Vol.2016-CVIM-200 No.33
2016/1/22

深層学習による特徴量データベース選択に基づくカメラ位置姿勢推定の高精度化

中島 由勝 [*1]　　斎藤 英雄 [*1]

Abstract – 本稿では，視点が大きく変化するようなシーンにも堅牢性のあるカメラ位置姿勢推定手法を提案する．VGL のような従来の学習型カメラ位置姿勢推定手法では SIFT 等に比べ視点の変化には堅牢性があるが，一方でデータベースの容量の問題上データを圧縮する必要があり，特徴量が大きく変化するような対象物体に対する角度の浅い視点には対応できないという問題点がある．そこで，本手法では対象物体やシーンに対し，45 度程度の角度範囲のカメラ視点毎に非圧縮の特徴量データベースを複数用意し，深層学習により視点が未知の入力画像のカメラ姿勢に対応するデータベースを選択し，そのデータベース内で最近傍探索によりマッチングを行う．評価結果では，VGL と比較して有効な特徴点数が大きく増加し精度が大幅に向上したことを示す．

Keywords : カメラ位置姿勢推定，深層学習，拡張現実感

1 はじめに

近年，実世界の平面を撮影した画像に対し仮想情報を表示する拡張現実感では，撮影された実世界シーンに対するカメラ位置姿勢を推定するために AR マーカー [1] を使用しないマーカーレストラッキングが盛んに研究されてきた．通常のマーカーレストラッキングでは，登録された平面パターンと入力画像間で正確な点対応を任意に変化するカメラの位置姿勢に対し取得する必要があり，画像から局所特徴量を取得することでその変化に対する堅牢性を得ている．Lowe の SIFT[2] は著名な局所特徴量検出アルゴリズムの一つであり，Laplacian of Gaussian(LoG) を近似した Differences of Gaussians(DoG) で特徴点を検出し，その周辺の画素情報により 128 次元の勾配ベクトルを特徴量として抽出する．ここで得られた特徴量は見え方によらず同じ特徴点に関しては近い値を示すことが要求されるが，Mikolajczyk らの研究 [3] により拡大縮小，及び回転に対し高い堅牢性を得られることが示されている．

SIFT の提案以降，数多の局所特徴量検出アルゴリズムが考案された．Yan らの PCA-SIFT[4] は SIFT と同様にして得られた特徴量に対し主成分分析を行い，128 次元勾配ベクトルを 36 次元に削減することでマッチング時におけるベクトル同士の距離計算を高速化した．また，Pablo らの AKAZE[5] は，特徴点検出に非線形で非等方的なスケールスペースを使用し，特徴量記述に Modified-Local Difference Binary(M-LDB) 等独自の手法を多数使用することで堅牢性の向上及び高速化を図った．

しかしこれらの局所特徴量検出アルゴリズムは，拡大縮小，及び回転に対する堅牢性については優れている一方，入力画像があらかじめ登録された平面パターンに対し射影的歪みを受けたような画像となると 2 つの画像間で正確な点対応が得られない場合が存在することが実験から確認できる．Yoshida らの Viewpoint Generative Learning(VGL)[6] ではこの問題点を解決すべく，あらかじめ入力された平面パターンに対し様々な視点から撮影されたかのような平面パターン群を生成し，複数のパターンから検出される特徴点についてそれらの特徴量を K-means 法により圧縮した後データベース化することで正確な点対応数を向上させた．Yoshida らはこの手法において，処理時間等の関係上，ある特徴点に対し得られた複数の特徴量である 128 次元勾配ベクトルを圧縮する際にクラスタ数を 5 とした K-means 法を推奨している．

しかしこの手法では，カメラ位置姿勢が平面パターンに対し非常に浅い角度となった場合，各特徴点に対する特徴量は大きく変化するためクラスタ数 5 の K-means 法で圧縮してしまうと情報が大きく失われ，大きな射影的歪みについては対応できないという問題点がある．この問題点の解決のため本研究では，あらかじめ浅い角度も含む複数の視点ごとに平面パターンの特徴量データベースを非圧縮で用意し，入力画像に対してその画像がおおよそどの角度から撮影されたかを判定できるように学習させた Convolutional Neural Network(CNN) により，データベース群から入力画像に近い角度での平面パターンの特徴量データベースを一つ選択しマッチングを行う手法を提案する．

[*1] 慶應義塾大学

Vol.2016-CVIM-200 No.33
2016/1/22

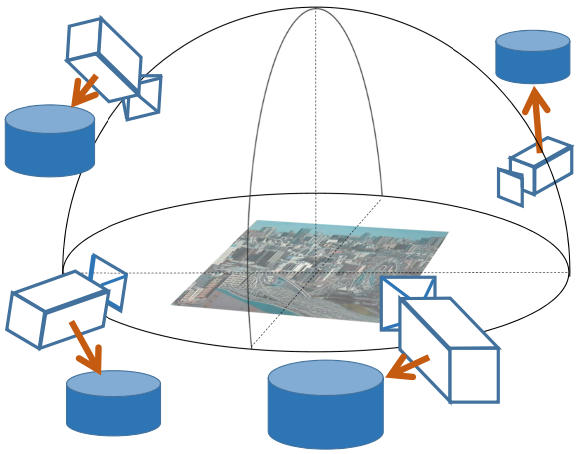

図 1　データベース生成の概念図
Fig. 1　Concept of generating database

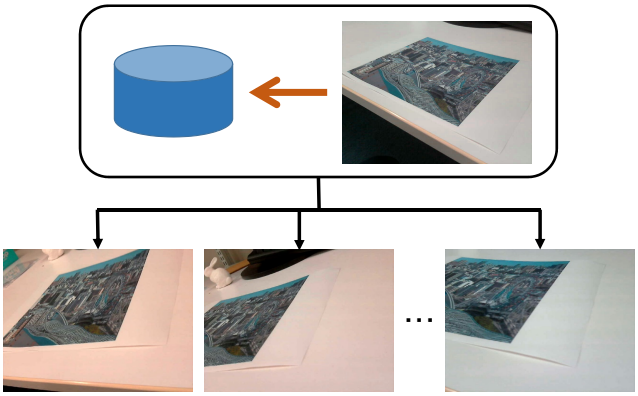

図 2　CNN の学習に使用する画像の生成例
Fig. 2　Images for deep learning

2　提案手法

局所特徴量アルゴリズムは拡大縮小及び回転に対して堅牢性があるが，視点の大きな変化には対応するとこが出来ない．また，従来の学習型のカメラ位置姿勢推定アルゴリズムではデータベースの圧縮により特徴量が大きく変化するシーンに対する堅牢性が低い．そこで我々は様々な視点ごとに特徴量データベースを非圧縮で作成することにより拡大縮小や回転に加え，視点変化についても更なる堅牢性を得る．本研究では VGL[6] に代表されるような，対象となる平面パターンの正面画像から仮想生成された画像により学習を行う手法ではなく，実際に複数の視点から撮影した画像群を用いて学習を行う．この時，対象となる平面パターンを正面から見た際の座標で特徴量データベース群を構成することでカメラ位置姿勢の推定を可能とした．本研究のような，固定された平面パターンに対して浅い角度からの視点に関しても精度の高いカメラ位置姿勢推定手法は，ポスターや広告，展示物などが対象物体となるような場合での応用が期待される．

2.1　データベース群の生成

はじめに，固定された平面パターンに対し複数の視点から画像を撮影し画像群を得る．次にそれらの画像群に対し，正面から見た際の座標が明確にわかる 4 点を与え，

$$\tilde{\boldsymbol{x}}' \sim \boldsymbol{H}\tilde{\boldsymbol{x}} \tag{1}$$

によりその画像中の平面パターンを正面画像のように変換する射影変換を表す 3×3 の行列 $\boldsymbol{H}$ をそれぞれ得る．ここで，$\tilde{\boldsymbol{x}}' \sim (x', y', 1)^{\mathrm{T}}$，$\tilde{\boldsymbol{x}} \sim (x, y, 1)^{\mathrm{T}}$ であり，$\tilde{\boldsymbol{x}}'$ は平面パターンの正面画像における座標，$\tilde{\boldsymbol{x}}$ は各視点から撮影された画像における座標である．以下，画像 i に対する射影変換行列 $\boldsymbol{H}$ を $\boldsymbol{H}_i$ とする．次に画像 i に対し適切な局所特徴の検出器により特徴点を検出する．画像 i における平面パターンを正面画像に変換する射影変換行列 $\boldsymbol{H}_i$ はすでに導出されているため，画像 i において検出された特徴点を p_{ij} とすると，p_{ij} は正面画像上の p'_{ij} に式 $p'_{ij} = \boldsymbol{H}_i p_{ij}$ により投影できる．最後に p_{ij} に対して適切な特徴量記述アルゴリズムにより p_{ij} に関する特徴量 d_{ij} を記述し，p'_{ij} と d_{ij} をひも付けたデータベースを各視点から撮影された画像ごとに生成する．ただしここで，p'_{ij} が正面画像における平面パターン内に含まれない場合，つまり平面パターン外の特徴点はデータベースに含まないものとする．上記の手法により生成された平面パターンに対するデータベース群の概念図を図 1 に示す．

2.2　視点に関する深層学習

与えられた任意の入力画像に対して，図 1 に示したデータベース群の中からその画像が撮影されたカメラ位置姿勢とより近いカメラ位置姿勢から撮影された画像により生成されたデータベースを適宜選択するため深層学習による画像分類を用いる．CNN は主に画像認識に応用される順伝播型ネットワークであり，多層の CNN は画像認識の問題全般に対する非常に重要な技術として位置づけられている．本研究ではこれを同一の平面パターンにおける視点に関する認識問題として応用する．この実現に際し，図 1 のようにして生成した各データベースについてそのデータベースを生成する際に利用した画像 i を撮影したカメラ位置姿勢と近いカメラ位置姿勢から撮影した平面パターンの画像を CNN の学習のため複数枚用意する．図 2 に学習用画像の生成例を示した．上部の画像がデータベースを生成する際に使用した画像 i であり，下部の画像群が CNN 学習のための画像群である．この操作を全てのデータベースについて行い，対応するデータベースと紐付けられた学習用の画像群を得る．

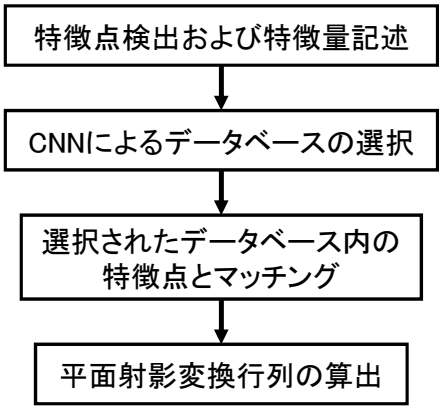

図 3　カメラ位置姿勢推定の流れ
Fig. 3　Flow of pose estimation

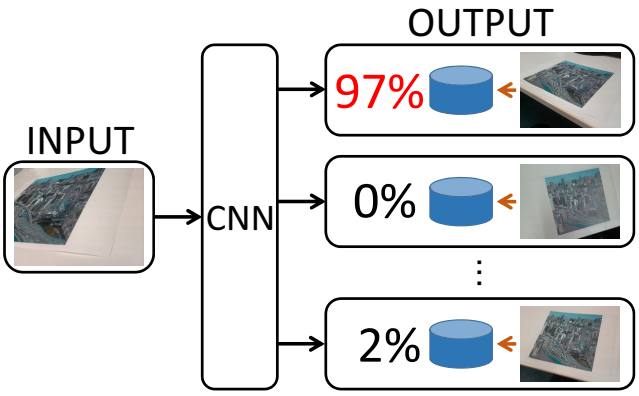

図 4　CNN によるデータベース推定の例
Fig. 4　Database estimation using CNN

次に CNN における input layer や convolution layer, pooling layer，local contrast normalization(LCN) layer, fully-connected layer，output layer の層数やユニット数，活性化関数などを適切に決定し CNN を構成する．ここで output layer のユニット数についてはデータベース群の個数と一致する必要がある．最後に，構成した CNN についてミニバッチサイズやエポック数を適切に決定し確率的勾配降下法や誤差逆伝播法を用いて各データベースを教師とした学習を行い LCN layer などのパラメータを決定する．この際，学習用に用意した各画像は input layer のユニット数に従い適宜リサイズを行う．

2.3　入力画像に対するカメラ位置姿勢推定

学習を終えると，実際の入力画像に対し対象となる平面パターンを検出し入力画像に対するカメラ位置姿勢を推定する．本研究における，入力画像に対しカメラ位置姿勢を推定する際の処理の流れを図 3 に示す．まず入力画像に対し，データベース生成時に用いた特徴点検出アルゴリズムを用いて特徴点を検出し，同様にして各特徴点の特徴量を記述する．次に入力画像を学習済みの CNN に入力する．この際，CNN の input layer のユニット数に従い入力画像を適宜リサイズする．CNN は学習済みのため，どのデータベースが入力画像とのマッチングを行う上で最適であるかが推定される (図 4 参照)．推定の結果は図 4 のように全データベースに対する割合で得られ，最も割合の高いデータベースを 1 つ選択する．次に，入力画像から得た特徴点を選択したデータベース内の特徴点とそれぞれの特徴量のユークリッド距離を用いて比較する．この時，最近傍である特徴点と 2 番めに近い特徴点の両方を探索する．これは Mikolajczyk らの，最近傍比を用いることで特徴点同士の誤対応を削減する手法 [7] であり，再近傍とのユークリッド距離が 2 番めの近傍とのユークリッド距離より十分に小さい時のみ正対応とする．入力画像のある特徴点における特徴量を D_A とし，対応するデータベース内の最近傍の特徴量を D_B，2 番めの近傍の特徴量を D_C とした時，以下の関係式が満たされる時のみ D_A と D_B はマッチングされるものとする．

$$\frac{\|D_A - D_B\|}{\|D_A - D_C\|} < t \qquad (2)$$

閾値 t を大きくすれば対応数が増加するが誤対応数も増加する．一方で閾値 t を小さくすれば誤対応数は減少するが対応数も減少する．条件式 (2) により対応するデータベース内の特徴点と入力画像の特徴点間で対応が十分量取れ特徴点同士のマッチングが終了すると，データベース内の各特徴点ごとに保持されている，対象となる平面パターンの正面画像における座標と入力画像における各特徴点同士で点対応が定まり，ロバスト推定法である RANSAC によって誤対応を更に削減した後，平面射影変換行列を算出することで入力画像に対するカメラ位置姿勢を推定する．

3　評価実験

本章では提案手法を評価するために行った実験について示す．与えられた平面パターンに対しカメラ位置姿勢が浅い角度となるようなシーンを含む，データベース群を用意する際に使用した画像と同じサイズである 800 × 600 の動画を用意し，各フレームにおける再投影誤差やカメラ位置姿勢推定に用いたマッチング数，フレームレート等を VGL と比較し評価する．

3.1　評価環境

本実験では，CPU: Intel Core i7-4770K 3.50GHz, GPU: GeForce GTX 760，メモリ: 16GB の環境下で処理時間を測定した．机上に平面パターンを固定した後，平面パターンに対し図 5 のようにまんべんなく複数の視点から撮影した 22 枚の画像群によりデータベース群を作成した．

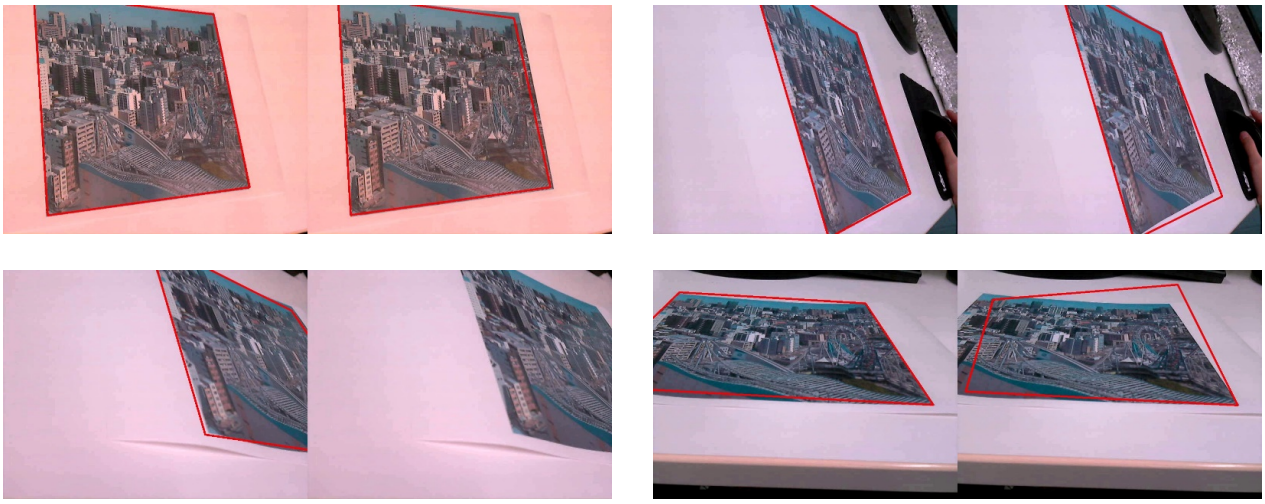

図 6　カメラ位置姿勢の推定例 (左: 提案手法，右: VGL)
Fig. 6　Examples of camera pose estimation (left:proposed method, rignt:VGL)

図 5　データベース生成に使用した画像
Fig. 5　Images for generating database

この時，非圧縮のデータベースを作成するため，検出された全ての特徴点について正面画像の座標に変換した後，各特徴点の特徴量と紐付けデータベースに格納した．ここで，データベース作成にあたり SIFT を特徴点検出及び特徴量記述アルゴリズムとして用いた．また，入力画像に対しその画像が撮影されたカメラ位置姿勢に最も近いデータベースを推定する CNN の構成には Min らの Network In Network(NIN)[8] を用いた．NIN は高い物体認識性能を保持しながらパラメータ数を大幅に削減するため物体認識に要する処理時間が短縮され，本研究のように各フレームに対し推定に要する時間を高速化する必要がある際に有効である．

次に，構成した CNN が入力画像についてのカメラ位置姿勢を推定できるように学習させるため，各データベースを生成するにあたり使用した画像が撮影されたカメラ位置姿勢の周辺から動画を撮影し，各フレームを切り出すことでそれぞれのデータベースにつき 600 枚程度の学習用画像群を図 2 のように用意した．

この画像群を用いエポック数を 10，ミニバッチサイズを 32 として学習を行った．この時，全画像のうち 97%を学習用として用い，残りの 3%をテスト用として用いたところ，約 2 時間程度でテスト用画像の推定誤差が 0 %となることを確認した．

また本評価では推定された平面パターンに対するカメラ位置姿勢を，コーナーの再投影誤差によって評価する手法を用いる．この手法では，テスト画像内の平面パターンについてその 4 隅の点が推定された位置により評価を行う．テスト画像内の平面パターンにおける，真値である 4 隅の点をそれぞれ p_i とし，データベースとテスト画像のマッチングにより算出された平面射影変換行列により投影した 4 隅の点を q_i とした時，再投影誤差 err は以下の式により表される．

$$err = \sqrt{\frac{1}{4}\sum_{i=1}^{4}\|p_i - q_i\|^2} \tag{3}$$

err が小さいほど正確にカメラ位置姿勢が推定されたことを示しており，err が 30 画素を超える場合はカメラ位置姿勢の推定は失敗したものとして扱う．

比較対象である VGL については，本提案手法のデータベース群生成に用いた画像 22 枚を用い，パラメータをクラスタ数: 5，stable keypoint 数: 2000 としてデータベースを生成した．

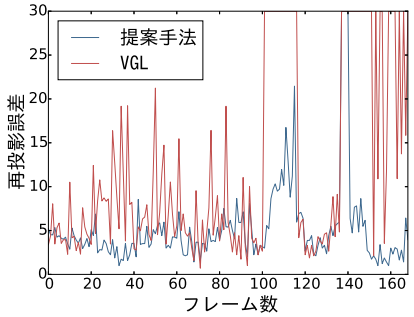

図 7　再投影誤差の比較
Fig. 7　Re-projection error

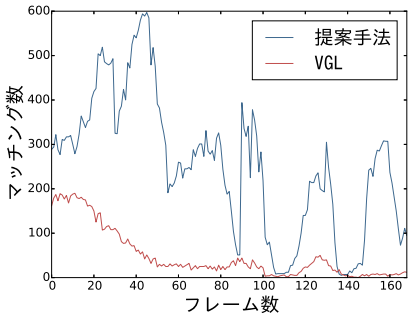

図 8　マッチング数の比較
Fig. 8　Number of matching

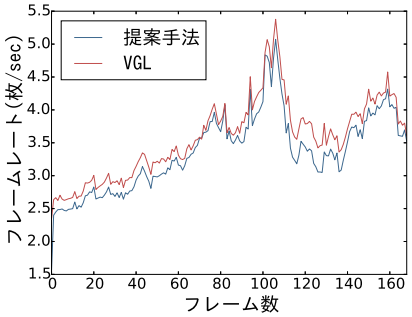

図 9　フレームレートの比較
Fig. 9　Frame rate

3.2　評価結果及び考察

評価のため，評価用の動画を用意した．図 6 は用意した評価用の動画に対し本提案手法及び VGL によりカメラ位置姿勢を推定した結果の一部である．算出された平面射影変換行列を用いて平面パターンの正面画像の 4 隅の点を再投影し赤い枠線によりカメラ位置姿勢推定の可視化を図った．また，赤い枠線が表示されていないものは射影変換行列の算出に足るマッチング数が得られなかったことを示す．図 6 を見ると，平面パターンに対し正面に近い位置姿勢から撮影されたフレームについてはほぼ性能差はないが，平面パターンに対するカメラ位置姿勢が浅い角度となると提案手法の方がより正確にカメラ位置姿勢を推定できていることがわかる．図 7 に評価用の動画の各フレームについて式 (3) により算出した再投影誤差を示す．ここで，評価用動画における各フレームでの平面パターンの 4 隅の点の真値については，すべて手動で検出した．図 7 を見ると特に浅い角度を含む後半部分のフレームでのカメラ位置姿勢推定精度について提案手法が VGL を大きく上回っていることが確認できる．また，図 8 に平面射影変換行列の算出に用いた入力画像の特徴点とデータベース内の特徴点のマッチング数を示す．提案手法のマッチング数が増減を繰り返している原因は，入力画像とのマッチングを行うデータベースが CNN による選択により変動したためである．図 6，図 7 及び図 8 を見ると，VGL は特徴量を圧縮するためカメラ位置姿勢が浅くなり各特徴点の特徴量が大きく変動した場合，正確なマッチングを行えずカメラ位置姿勢推定の精度が低下した事がわかる．一方で提案手法は，入力画像に対し適切な非圧縮のデータベースが適宜選択されるためマッチングがより正確に行われカメラ位置姿勢推定がより高精度に行われたことが確認できる．

次に処理時間についての評価結果を示す．図 9 は各フレームにおける処理時間の逆数を提案手法及び VGL について図示したものである．この結果を見ると，提案手法における CNN を用いたデータベース選択によるオーバーヘッドは十分に小さいことが確認できる．

4　結論

本稿では，対象物体に対し，45 度程度の角度範囲のカメラ視点毎に非圧縮の特徴量データベースを複数用意し，深層学習により視点が未知である入力画像のカメラ姿勢に対応するデータベースを選択し，そのデータベース内で最近傍探索によりマッチングを行うことで入力画像におけるカメラ位置姿勢を推定する手法を提案した．この提案手法により，従来手法に比べ入力画像とデータベース間のマッチング数が大幅に向上し，正確な点対応が複数得られることでカメラ位置姿勢推定の精度が向上した．また，処理時間についても CNN を用いたデータベース選択によるオーバーヘッドが小さく，即応性が十分であることを確認した．

今後の課題としては，3 次元物体への応用や，平面パターンの正面画像 1 枚により自動的に学習を行う生成型学習への発展などがあげられる．

参考文献

[1] H. Kato and Mark Billinghurst: Marker Tracking and HMD Calibration for a Video-based Augmented Reality Conferencing System; IWAR, 1999
[2] David G. Lowe: Distinctive image features from scale-invariant keypoints; IJCV, Vol.60, pp.91-110, 2004
[3] Krystian Mikolajczyk and Cordelia Schmid: A performance evaluation of local descriptors; TPAMI, Vol.27, pp.1615 − 1630, 2005
[4] Yan Ke and Rahul Sukthankar: Pca-sift： A more distinctive representation for local image descriptors; CVPR, 2004
[5] P. F. Alcantarilla and J. Nuevo and A. Bartoli: Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces; BMVC, 2013
[6] 吉田拓洋, 斎藤英雄, 清水雅芳, 田口哲典視点生成型学習による頑健な平面位置姿勢推定日本バーチャルリアリティ学会論文誌, Vol.17, No.3, 2012
[7] K．Mikolajczyk, T. Tuytelaars, Cordelia Schmid, Andrew Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. Van GooL: A comparison of affine region detectors; IJCV, Vol.65, pp.43-72, 2005
[8] Lin Min; Chen Qiang, Yan Shuicheng: Network In Network; ICLR, 2014