CH-001

パラメータ推定とピクセルラベリングの同時学習に基づく競技コート認識
Sports Court Recognition with Multi-task Learning

田良島 周平†
Shuhei Tarashima

1. はじめに

スポーツ映像解析において，映像に写る競技コートを認識することは，試合状況の可視化や選手のパフォーマンスの定量化，戦術の分析，リクルーティングといった応用を実現する上で重要な役割を担う[1-2]．本稿において競技コート認識とは，映像フレーム(e.g. 図 1(a))に写る競技コートを所定のモデル(e.g. 図 1(b))に対応付ける幾何パラメータを推定することと定義する．スポーツ中継やホームビデオの映像は，試合の状況に応じてカメラポーズを変化させながら撮影されたものが多く，これらの映像を対象とする場合競技コート認識はフレーム毎に行う必要がある．また，特に可視化や戦術分析といった応用を想定すると，コート認識は高い精度に加え，リアルタイム相当の速度で処理可能であることも求められる．したがって，競技コート認識処理は，認識精度と処理速度とを高いレベルで両立できるアルゴリズムであることが望ましい．

多くの競技コート認識技術[3-5]は人手の介在を前提としており，時刻とともにポーズの変化するカメラから得られる映像をリアルタイムに解析するという用途には不向きである．人手を介さず自動でコートを認識する手法もいくつか提案されている[6-8]ものの，これらのアルゴリズムでは精度と速度がトレードオフの関係にあり，両者をバランスさせることが容易ではない．精度と速度が両立しうるアプローチの一つとして，モデルの順伝播のみで競技コート認識を実現させるという方法が考えられる．現時点で，数十層の CNN がリアルタイムを上回る処理速度で順伝播可能であるという報告は多くなされており(e.g. [10,17])，このアプローチを採用することで高いスループットが得られると期待できる．しかし単純なモデル学習，推論アプローチを適用するのみでは，入力フレームの自由度に対して十分な汎化性能が得られないことが懸念される．実際に我々は，競技コート認識のみを考慮した学習から得られるモデルでは十分な精度が得られないことを実験的に確認している(cf. §4.4).

そこで本研究では，順伝播 CNN による競技コート認識の汎化性能を向上させることを目的として，関連する別タスクとの同時学習によりモデルを獲得することを提案する．具体的には，入力フレーム中の競技コート領域をピクセル毎にラベリングするタスク(領域分割タスク)を考え，これを競技コート認識タスクと共に解く CNN(図 1(c)の提案モデル)を教師データ所与のもと学習する．入力フレームのラベリングは幾何変換パラメータ推定に有効な領域を検出する処理と解釈できるため，その推論過程で得られる中間特徴はパラメータ推定にも有効であると考えられる．加えて提案手法では，ラベリング精度，ひいては競技コート認識精度を向上させるため，コートを構成するラベル間の空間関係の特性をモデル学習時に陽に考慮する手法を提案する(cf. §3.3)．上記の学習に必要な教師データは，映像フレームに対応する正解幾何パラメータを算出するために必要な人手のアノテーションがあれば自動構築が可能であり，追加のアノテーションの必要はない．本研究では，バスケットボール映像を対象としたデータセットを新たに構築(cf. §4.1)し提案手法を評価する．実験の結果，提案手法が 30fps を超える速度で処理可能であり，かつベースラインを上回る性能が得られることを示す．

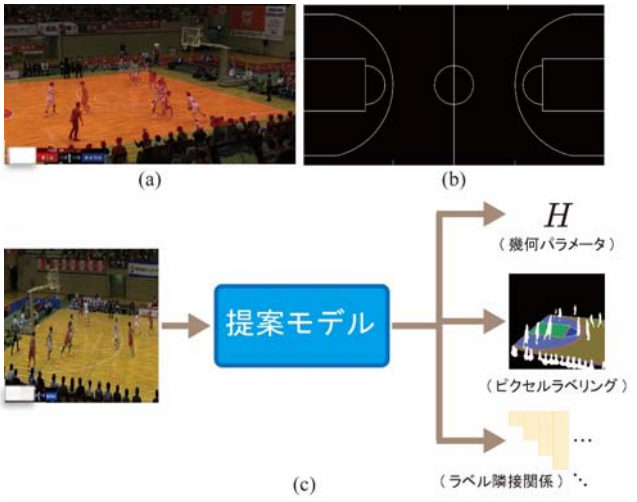

図 1. 競技コート認識

2. 関連研究

多くのスポーツにおける競技コートは平面である．この場合，コートとモデルを対応付ける幾何パラメータは 3x3 の射影変換行列(自由度 8)で定義することができる．射影変換行列は，4 つの対応点あるいは直線から DLT アルゴリズムによって推定できることが広く知られている．一方で入力フレームからモデルと対応する領域を自動で検出することは難度が高いため，多くのアプローチでは人手で対応領域を指定することを前提としている．人手の負荷を軽減するため，例えば[3-4]では，映像中のいくつかのフレームにのみ人手のアノテーションで射影変換行列を推定しておき，残りのフレームについては連続するフレーム間の射影変換行列を特徴点マッチング等で推定するより補間している．

†NTT コミュニケーションズ(株) 技術開発部 /
Technology Development, NTT Communications Corp.

Copyright © 2019 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

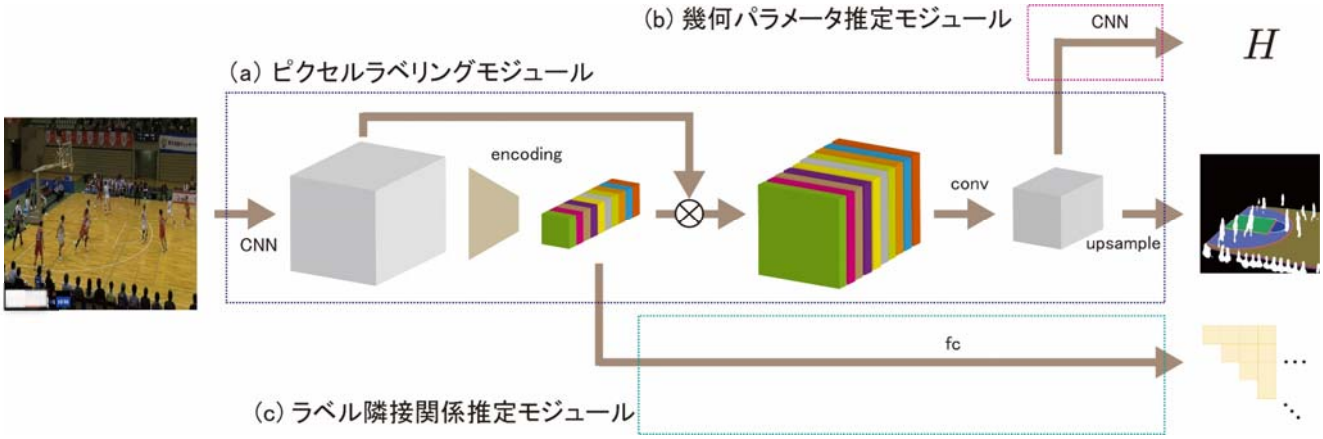

図 2. 提案モデル

また[5]では，PTZ カメラで取得された映像であること仮定することにより人手で指定が必要な領域の数を削減する手法が提案されている．しかしこれらの改善はあくまでオフラインでの使用を想定したものであり，例えば試合状況の可視化やオンラインでの戦術分析など，リアルタイム性が強く求められる用途に応用することは現実的ではない．

我々の知る限り，人手を介さずに競技コートを認識する既存手法は少ない．例えば[6]では，競技コート認識問題を，フレーム座標系で競技コートの平行線が交わる消失点の位置推定問題に帰着させ，それをマルコフ確率場におけるエネルギー最小化の枠組みで推定する方法が提案されている．また[7-8]では，対象フレームから抽出した画像特徴でデータベースを検索し，得られた類似画像に対応する座標変換パラメータから競技コートを認識する手法が提案されている．上記のアプローチでは，粒度の細かいラベル空間や大規模なデータベースを用いることで認識精度の向上が見込める．一方で，ラベルやデータベースのサイズを大きくすることはオンラインでの計算コストを増大させることに直結する．すなわち，推論精度を向上させるためのチューニングが計算コストとトレードオフの関係にあるため，これらをバランスさせることが難しいという問題がある．

上記の考察に基づき，本研究では，入力フレームを提案モデルへ順伝播させることで幾何パラメータを直接推定するアプローチを採用する．次章では，この提案モデルの具体的なアーキテクチャについて述べる．

3. 提案手法

図 2.に提案モデルの概要を示す．提案モデルは，大きく(a)ピクセルラベリングモジュール，(b)幾何パラメータ推定モジュール，(c)ラベル隣接関係推定モジュールから構成されている．以下では各モジュールについて詳細を述べた後に，モデル学習のための誤差関数について説明する．

3.1 ピクセルラベリングモジュール

ピクセルラベリングモジュールは，入力フレームの各ピクセルを予め決められたラベルのいずれかに割り当てるタスクを担う．本研究では，[9-10]に基づき当該モジュールを構築する．具体的には，事前学習済み CNN の最終畳み込み層の後にチャネル間の重要度を推定するエンコーダ(encoder)と畳み込み層(conv)を設置し，最終畳み込み層出力の特徴マップをエンコーダ出力で重み付けした上で畳み込み層へ入力，その出力を空間方向へアップサンプリング(upsample)することで入力フレーム各ピクセルのラベルを推定する．本研究では，事前学習済み CNN として ResNet-50[11]，エンコーダとして Context Encoding Module[9]を用いる．また ResNet-50 の第 3，第 4 ステージの畳み込み層には Joint Pyramid Upsampling Module[10]を導入し，出力特徴マップの空間解像度を向上させた．

3.2 幾何パラメータ推定モジュール

幾何パラメータ推定モジュールは，入力フレームに写る競技コートをモデルへと対応付ける射影変換行列パラメータを推定するタスクを担う．提案モデルでは，ピクセルラベリングモジュールが出力するアップサンプリング前の特徴マップを入力とする小規模な CNN を設置する．この CNN の具体的なアーキテクチャは，いくつかの選択肢の中から実験的に決定した(cf. §4.3)．なお本モジュールが出力する射影変換行列のパラメータは，縦横共に[-1,1]の区間で正規化されたフレームとモデルを対応付ける射影変換行列パラメータのうち，3 行 3 列目(常に 1)以外の 8 要素とする．このようにすることで推定する値間の分散が小さくなり，回帰問題の難度を緩和することができる[12-13].

3.3 ラベル隣接関係推定モジュール

ピクセルラベリングの学習で用いられる誤差関数はピクセル毎の誤差値の和で定義されることが一般的である．この場合，入力フレーム中で占める割合の低いラベルほど対応する誤差が考慮されなくなるため，小さい，あるいは細長い領域の分割精度を向上させることは容易ではない．[9]では，この問題に対し各ラベルが入力フレームの中に出現するか否かを推定するモジュールを導入し，それと領域分割モジュールとを同時学習することが提案されている．

Copyright © 2019 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

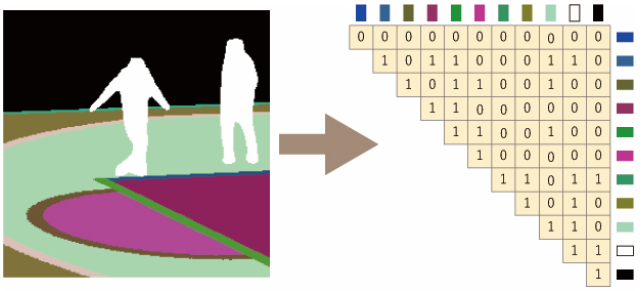

図 3. ラベルの隣接関係

本研究では，[9]のアイデアを競技コート認識の文脈をふまえ拡張する．具体的には，競技コートそのものは同一映像内で変形することがなく，構成する領域間の位置関係に常に一貫性があることに着目し，入力フレーム内でラベルが存在するか否かに加え，任意のラベルペアがフレーム内で隣接するか否かを推定するモジュール(ラベル隣接関係推定モジュール)を導入する．図 3 に，ある入力フレームの正解マスク(左部)に対応するラベル隣接関係(右部)を示す．図右部の上三角行列の各行各列は一つのラベルに対応し，対角成分の各要素は対応するラベルがマスクに含まれるか否か(含まれる場合に 1 をとる)を示している．また対角成分以外の各要素は，行と列に対応するラベルが正解マスク中で隣接するか否か(隣接する場合に 1 をとる)を示している．[9]ではこのラベル隣接関係のうち対角成分のみが考慮されていたと解釈することができる．これに加えラベル間の隣接関係を明示的が推定できるようモデルを学習することによって，より認識対象の構造に基づくコンテキストを考慮することができるようになり，結果領域分割や幾何パラメータ推定の精度向上が見込める．

本研究では，ラベル隣接関係推定モジュールを，エンコーダ出力を入力とする全結合層で定義した．本モジュールの最終出力の次元数は，ラベル数を N_{label} として $N_{label}(N_{label}+1)/2$と計算できる．

3.4 目的関数

提案モデルの学習時には，上記 3 モジュールの各出力と対応する正解データとを入力とする誤差関数の出力の和 $L = L_{param} + w_{label}L_{label} + w_{spatial}L_{spatial}$ を目的関数と定義し，これを誤差逆伝播法によって最小化した．ピクセルラベリングモジュールの誤差関数L_{label}にはピクセル毎のクロスエントロピーの和，幾何パラメータ推定モジュール出力の誤差関数L_{param}には Smoothed L1 Loss[14]，ラベル隣接関係推定モジュールの誤差関数$L_{spatial}$には要素ごとのバイナリクロスエントロピーの和を用いる．重みw_{label}，$w_{spatial}$はそれぞれ 1.0，0.2 と設定した．

4. 評価

本章の評価では，競技コート認識精度の評価指標として，[-1,1]間で正規化された座標空間における対応点誤差(NPPE)[6]を，ピクセルラベリングの評価指標として平均

図 4. 対応点とアノテーション結果の例

IoU(mIoU)を用いる．NPPE は低いほど，mIoU は高いほど高い性能を示す指標である．

4.1 データセットの構築

提案手法を評価するため，本研究ではバスケットボール映像を対象としたデータセットを新たに構築した．まず，異なる会場で開催された計 23 の試合映像を取得し，その中からランダムで画像フレームを抽出し，得られた各フレームについて，モデルに対し定義された計 28 の点のうち写り込む対応点を人手で指定した．対応点とアノテーション結果例を図 4．に示す．4 点以上の対応点が得られたフレームについて DLT アルゴリズムを用いて射影変換行列パラメータを推定した結果，計 615 フレームとそれに対応する射影変換行列からなるデータセットを得た．映像あたりに含まれるフレーム数は 20～50 である．

ピクセルラベリングモジュールの学習/評価に必要な正解マスクは，モデルに対応する正解マスクを射影変換行列を用いてフレームへマッピングしたうえで，人物領域抽出手法[15]を適用することで得られたフレーム内人物領域を重畳することで生成した．またラベル隣接関係の正解データは，正解マスクから容易に生成可能である．いずれの正解データも，競技コート認識を目的とするデータセットから自動で生成することができる．

本稿では，上記で構築したデータセットを，3 会場 60 フレームからなるテスト用データと，それら以外の会場の全フレームからなる学習用データに分け評価を行う．

4.2 学習

§3 で示したモデルを学習するにあたり，本研究ではまずピクセルラベリングモジュールの誤差関数L_{label}とラベル隣接関係推定モジュール$L_{spatial}$の誤差関数を用いて対応するモジュールを学習させた(ステップ 1)のちに，幾何パラメータ推定モジュールの誤差関数L_{label}も含めて全体を学習させた(ステップ 2)．ステップ 1 は 50 エポック，ステップ 2 は 30 エポック学習させる．いずれの最適化にも momentum-SGD(モーメントは 0.9 に設定)を使用，学習率の初期値をステップ 1 では 0.04，ステップ 2 では 0.01 と設定し，[16]と同様の方法でエポック毎にそれらを減衰させた．なお両ステップにおいて，事前学習済モジュールの学習率には上記を更に 1/10 させて用いている．バッチサイズは 16 とした．また学習時には，入力フレームをそのアスペクト比を保ちながらランダムにクロッピングし，それに対応する幾何パラメータ，正解マスク，ラベル隣接関係を算出することでデータを水増しした．

Copyright © 2019 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

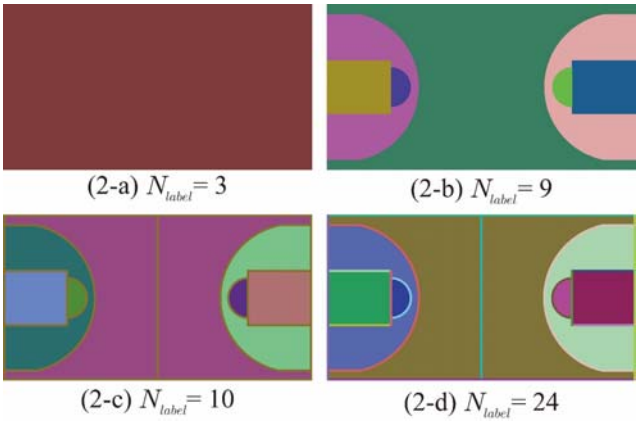

図 5. 競技コートのラベルパターン

4.3 Ablation Study

本節では，本研究のアイデアの検証やチューニングに関する 3 つの実験結果を示す：

幾何パラメータ推定モジュールのアーキテクチャ　異なる幾何パラメータ推定モジュール(cf. §3.2)アーキテクチャを採用した際の精度を表 1.に示す．いずれのアーキテクチャについても，畳み込み層のフィルタサイズは 3x3，ストライドは 1，パディングは 1 を採用し，畳込み層の後には ReLU と最大値プーリング層(フィルタサイズは 2x2，ストライドは 2)を設置した．表 1.各セル内の括弧はチャネルサイズを表している．

表 1 から，パターン(b)のケースで最も高い精度が得られることが分かる．以下の実験では，幾何パラメータ推定モジュールは(1-a)に固定して実験を進める．

表 2. 幾何パラメータ推定モジュールの設計

	アーキテクチャ	NPPE↓
1-a	conv(30)-conv(36)-conv(42)-fc(8)	**0.022**
1-b	conv(36)-conv(48)-conv(60)-fc(8)	0.026
1-c	conv(48)-conv(72)-conv(96)-fc(8)	0.032

競技コート領域のラベルパターン　競技コート領域のラベルパターンには任意性があるため，本稿では図 5.に示す 4 つのパターンを考え，各々ケースでの精度評価を行った．教師フレームへのアノテーションはモデルのラベルを射影変換行列でマッピングするため，パターン間でアノテーションコストに差異はない．結果を表 2.に示す．表 2 第二列の mIoU について，パターン間ではラベル構成が異なるため単純比較はできないことに注意されたい．表 2.から，いずれのパターンでもピクセルラベリングの学習は精度良くなされている一方で，NPPE の精度は最も細かく競技コートを分割したパターン 2-d が最も高いことが分かる．そこで以下の実験では，競技コートのラベル構成を 2-d に固定して議論を進める．

ラベル隣接関係推定モジュールの効果　ラベル隣接関係推定モジュールの効果を評価するため，提案モデルの構成(3-c)を，ラベル隣接関係推定モジュールを用いない構成(3-a)

表 2. 競技コートのラベル構成

	N_{label}	mIoU	NPPE↓
2-a	3	0.898	0.064
2-b	9	0.835	0.035
2-c	10	0.788	0.023
2-d	24	0.577	**0.022**

およびラベル隣接関係推定モジュールを[9]で提案されているモジュールに置き換えた構成(3-b)と比較した．結果を表 3.に示す．表 3 から，[9]で提案されているモジュールを導入することによりピクセルラベリング性能，競技コート認識性能が向上しており，またそれらの性能は提案モジュールを用いることで更に改善していることがわかる．

このことから，各ラベルの共起のみでなく，ラベル間の隣接関係を考慮することで，ピクセルラベリングおよび幾何パラメータ推定精度が向上することが分かる．

表 3. ラベル隣接関係推定モジュールの効果

	mIoU↑	NPPE↓
3-a	0.498	0.028
3-b	0.541	0.024
3-c (Ours)	**0.577**	**0.022**

4.4 ベースラインとの比較

§4.3 の結果をふまえ，本節では，提案手法を後述する 2 つのベースライン手法との比較評価を行う：

Baseline A　本ベースラインでは，事前学習済の CNN を転移学習することで，入力フレームの順伝播によって直接幾何パラメータを推定する．事前学習済モデルには提案モデル同様 ResNet-50 を適用し，その最終畳み込み層の後に表 1 (1-a)に相当する CNN を結合することでモデルを構築した．学習は，§4.2 のステップ 2 と同様に実施した．

Baseline b　本ベースラインでは，入力フレームから抽出した画像特徴を用いて学習データを検索し，最も類似したフレームに対応する座標変換パラメータを推定結果とする．フレームの画像特徴抽出には ResNet-50 を用い，その平均値プーリング層が出力する 2048 次元の特徴を L2 正規化した上で適用した．画像特徴は正規化した上で L2 ノルムを用いて距離計算を行った．

競技コート認識精度および処理速度の結果を表 4.に示す．表 4.の処理時間の結果は，Core i9 CPU，GTX1080Ti を搭載するデスクトップ PC を用いて測定した．表 4.から，提案手法の競技コート認識性能が，ベースラインと比べより高いことが分かる．また，本研究の評価環境における提案モデルの処理速度が 30fps 上回っていることがわかり，高いリアルタイム性が求められる用途にも応用可能であることが示唆されている．

図 6.に，各手法の定性的な結果例を示す．ここからも，提案手法の優位性が確認できる．

Copyright © 2019 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

図 6. ベースライン手法との定性的な比較

Copyright © 2019 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

図 7. 提案手法の失敗例

表 4. ベースライン手法との定量的な比較

	NPPE↓	FPS↑
Baseline A	0.127	23.1
Baseline B	0.287	13.6
Ours	**0.022**	**32.3**

5. 結論

本研究では，入力フレームに写る競技コートを認識する順伝播 CNN モデルを，ピクセルラベリングとの同時学習により獲得する手法を提案した．バスケットボール映像を対象とした新たなデータセットで手法を比較評価し，本研究のアイデアの有効性をコート認識精度および処理速度の観点で示した．

最後に，図 7.に提案手法で認識が失敗してしまった例を示す．コート中央部が写り込んだフレームで精度が低下する傾向があり，この原因の一つとして，このようなフレームではアノテーションで十分な対応点が取得できず，データセットの中にコート中央部が写り込んだデータの数が少なくなってしまったことが考えられる．今後は，このような隔たりを小さくするようデータセットの拡張を進めていく予定である．

参考文献

[1] R. Theagarajan, F. Pala, X. Zhang and B. Bhanu, "Soccer: Who Has The Ball? Generating Visual Analytics and Player Statistics", in Proc. CVPR Workshop (2018).

[2] S. Giancola, M. Amine, T. Dghaily, B. Ghanem, "SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos", in Proc. CVPR Workshop (2018).

[3] A. Gupta, J. J. Little, R. J. Woodham, "Using Line and Ellipse Features for Rectification of Broadcast Hockey Video", in Proc. CRV (2011)

[4] 大内一成，小林大祐，中洲俊信，青木義満，“ラグビー映像解析システムの開発”，電子情報通信学会通信ソサイエティ和文論文誌，J100-B，No.12 (2017).

[5] J. Chen, F. Zhu, J. J. Little, "A Two-point Method for PTZ Camera Calibration in Sports", in Proc. WACV (2018).

[6] N. Homayounfar, S. Fidler, R. Urtasun, "Sports Field Localization via Deep Structured Models", in Proc. IEEE CVPR (2017).

[7] J. Chen, J. J. Little, "Sports Camera Calibration via Synthetic Data", in arXiv (2018).

[8] R. A. Sharma, B. Bhat, V. Gandhi, C. V. Jawahar, "Automated Top View Registration of Broadcast Football Videos", in Proc. WACV (2018).

[9] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, "Context Encoding for Semantic Segmentation", in Proc. CVPR (2018).

[10] H. Wu, J. Zhang, K. Huang, "FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation", in arXiv (2019).

[11] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", in Proc. CVPR (2016).

[12] D. DeTone, T. Malisiewicz, A. Rabinovich, "Deep Image Homography Estimation", in arXiv (2016).

[13] C.-H. Lin, S. Lucey, "Inverse Compositional Spatial Transformer Networks", in Proc. CVPR (2017).

[14] Ross Girshick, "Fast R-CNN", in Proc. ICCV (2015).

[15] K. He, G. Gkioxari, P. Dollár, R. Girshick, "Mask R-CNN", in Proc. ICCV (2017).

[16] W. Liu, A. Rabinovich, A. C. Berg, "ParseNet: Looking Wider to See Better", in arXiv (2016).

[17] X. Zhou, D. Wang, P. Krähenbühl, "Objects as Points", in arXiv (2019).

Copyright © 2019 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.