないため，従来手法と比較して計算量を大幅に削減できる．提案手法の精度を評価するため Panoramic Change Detection データセットを作成した．実験結果から提案手法は CNN の高い識別力とスーパーピクセルの正確なセグメンテーションを効果的に統合できることが確認できた．

さらに，シーン変化の検出に対する CNN の上位層と下位層の有効性の違いを評価した．その結果，推定したい対象の抽象度に応じて CNN の層のレベルを選択できることを確認した．物体などの抽象度の高いシーンの変化を識別したい場合は上位のプーリング層を，逆に，エッジなどのローレベルな画像特徴の違いを検出したい場合は下位のプーリング層を利用すればよい．

提案手法の推定精度を向上させ，広範囲のシーン変化を可視化するためには，(i)CNN の上位層と下位層の特徴の統合，(ii) 変化を検出した物体の統計的分類を行う必要がある．上位層は抽象度の高い物体を，下位層はローレベルな画像特徴を識別することができる．そのためこれらを統合することで変化検出の精度を向上できると考えられる．さらに，本研究の最終目標は津波被災地全体のシーン変化の可視化である．被災地全域の被害を把握するために，シーンの変化量を物体のカテゴリーごとに分類し統計量を計算する．

謝辞

本研究は JSPS 科研費 25135701,25280054 の助成を受けたものです。

表 1 各特徴のシーン変化検出の F_1 スコアとそのときの閾値 t_{dist}．下段は各特徴の次元数（行，列，特徴量）．入力画像の解像度はは 224×1024．本実験の CNN では 19 の重み層を持つ VGG モデルを採用 [27].

	pool-5	pool-4	pool-3	pool-2	pool-1	Dense-SIFT	Patch
F_1 score	0.722	0.722	0.688	0.629	0.592	0.592	0.599
Threshold	0.75	0.75	0.70	0.65	0.35	0.25	0.90
Feat Dim (y,x,d)	(7,32,512)	(14,64,512)	(28,128,256)	(56,256,128)	(112,512,64)	(7,32,512)	(7,32,256)

参考文献

[1] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Advances in Neural Information Processing Systems 25* (Pereira, F., Burges, C., Bottou, L. and Weinberger, K., eds.), Curran Associates, Inc., pp. 1097–1105 (online), available from ⟨http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf⟩ (2012).

[2] Bengio, Y.: Learning deep architectures for AI, *Foundations and trends® in Machine Learning*, Vol. 2, No. 1, pp. 1–127 (2009).

[3] Zeiler, M. D. and Fergus, R.: Visualizing and understanding convolutional networks, *ECCV*, Springer, pp. 818–833 (2014).

[4] Agarwal, S., Snavely, N., Simon, I., Seitz, S. M. and Szeliski, R.: Building Rome in a day, *ICCV*, pp. 72–79 (2009).

[5] Crandall, D., Owens, A., Snavely, N. and Huttenlocher, D.: Discrete-Continuous Optimization for Large-Scale Structure from Motion, *CVPR*, pp. 3001–3008 (2011).

[6] Pollefeys, M., Nistér, D., Frahm, J.-M., Akbarzadeh, A., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Kim, S.-J., Merrell, P., Salmi, C., Sinha, S., Talton, B., Wang, L., Yang, Q., Stewénius, H., Yang, R., Welch, G. and Towles, H.: Detailed Real-Time Urban 3D Reconstruction from Video, *IJCV*, Vol. 78, No. 2-3, pp. 143–167 (2008).

[7] Snavely, N., Seitz, S. M. and Szeliski, R.: Modeling the World from Internet Photo Collections, *IJCV*, Vol. 80, No. 2, pp. 189–210 (2007).

[8] Zhang, C., Wang, L. and Yang, R.: Semantic Segmentation of Urban Scenes Using Dense Depth Maps, *ECCV*, pp. 708–721 (2010).

[9] Zhang, G., Jia, J., Xiong, W., Wong, T.-T., Heng, P.-A. and Bao, H.: Moving Object Extraction with a Hand-held Camera, *ICCV*, pp. 1–8 (2007).

[10] Pollard, T. and Mundy, J. L.: Change Detection in a 3-d World, *CVPR*, pp. 1–6 (2007).

[11] Radke, R. J., Andra, S., Al-Kofahi, O. and Roysam, B.: Image Change Detection Algorithms: A Systematic Survey, *Transactions on Image Processing*, Vol. 14, No. 3, pp. 294–307 (2005).

[12] Crispell, D., Mundy, J. and Taubin, G.: A Variable-Resolution Probabilistic Three-Dimensional Model for Change Detection, *Geoscience and Remote Sensing*, Vol. 50, No. 2, pp. 489–500 (2012).

[13] Ibrahim Eden, D. C.: Using 3D Line Segments for Robust and Efficient Change Detection from Multiple Noisy Images, *ECCV*, pp. 172–185 (2008).

[14] Schindler, G. and Dellaert, F.: Probabilistic temporal inference on reconstructed 3D scenes, *CVPR*, pp. 1410–1417 (2010).

[15] Matzen, K. and Snavely, N.: Scene Chronology, *Proc. European Conf. on Computer Vision* (2014).

[16] Huertas, A. and Nevatia, R.: Detecting Changes in Aerial Views of Man-Made Structures, *ICCV*, pp. 73–80 (1998).

[17] Taneja, A., Ballan, L. and Pollefeys, M.: Image based detection of geometric changes in urban environments, *ICCV*, pp. 2336–2343 (2011).

[18] Taneja, A., Ballan, L. and Pollefeys, M.: City-Scale Change Detection in Cadastral 3D Models Using Images, *CVPR*, pp. 113–120 (2013).

[19] Derek Hoiem, Alexei A. Efros and Martial Hebert: Geometric context from a single image, *ICCV*, pp. 654–661 (2005).

[20] Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints, *IJCV*, Vol. 60, No. 2, pp. 91–110 (2004).

[21] Liu, C., Yuen, J., Torralba, A., Sivic, J. and Freeman, W. T.: Sift flow: Dense correspondence across different scenes, *ECCV*, Springer, pp. 28–42 (2008).

[22] Vedaldi, A. and Fulkerson, B.: Vlfeat: An Open and Portable Library of Computer Vision Algorithms, *International Conference on Multimedia*, MM '10, ACM, pp. 1469–1472 (2010).

[23] Yang, J., Jiang, Y.-G., Hauptmann, A. G. and Ngo, C.-W.: Evaluating Bag-of-visual-words Representations in Scene Classification, *International Workshop on Workshop on Multimedia Information Retrieval*, ACM, pp. 197–206 (2007).

[24] Perronnin, F. and Dance, C.: Fisher kernels on visual vocabularies for image categorization, *CVPR*, IEEE, pp. 1–8 (2007).

[25] Perronnin, F., Sánchez, J. and Mensink, T.: Improving the fisher kernel for large-scale image classification, *ECCV*, Springer, pp. 143–156 (2010).

[26] Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D. and DiCarlo, J. J.: Performance-optimized hierarchical models predict neural responses in higher visual cortex, *Proceedings of the National Academy of Sciences*, p. 201403112 (2014).

[27] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *CoRR*, Vol. abs/1409.1556 (online), available from ⟨http://arxiv.org/abs/1409.1556⟩ (2014).

[28] Felzenszwalb, P. F. and Huttenlocher, D. P.: Efficient graph-based image segmentation, *International Journal of Computer Vision*, Vol. 59, No. 2, pp. 167–181 (2004).

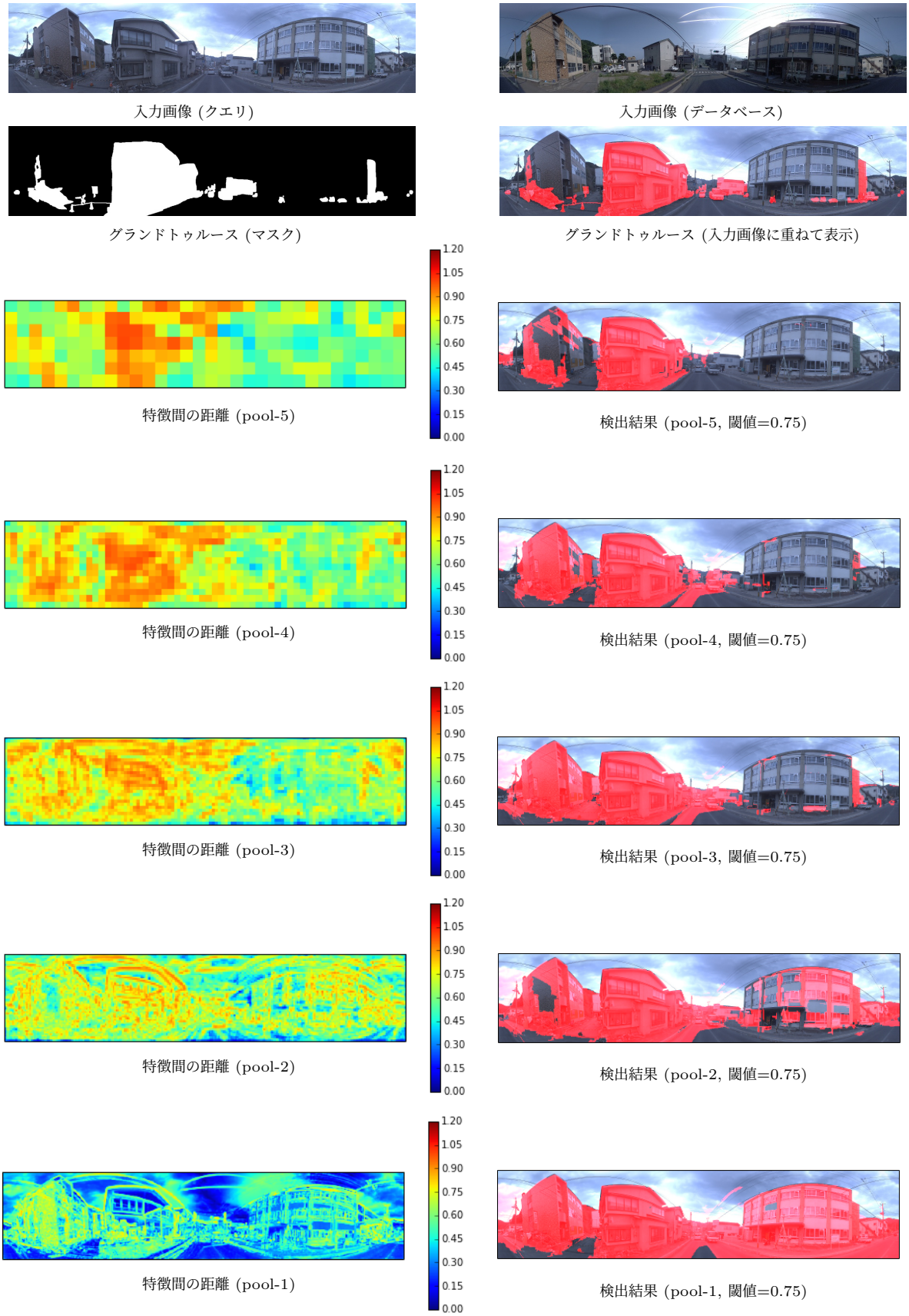


図 4 各グリッドの特徴間の距離（CNN のプーリング層）．右の検出結果の閾値は表 1 に示す．プーリング層の特徴の各要素は全て非負の値を取るため，各グリッドの正規化された特徴間の距離 $d_i \in \mathbf{d}_g$ は $0 \leq d_i \leq \sqrt{2}$ の範囲の値を取る．

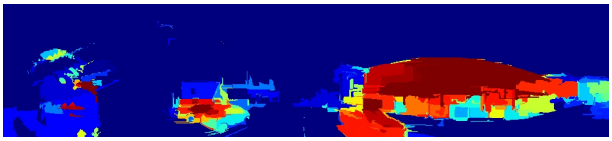

入力画像 (クエリ)

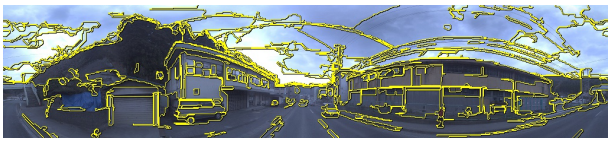

入力画像 (データベース)

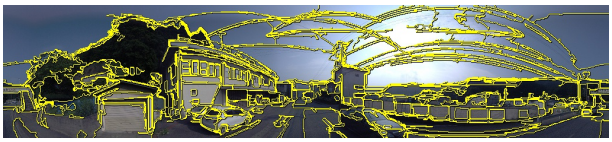

グランドトゥルース (入力画像に重ねて表示)

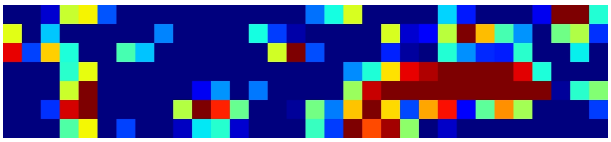

グランドトゥルース (マスク)

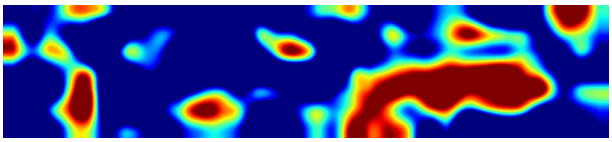

検出結果 (2 値化)

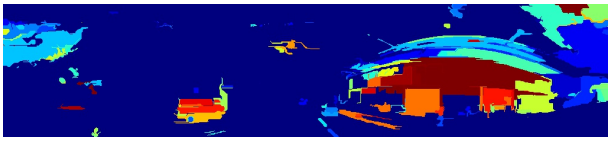

検出結果 (距離)

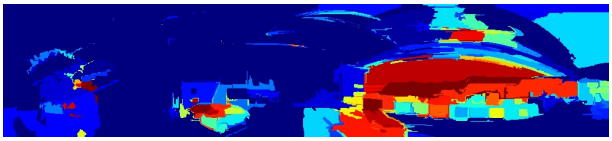

スーパーピクセルセグメンテーション (クエリ)

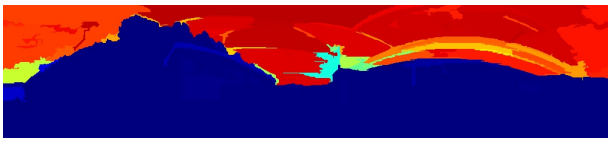

スーパーピクセルセグメンテーション (データベース)

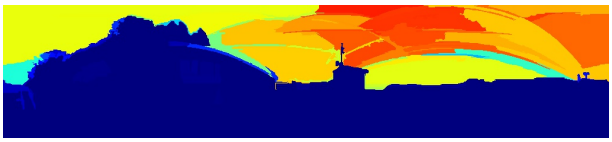

特徴の距離 (グリッド)


特徴の距離 (補間)


各スーパーピクセルの特徴間の距離 (クエリ)


各スーパーピクセルの特徴間の距離 (データベース)


空の確率 (クエリ)


空の確率 (データベース)


地面の確率 (クエリ)

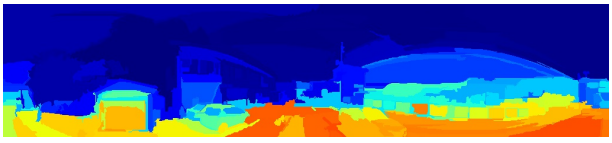

地面の確率 (データベース)

図 5 CNN のプーリング 5 層の特徴を用いたシーン変化の検出結果