3次元 CNN を利用した Wi-Fi CSI によるジェスチャ認識

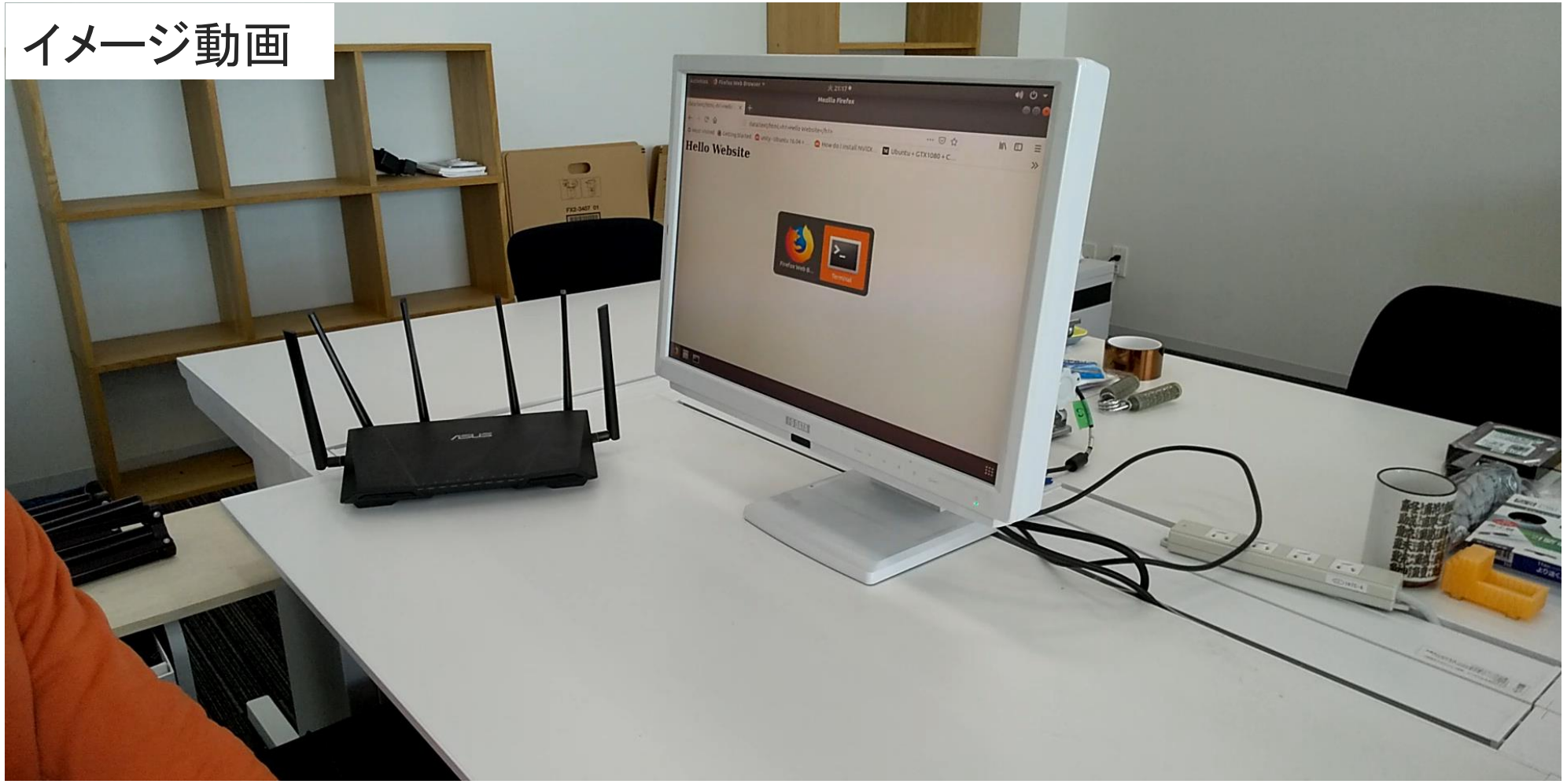
HCI研究会 182

明治大学 宮下研究室

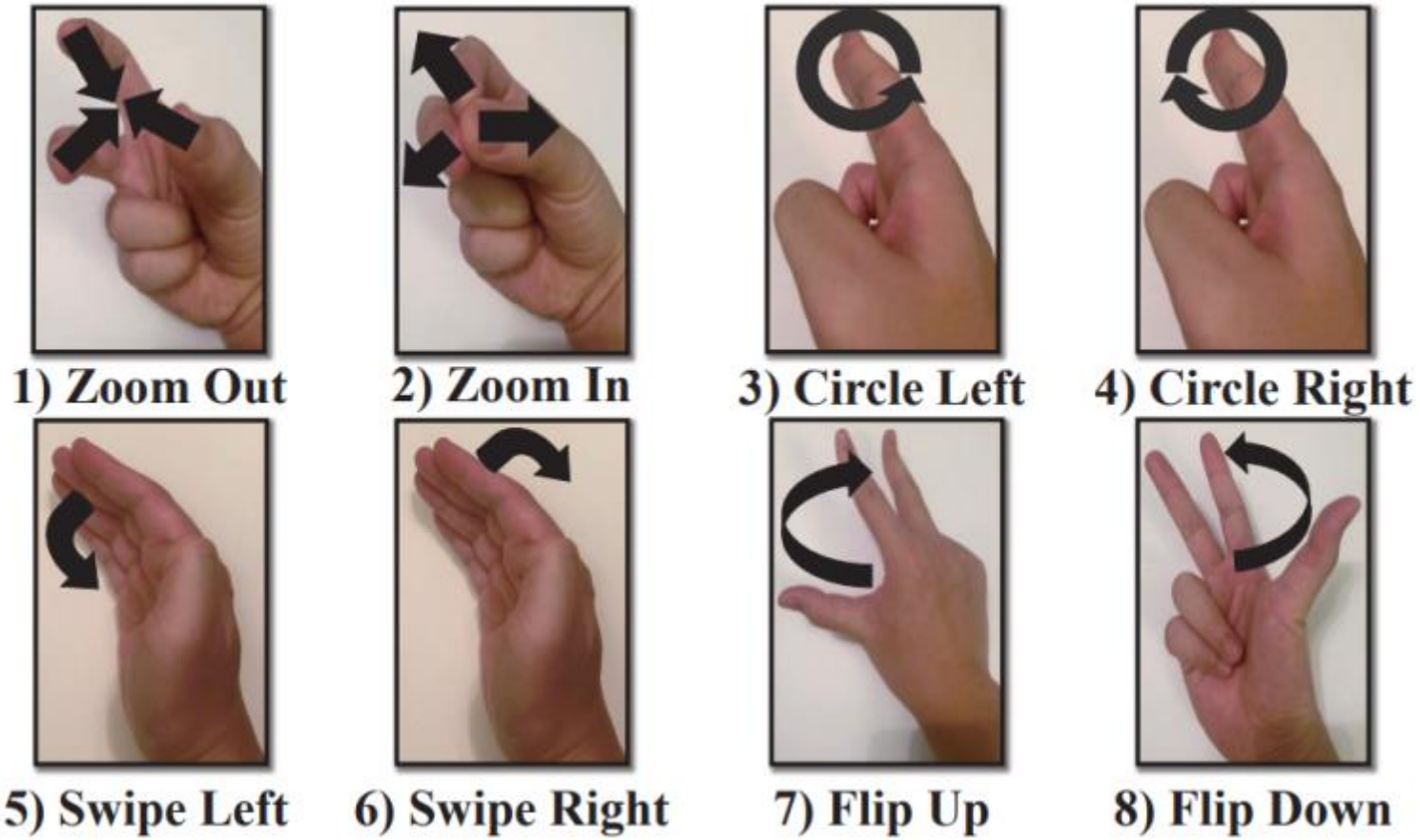
宮代 理弘 / 宮下 芳明

2台の Wi-Fi デバイス間でジェスチャ認識

イメージ動画



8つのジェスチャ



先行研究 WiFinger [Tan+, 2016] にならったジェスチャ

自分のデータから
学習させたモデルでは，
最低でも 0.932 の精度で
8つのジェスチャを認識

[Tan+, 2016]: Tan, S. and Yang, J.: WiFinger: Leveraging Commodity WiFi for Fine-grained Finger Gesture Recognition, Proceedings of the 17th ACM International Symposium on Mobile AdHoc Networking and Computing, MobiHoc '16, New York, NY, USA, ACM, pp. 201-210 (on-line), DOI: 10.1145/2942358.2942393 (2016).

提案手法

3次元 CNN を利用した Wi-Fi CSI によるジェスチャ認識

- Wi-Fi CSI を使う利点
- CNN（畳み込みニューラルネットワーク）を使う利点
- ジェスチャ認識にフォーカスした理由

Wi-Fi CSI の利点

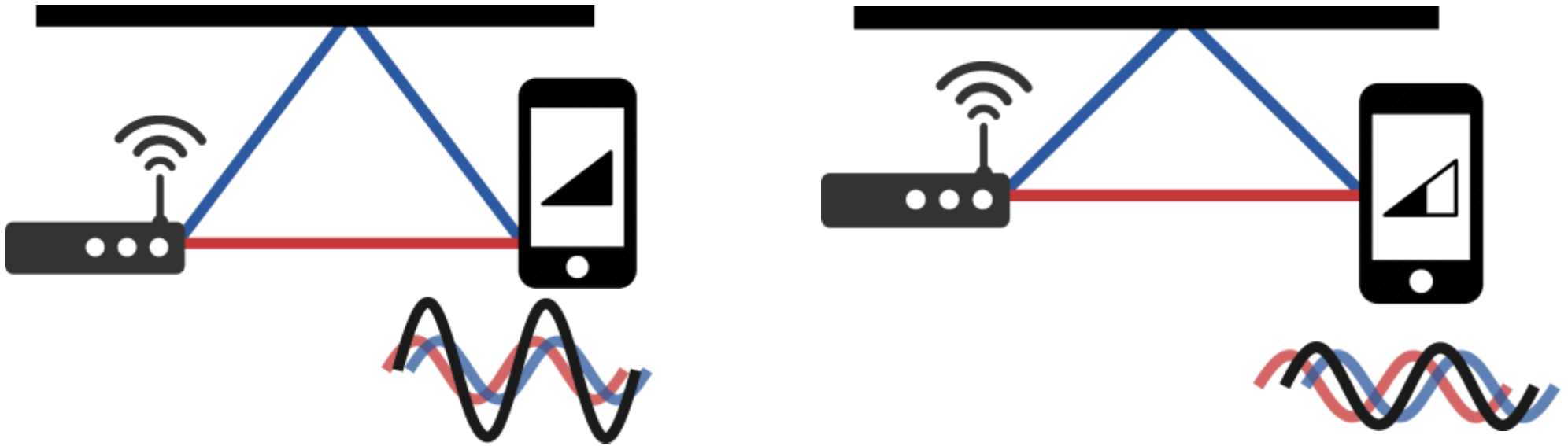
Wi-Fi から取れる変位情報

RSSI
受信信号強度

CSI
チャネル状態情報

RSSI (Received Signal Strength Indicator)

- 受信した電波の強度（振幅）のこと（よくで表示されるもの）
  - マルチパス（反射・回折による他経路）による影響を受けやすい



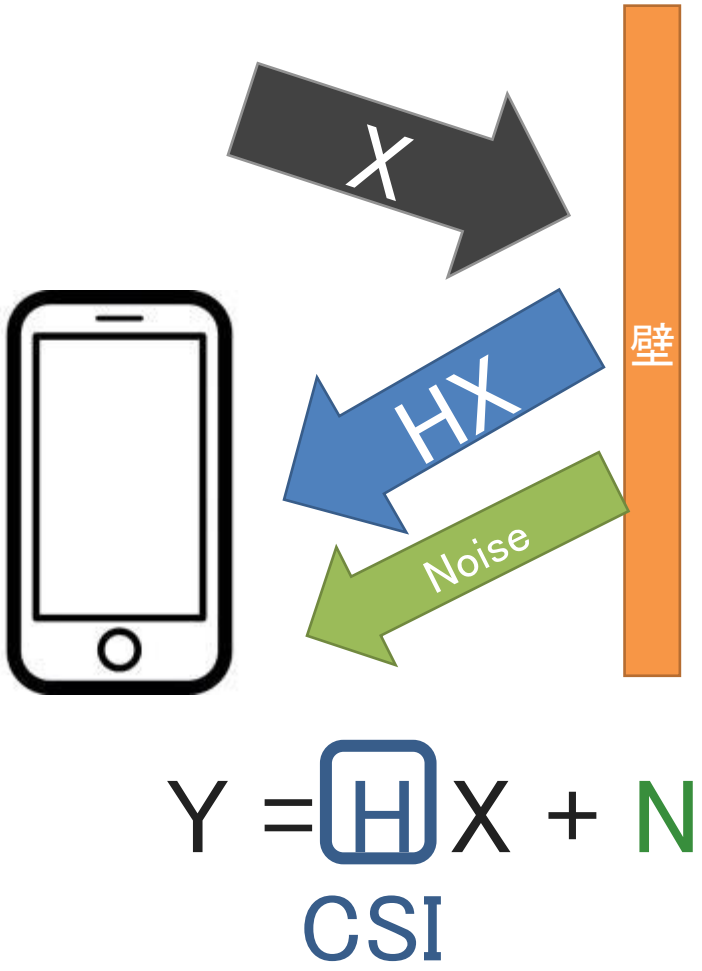
ルータから同じ距離の端末でも RSSI は変わってしまう

CSI (Channel State Information)

より詳細な情報が取れる CSI

- IEEE 802.11 の物理層で定義されている
- どの Wi-Fi デバイスでも取得できる
  - 一般にはドライバの改変が必要
- ルータから取得できる振幅・位相の変位を電波の経路ごと集めた多次元データ

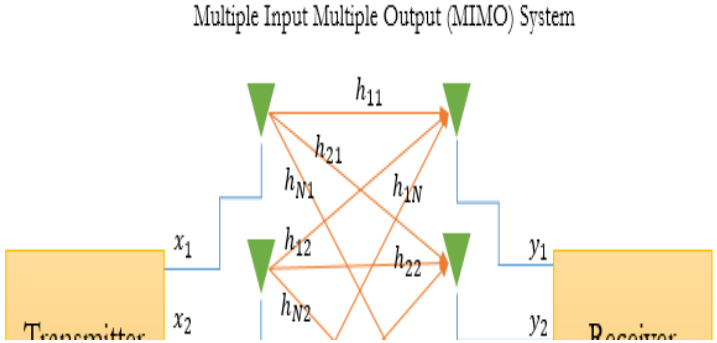
$$\mathrm{H} = \underbrace{\|H\|}_{\text{振幅}} e^{j\underbrace{\angle X}_{\text{位相}}}$$



CSI は多要素数のデータ郡

MIMO (Multiple-Input Multiple-Output)

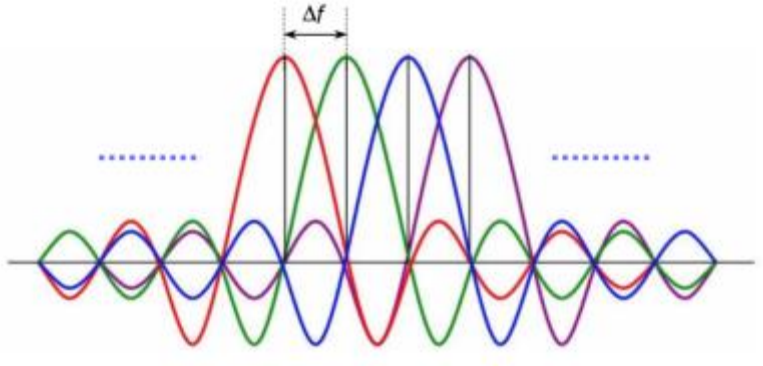
複数のアンテナで通信するため、電波経路が多い



http://www.gaussianwaves.com/2014/08/characterizing-a-mimo-channel/

OFDM (Orthogonal Frequency Division Multiplexing)

複数の周波数帯で通信するため、電波経路が多い



(受信アンテナ数) ×(送信アンテナ数) ×(サブキャリア数) 要素のデータが取得できる

- 例えば，受信アンテナ 3つ，送信アンテナ 1つであれば，
  3 × 1 × 30 = 90 の振幅・位相が取得できる

CNN を使う利点

Wi-Fi 電波による室内センシング

位置測位

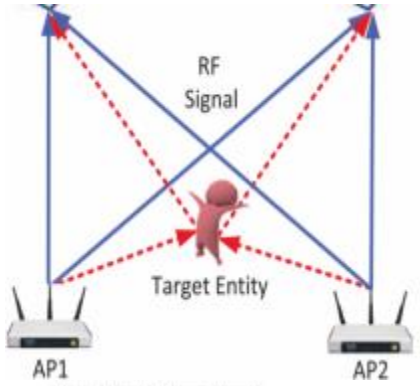
- 室内のどこに人がいるかを推定する

行動認識

- 室内でどのような行動か推定する

個人認識

- 室内に誰がいるか推定する



位置測位の例
（Pilot [Xiao+, 2013]）

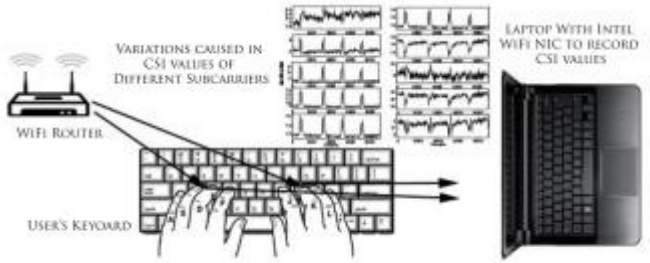
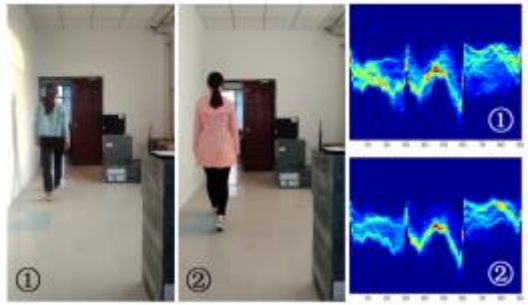


Figure 1: WiKey System

行動認識の例
（WiKey [Ali+, 2015]）



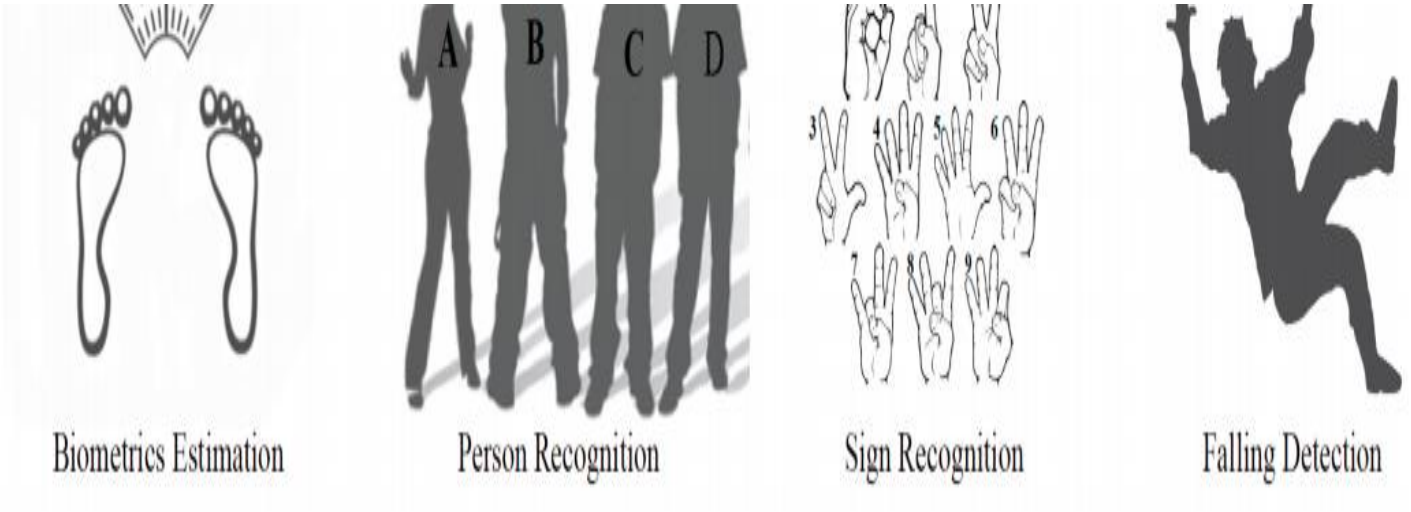
個人認識の例
（WFID [Feng+, 2016]）

[Xiao+, 2013]: Xiao, J., Wu, K., Yi, Y., Wang, L. and Ni, L. M.: Pilot: Passive Device-Free Indoor Localization Using Channel State Information, Proceedings of the 2013 IEEE 33rd International Conference on Distributed Computing Systems, ICDCS '13, Washington, DC, USA, IEEE Computer Society, pp. 236-245 (online), DOI: 10.1109/ICDCS.2013.49 (2013).
[Ali+, 2015]: Ali, K., Liu, A. X., Wang, W. and Shahzad, M.: Keystroke Recognition Using WiFi Signals, Proceedings of the 21st Annual International Conference on Mobile Computing and Networking, MobiCom '15, New York, NY, USA, ACM, pp. 90-102 (online), DOI: 10.1145/2789168.2790109 (2015).
[Feng+, 2016]: Feng Hong, Xiang Wang, Yanni Yang, Yuan Zong, Yuliang Zhang, and Zhongwen Guo. WFID: Passive Device-free Human Identification Using WiFi Signal. In Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services ACM, New York, NY, USA, 47-56. DOI: https://doi.org/10.1145/2994374.2994377

Wi-Fi CSI と CNN

Wi-Fi CSI は多次元データであるため，
何らかの方法で次元を減らしつつ特徴量を掴む必要がある

今まで目的によって特徴量の計算方法が多種多様であったが，
CNN から共通の処理で特徴量を得る研究が出てきた [Wang+, 2018]

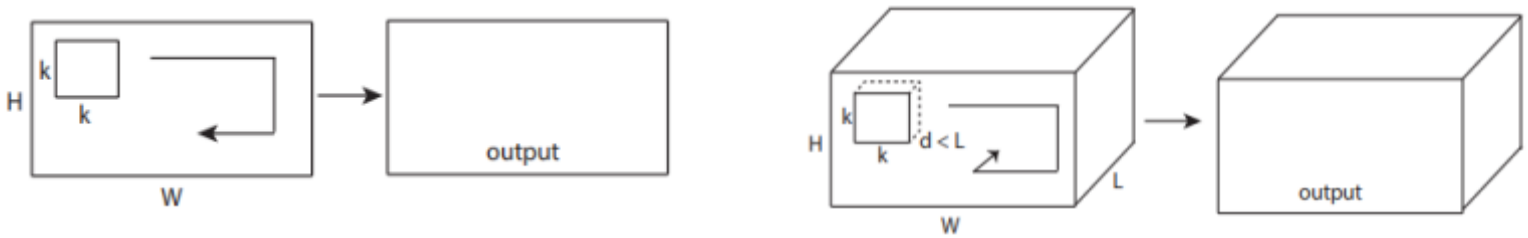


[Wang, 2018]: Wang, F., Han, J., Zhang, S., He, X. and Huang, D.: CSI-Net: Unified Human Body Characterization and Action Recognition, arXiv preprint arXiv:1810.03064 (2018).

ジェスチャ認識にフォーカスする理由

CNN は時系列情報が欠落する

- 既存手法 [Wang+, 2018] の CNN は時系列情報が欠落している
  - 瞬間の状態しか推測できない（個人認識，サイン認識 etc.）
- 時系列情報が必要になる認識としてジェスチャ認識にフォーカス
- 時系列方向にも畳込み処理をする 3次元 CNN [Tran+, 2015] を採用
  - 主に動画の内容推定に利用される手法



2次元 CNN と 3次元 CNN の違い [Tran, 2015]

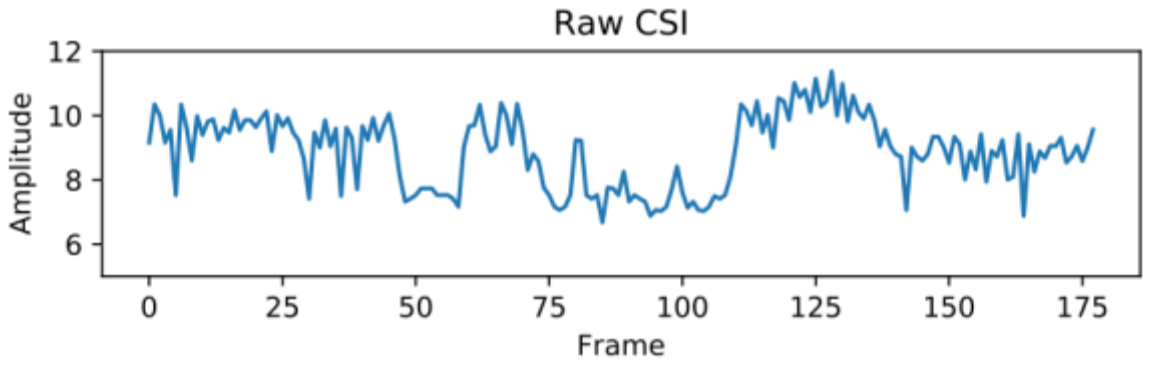
[Wang+, 2018]: Wang, F., Han, J., Zhang, S., He, X. and Huang, D.: CSI-Net: Unified Human Body Characterization and Action Recognition, arXiv preprint arXiv:1810.03064 (2018).
[Tran+, 2015]: Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M.: Learning Spatiotemporal Features with 3D Convolutional Networks, Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, Washington, DC, USA, IEEE Computer Society, pp. 4489–4497 (online), DOI: 10.1109/ICCV.2015.510 (2015).

実装

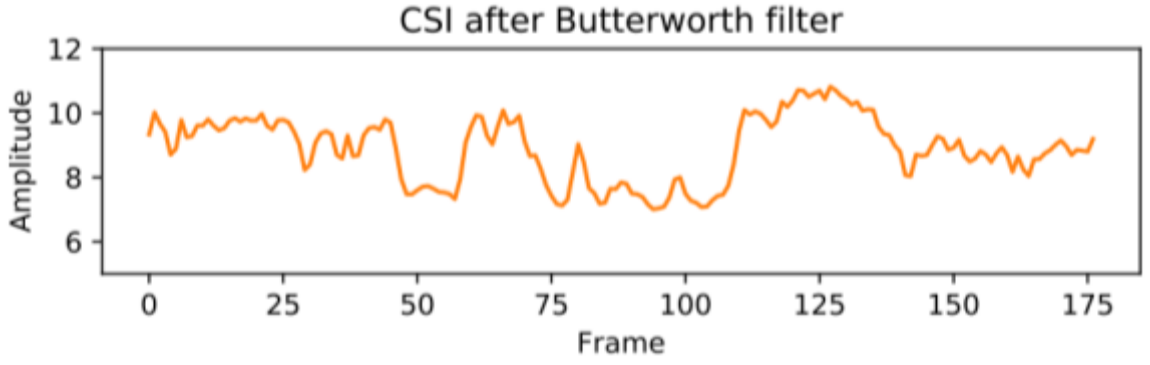
ノイズ処理

Wi-Fi CSI を CNN で学習する先行研究 [Wang+, 2018] にて,
ローパスバターワースフィルタによるノイズ処理が
CNN による学習結果を向上させると報告されている

- 本提案手法においても同様の傾向が見られたため、
  ローパスバターワースフィルタを採用する



取得した CSI（ノイズ処理前）



ローパスバターワースフィルタ後

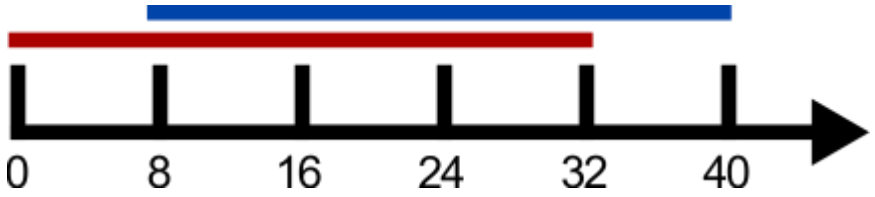
[Wang+, 2018]: Wang, F., Han, J., Zhang, S., He, X. and Huang, D.: CSI-Net: Unified Human Body Characterization and Action Recognition, arXiv preprint arXiv:1810.03064 (2018).

データサンプルの作成

CSI は 0.01 sec ごとに取得する

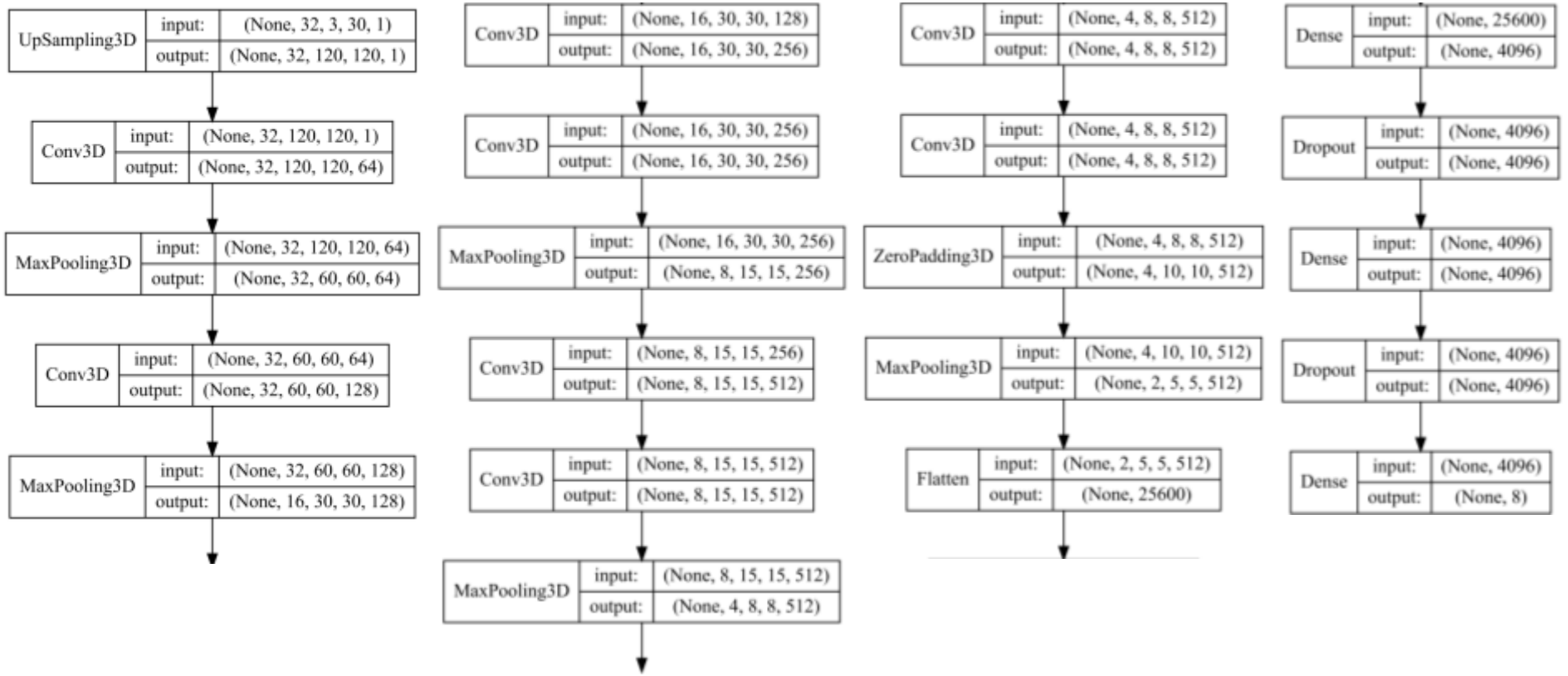
取得した CSI は，一定フレーム N_{frame} ごとで 1 sample にする

- ただし，サンプルを作る際は，データの連続性を保つため
  8 frame ずつシフトさせる
- 例えば，$N_{frame} = 32$ のとき，
  最初のサンプルは 1 ～ 32 frame をまとめ，
  次のサンプルは 9 ～ 40 frame をまとめる



1 sample は $\boldsymbol{N_{frame} \times 3_{[routes]} \times 30_{[subcarriers]}}$ のデータとなる

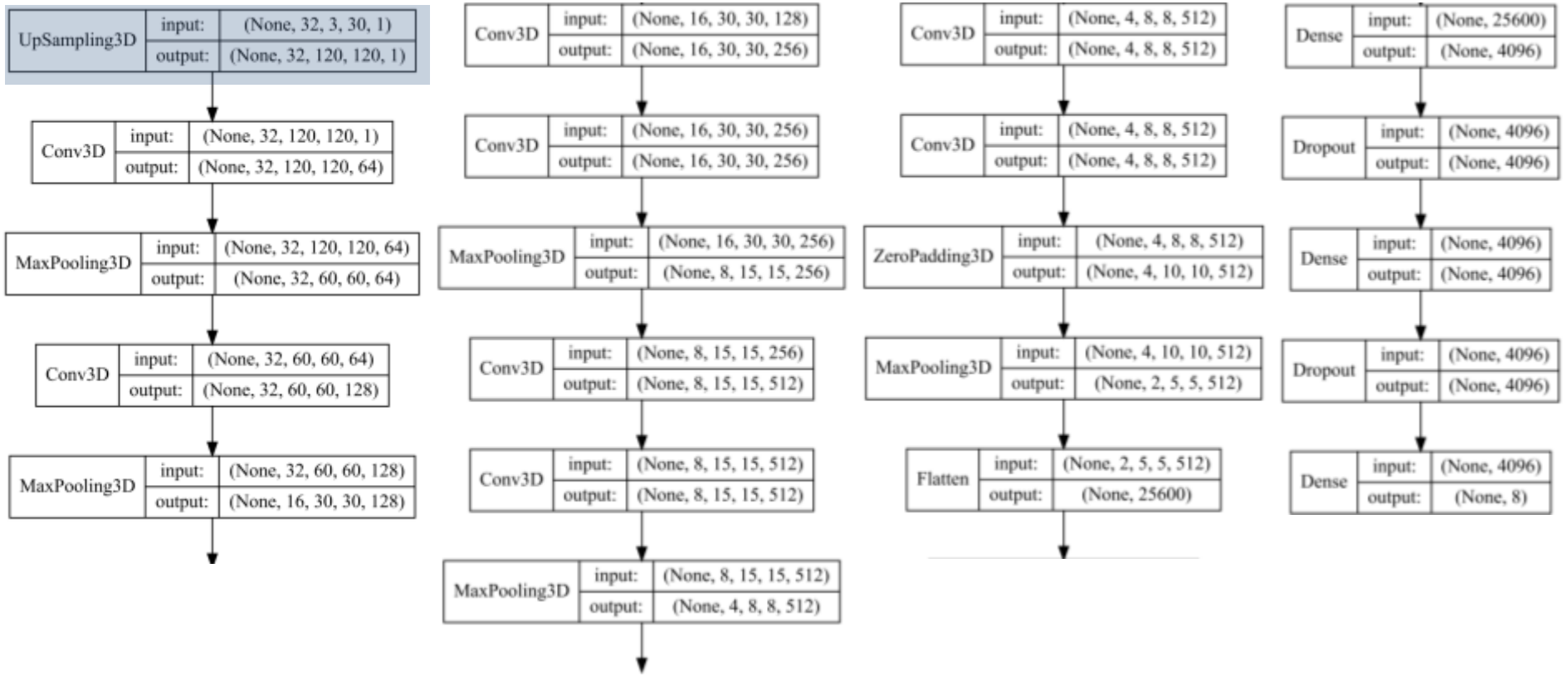
CNN モデル概要



動画認識の先行研究 C3D [Tran+, 2015] にならい，パラメタを設定
実装は Keras + Tensorflow で行った

[Tran+, 2015]: Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M.: Learning Spatiotemporal Features with 3D Convolutional Networks, Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, Washington, DC, USA, IEEE Computer Society, pp. 4489–4497 (online), DOI: 10.1109/ICCV.2015.510 (2015).

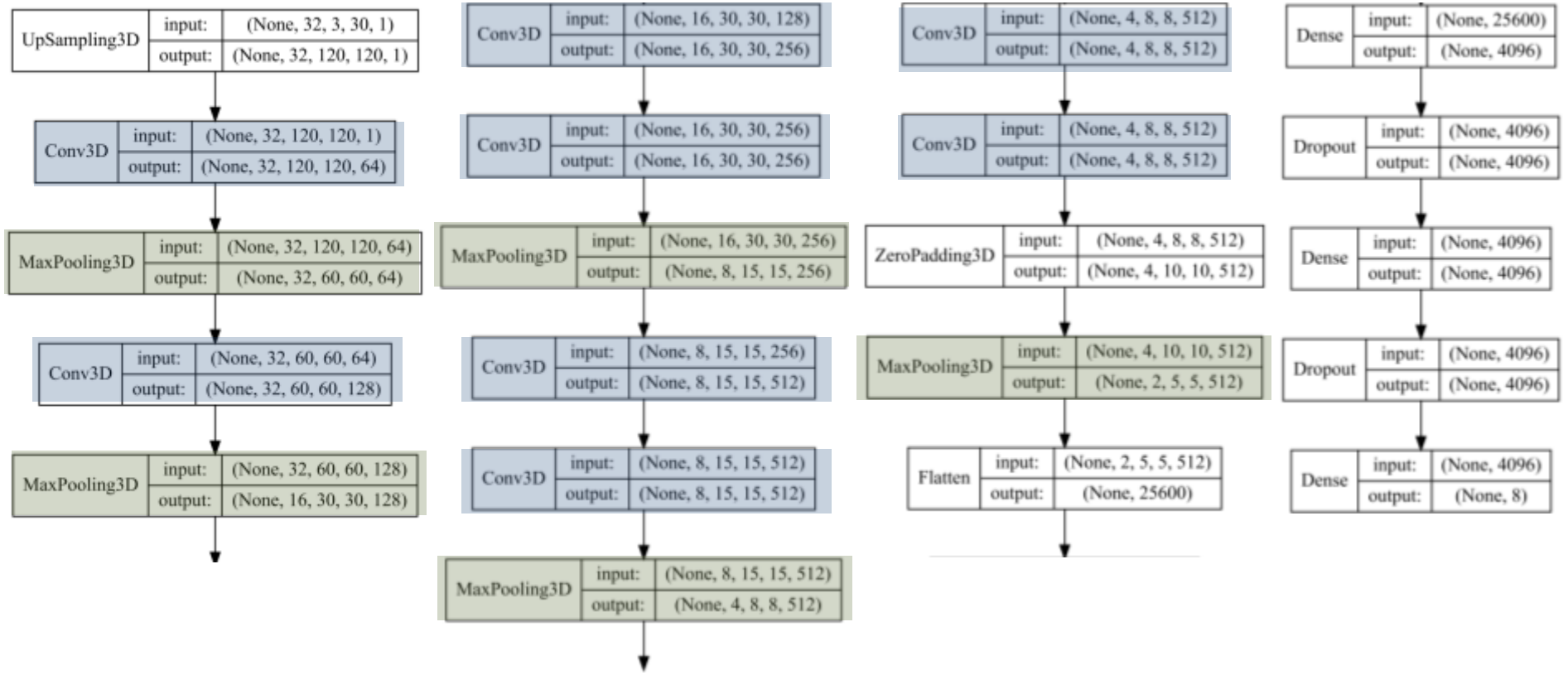
アップサンプリング



畳込みをするためには次元数が少ないため，

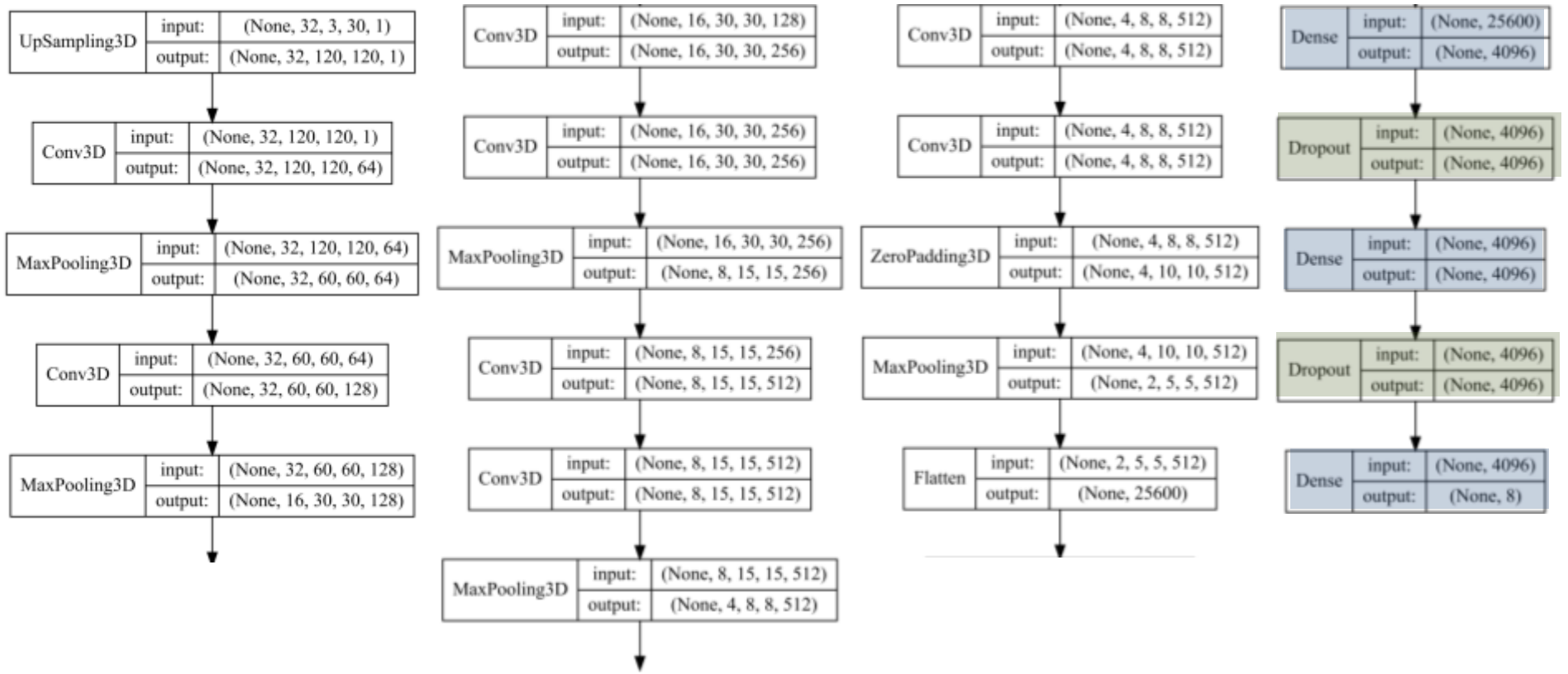
N_{frame} × 120 × 120 になるようにアップサンプリングをする

畳込み と Pooling



活性化関数 ReLU，ゼロパディングを有効にして 3次元に畳込みをする（青色）
Max Pooling 方式による Pooling を行う（緑色）

Dense と Dropout

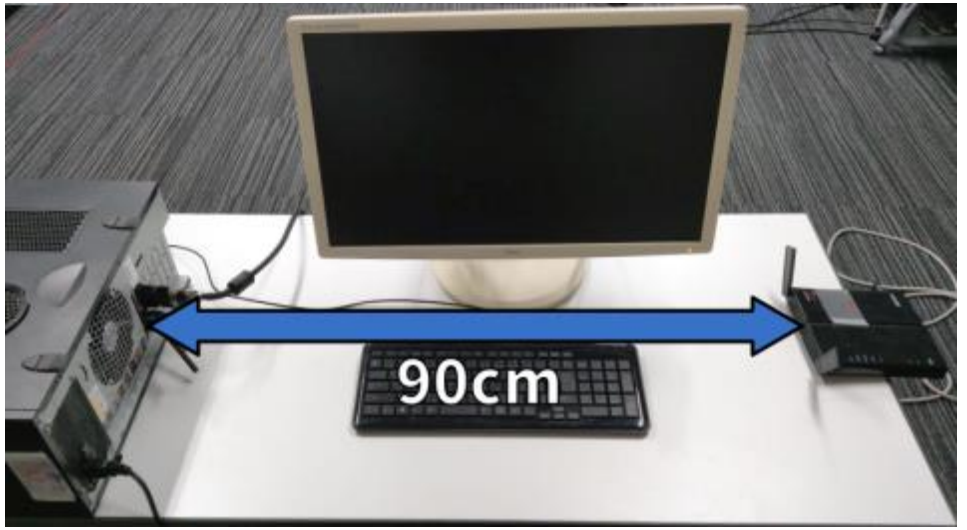


Dense では，活性化関数 ReLU で全結合処理をする（青色）
Dropout では，過学習防止のため 0.5 の確率でデータを無効化する（緑色）

評価実験

実験概要

- 23～24歳の男性5名のデータ
- 送受信機の中間でジェスチャをする
  - 送受信機間は 90 cm
- 1ジェスチャにつき 50 回取得する
  - うち 90% を学習データ，
    10% を検証データとする
- 試行前にジェスチャを実際に見せた



8つのジェスチャ

Wi-Fi CSI によるジェスチャ認識の先行研究 WiFinger [Tan+, 2016] にならったジェスチャ

[Tan+, 2016]: Tan, S. and Yang, J.: WiFinger: Leveraging Commodity WiFi for Fine-grained Finger Gesture Recognition, Proceedings of the 17th ACM International Symposium on Mobile AdHoc Networking and Computing, MobiHoc '16, New York, NY, USA, ACM, pp. 201-210 (on-line), DOI: 10.1145/2942358.2942393 (2016).

予備実験｜最適なフレーム数

最適なフレーム数を調べるために参加者 A のデータで学習をさせた

- 1 sample のフレーム数は $N_{frame} = \{16, 24, 32\}$ の3つで検証

結果，フレーム数が多いほど精度が向上した

- 24 frames と 32 frames では，正解率 0.975 以上となった
- 32 frames のほうが epoch 数（学習回数）が少ないため，採用した

	サンプル数 (学習/検証)	Early Stopping	正解率	損失
16 frames	6359/717	4 epoch	0.845	0.227
24 frames	5951/673	6 epoch	0.975	0.092
32 frames	5539/627	5 epoch	0.979	0.078

評価実験1｜自分のデータへの適合

自分のデータのみで学習させた場合，平均正解率 0.978 であった

- 一番精度が低い実験参加者でも，正解率 0.932 であった
- 既存手法 [Tan+, 2016] は，正解率 0.93 以上であり，同程度の精度を実現

	サンプル数 (学習/検証)	Early Stopping	正解率	損失
実験参加者 A	5539/627	5 epoch	0.979	0.078
実験参加者 B	4727/555	7 epoch	1.000	0.000
実験参加者 C	7590/820	7 epoch	0.978	0.030
実験参加者 D	6065/676	6 epoch	0.932	0.151
実験参加者 E	4966/551	6 epoch	1.000	0.048

[Tan+, 2016]: Tan, S. and Yang, J.: WiFinger: Leveraging Commodity WiFi for Fine-grained Finger Gesture Recognition, Proceedings of the 17th ACM International Symposium on Mobile AdHoc Networking and Computing, MobiHoc '16, New York, NY, USA, ACM, pp. 201–210 (on-line), DOI: 10.1145/2942358.2942393 (2016).

評価実験1｜自分のデータへの適合

ジェスチャの類似性と誤認識は関係しない

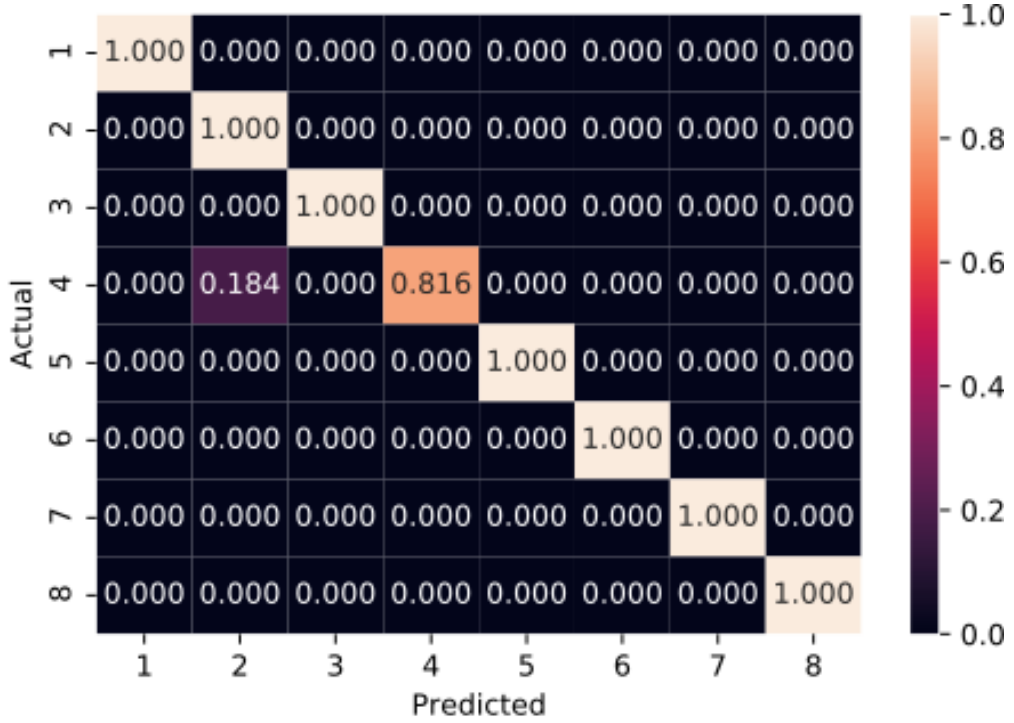
- 実験参加者 C では，
  動作が全く違うジェスチャが誤認識された

2) Zoom In

4) Circle Right

どの実験参加者でも誤認識は 1つのみ

- 連続して認識にかければ
  さらに高水準で予測できる可能性がある



実験参加者 C

評価実験2｜他人のデータへの適合

他人のデータで学習させた場合，平均正解率 0.184 であった

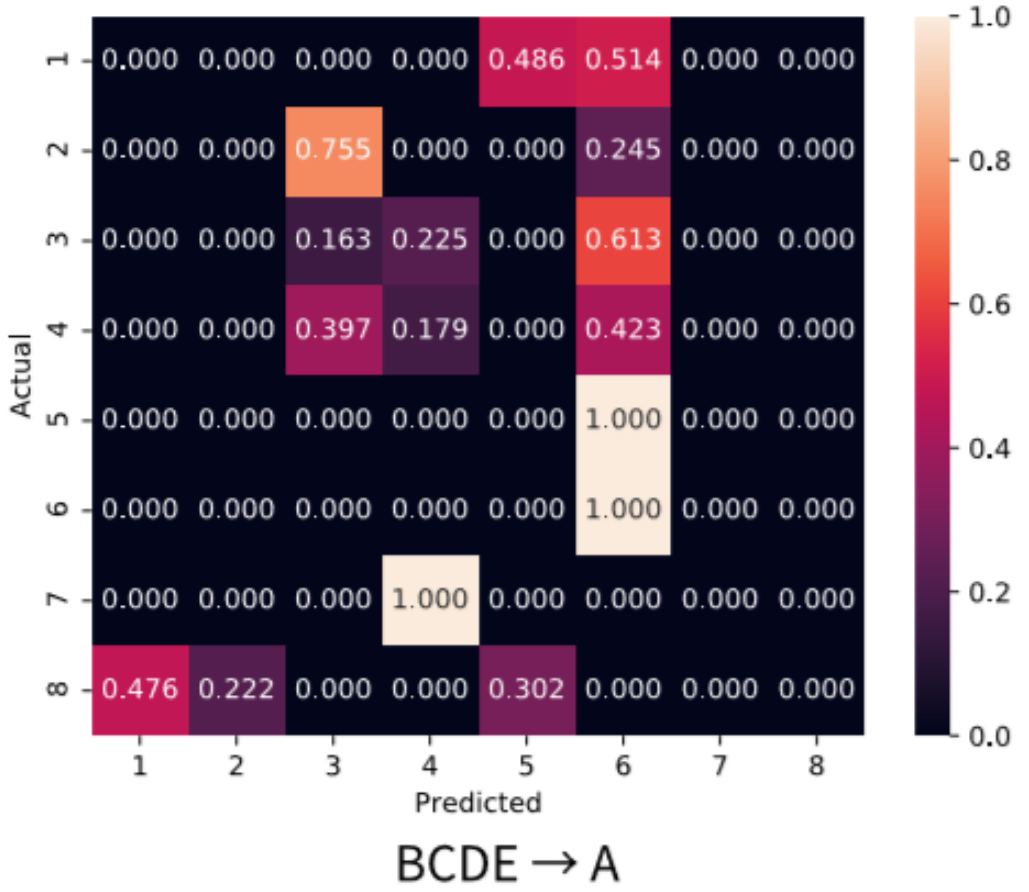
- 学習時には自分以外の 4 名のデータで学習させる
  例えば，実験参加者 ABCD のデータから，実験参加者 E の動きを推定する

学習 → 評価	Early Stopping	正解率（学習時）	損失（学習時）	正解率（評価時）
BCDE → A	6 epoch	0.980	0.081	0.168
ACDE → B	8 epoch	0.964	0.077	0.174
ABDE → C	5 epoch	0.930	0.135	0.230
ABCE → D	7 epoch	0.968	0.067	0.296
ABCD → E	6 epoch	0.962	0.089	0.050

評価実験2｜他人のデータへの適合

実験参加者 A は,
ジェスチャ 5・7 において
間違ったジェスチャで
誤答率 1.000

- 極端な誤答が目立つ結果



BCDE → A

評価実験2｜他人のデータへの適合

自分のデータを含めば，他人のデータを混ざっていても精度は高い

- 4人の学習モデルに対して，4人のテストデータは9割を超えているため，他人のデータを混ぜることによる影響は少ないと考えられる

学習 → 評価	Early Stopping	正解率（学習時）	損失（学習時）	正解率（評価時）
BCDE → A	6 epoch	0.980	0.081	0.168
ACDE → B	8 epoch	0.964	0.077	0.174
ABDE → C	5 epoch	0.930	0.135	0.230
ABCE → D	7 epoch	0.968	0.067	0.296
ABCD → E	6 epoch	0.962	0.089	0.050

考察

学習データの取得

学習に必要なデータを取るためには，時間が必要

- 提案手法では，50 回 × 8 ジェスチャのデータが必要
- 今回の実験では，ひとり当たり 30 分程度でデータを集めた

学習済みモデルと組み合わせることで，
必要なデータ量を減らすことができる

- 他人のデータが多くても自分のデータを含めれば，精度は高い
- しかし，他人のデータのみの精度は著しく低いため，
  どの程度の個人データを必要とするかは調査が必要

ジェスチャの動作速度

ジェスチャの動作速度はひとによって多少異なる

- 速度によって 1 sample に含まれる
  特徴量が変わってくる可能性がある

今回の実験での差異程度であれば，精度を高く対応できた

- 今回は最大でも 0.5 sec 程度の差だった
- 実験前にジェスチャを見せたことで，
  動作時間を揃えられたのではないか

	平均ジェスチャ時間 (ms)
実験参加者 A	1105
実験参加者 B	962
実験参加者 C	1420
実験参加者 D	1231
実験参加者 E	1001

アプリケーションへの活用

2台の Wi-Fi デバイスさえあれば，操作できる空間を即座に作れる

- 電波のみで識別できるため，公共の場でも使いやすい

既存の CNN による CSI センシング手法との併用も可能

- CSI-Net [Wang+, 2018] による個人認識などと組み合わせれば，
  同じジェスチャでも，ユーザごとに違う処理を割り当てることができる

[Wang+, 2018]: Wang, F., Han, J., Zhang, S., He, X. and Huang, D.: CSI-Net: Unified Human Body Characterization and Action Recognition, arXiv preprint arXiv:1810.03064 (2018).

結論

3次元 CNN を利用して，時系列情報を保持した
Wi-Fi CSI によるジェスチャ認識手法を提案した

評価実験により，以下のことがわかった

- 1 sample に含めるフレーム数は多いほうが精度が高い
- 自分のデータによる学習であれば，正解率 0.932 以上で認識できる
- 他人のデータのみで学習すると，著しく精度が落ちてしまう
  - ただし，自分のデータが含まれていれば，
    他人のデータの影響を受けることは少なく，精度は高いままであった