CH-004

BLE ビーコンを利用したマルチモーダル畳み込みニューラルネットワーク
Multimodal Convolutional Neural Networks Using BLE Beacon

橋本 和幸[1)]　堀川 三好[1)]　岡本 東[1)]
Kazuyuki Hashimoto　Mitsuyoshi Horikawa　Azuma Okamoto

1　はじめに

実社会におけるスマートフォンの利用は多岐に渡り，端末に搭載されたセンサを活用した広告サービスやナビゲーションシステムの構築がされている．例えば，駅構内や商業施設に Bluetooth Low Energy ビーコン（以下，ビーコン）を設置し，広告やナビゲーションを行う実証実験が行われている．スマートフォンに搭載されているセンサを利用したマーケティングへの応用は，今後も様々な分野での活用が期待される．

本研究はスマートフォンで受信されるビーコン受信強度 (RSSI:Received Signal Strength Indication) とカメラで得られる画像を用いて，物体をインスタンスレベルで認識し，関連する情報を表示するための技術開発を目的とする．そのため，異なるドメインにあるデータを統合的に扱う方法として，マルチモーダル学習を取り入れる．近年の深層学習の発展に伴い，マルチモーダル学習は MLP(Multi-layer Perceptron) や CNN(Convolutional Neural Network) を中心技術に取り入れ，字幕生成や翻訳など様々な場面で応用されている．本稿ではマルチモーダルな画像認識モデルとして，ビーコンの RSSI とカメラで撮影した画像を併せて学習することにより，現実空間に配置されている物体を高精度かつ容易に認識するための方法を提案する．この技術を活用することで，広告やポスター以外の物体についてインスタンス認識を行い，EC サイトへの誘導が可能となり，新たな購買機会の創出やオムニチャネル化の一助になることが期待される．

本研究に先立ち，著者ら [1] は画像認識の高性能化を目的とし，ビーコンの RSSI とカメラで撮影した画像を組み合わせたマルチモーダル学習によるインスタンス認識モデルを提案している．しかしながら，訓練データの収集には労力を要する．そのため，本稿では動画撮影からフレーム分割して画像生成を行うと同時に，収集した RSSI を統計モデルに基づいてデータオーギュメンテーションを行う手法を取り入れる．また，ビーコンが設置された空間において類似商品を配置した場合に，提案手法を取り入れる妥当性と分類精度を検証実験とし報告する．提案手法は従来の画像認識モデルに比べて高い分類精度を示し，データの収集に要する時間を大幅に抑える事ができた．

2　関連研究

2.1　マルチモーダル学習

Baltrusaitis ら [2] はマルチモーダル学習を表現学習 (Representation)，変換 (Translation)，アラインメント (Alignment)，融合 (Fusion)，および共学習 (Co-Learning) の 5 つに分類している．本研究と関連する表現学習と融合の 2 つについて以下に述べる．

1)　岩手県立大学大学院ソフトウェア情報学研究科

2.1.1　表現学習

表現学習は目的に適した特徴量を学習を通じて自発的に獲得するアプローチである．各モーダル情報ごとに表現学習が進められてきたが，近年マルチモーダル情報に対する表現学習の研究が進められている．その手法は Joint Representations と Coordinated Representations に大別される．Joint Representations は各モーダル情報を同じ表現空間に結合する手法である．また，Coordinated Representations は各モーダル情報を別々に処理するが互いに制約を与えることによりコーディネートされた空間を得る手法である．Ngiam ら [3] は動画に映る唇の動きから話された単語を予測するための表現を，Kahou ら [4] は話者の感情を予測するための表現を学習するモデルを提案している．

2.1.2　融合

融合は，2 つ以上のモーダル情報を結合して予測を行うことである．異なるモーダル情報を組み合わせることで，様々な予測が可能になると期待されている．また，深層学習の分野では表現学習と融合は密接な関係がある．例えば，教師あり学習におけるニューラルネットワークの最後の層はソフトマックス回帰分類器などの線形分類器であり，ネットワークの残りの部分はこの分類器に与える表現を学習している．つまり，表現学習は良い表現を得ることを目的とするが，その表現を融合による予測をもとに得ることも可能である．

マルチモーダル融合には 3 つの利点がある [2]．まず，同じ現象を観察する複数のモダリティへのアクセスを有することで，より堅牢な予測が可能になる．次に複数のモダリティへのアクセスを持つことで，独自のモダリティでは見えないような補足情報を取得することができる．そして，モダリティの 1 つが欠落している場合でも，マルチモーダルシステムは動作することである．融合を行っている例として，画像と自然言語を学習させることにより，Gao ら [5] や Malinowski ら [6] は画像に対する質問を自然言語で返答するモデルを，Vinyals ら [7] は画像に対して説明を付与するモデルを提案している．

2.2　データオーギュメンテーション

機械学習では大量のデータが求められるが，様々な要因から収集が困難な場合がある．そのため，既存のデータをもとに訓練データを増やす，データオーギュメンテーションを用いる．例えば，画像を入力とする際は入力画像の回転，クリッピング，左右反転などが利用される．また画像を部分的に隠す方法 [8][9] や 0 ~ 1 で表現されたピクセル強度を，それぞれに等しい確率で 1 を選択し二値で各ピクセルを置き換える，つまり数字をネットワークに渡すたびに再サンプリングする手法 [10] がある．一方センシングデータに関しては観測値に僅かな誤差を与える方法，時系列データであれば時間方向にずらす方法 [11] および部分的に拡大縮小する方法 [12] が取られている．

3　データセット

3.1　実験環境

本稿は，物体のインスタンス認識を行うことでマーケティングへ応用することを目的としている．そのため，図 1 に示すビーコンが設置された空間において類似商品（図 2:紙パック飲料）を配置し，インスタンス認識する実験を行う．

3.2　データの収集方法

2 つの方法でデータを収集する．収集方法 1 は画像と RSSI を同時に 1 枚ずつ収集する方法である．収集方法 2 は **3.3** で説明するデータオーギュメンテーションの使用を念頭に，動画と RSSI を収集する方法である．

(1) 収集方法 1

ビーコンの RSSI と写真を同時に収集する．学習時に読み込みやすいように，撮影した写真のファイル名とビーコンの ID と RSSI を記録したファイルが紐付けられるファイル名にする．この方法では一分間に 10 枚の画像を得ることができる．

(2) 収集方法 2

物体がフレームに入るように動画を一分間撮影し，同時に各ビーコンの RSSI を収集する．この方法では，一分間に 500 枚の画像を得ることができる．撮影した動画はフレーム分割し，画像として扱う．また，RSSI については，収集した RSSI を用いて以下に述べるデータオーギュメンテーションを行う．

3.3　データオーギュメンテーション

各物体の撮影時に収集した RSSI の分布を図 3 に示す．シャピロ・ウィルク検定における p 値（表 2），歪度（表 3），尖度（表 4），標準偏差（表 5）および各ビーコンの平均（表 6）を示す．物体撮影時の RSSI からシャピロ・ウィルク検定における p 値を求めると，物体 B 撮影時のビーコン 2 以外は $p < 0.01$ であり，統計的有意である．歪度は全て負の数であるため，全体的にやや右に偏っている傾向があるが標準偏差は 4.999 ~ 6.494 の値を取る．従ってガウス分布に近似できると言える．

各撮影物体 A，B，C，D に対して，対応付けられる RSSI を r_A, r_B, r_C, r_D とする．各 RSSI ベクトルはビーコン ID を表す 1，2，3，4 を用いて $r_x = [r_{x_1}, r_{x_2}, r_{x_3}, r_{x_4}]$ となる．これを $r_x \sim \mathcal{N}(\mu_{x_{id}}, \sigma^2_{x_{id}})$ として各 ID ごとに RSSI を生成する．

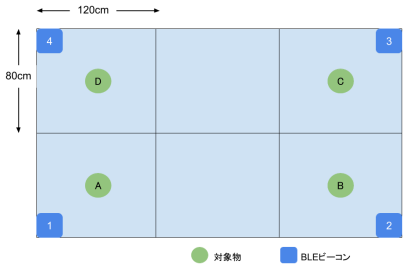

図 1: データを収集するための空間

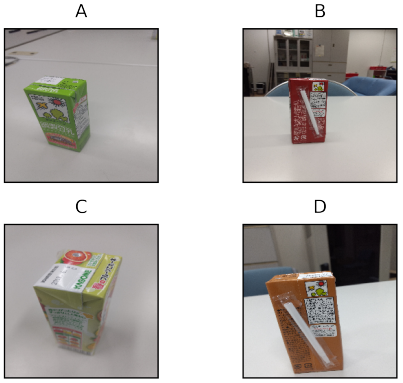

図 2: 分類するために収集したデータの一部とラベル

3.4　データフレームの作成

収集したデータは Pandas を用い，データフレームとしてメモリに展開する．収集方法 1 で集めたデータはファイル名から対応する画像と RSSI を選び，対応するラベルのついたデータフレームを作成する．これをデータフレーム 1 とする．収集方法 2 で集めたデータはまず，ラベルごとに，各ビーコンの平均と分散を求める．ラベルに対応する RSSI を学習時に毎回サンプリングさせるため，ここでは画像とそのラベルがわかるデータフレームを作成する．これをデータフレーム 2 とする．

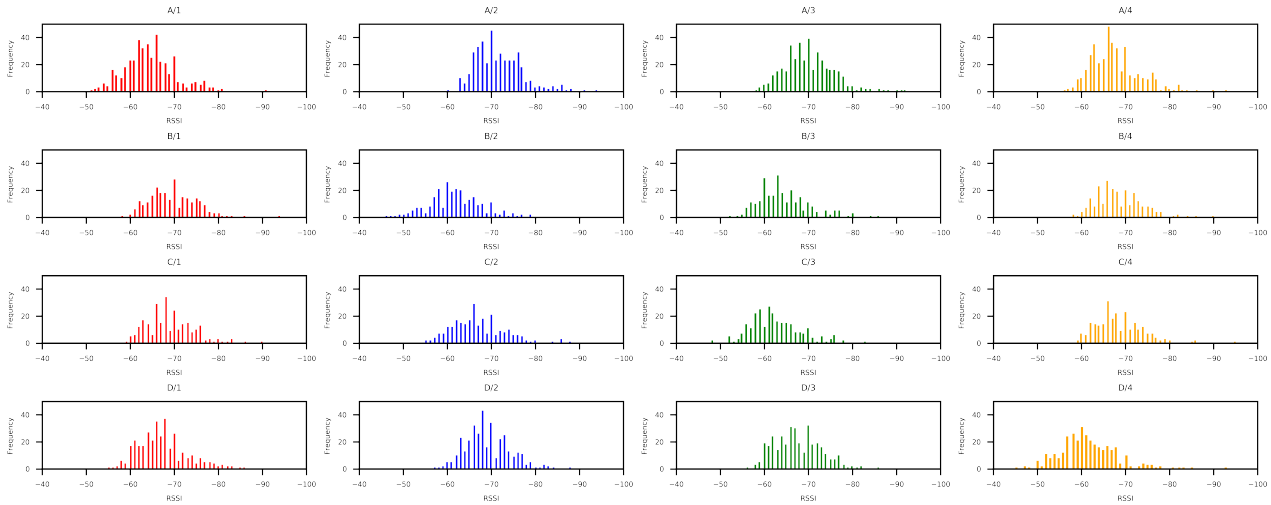

図 3: RSSI の分布

表 1: それぞれの方法で収集した枚数

		場所			
		A	B	C	D
撮影枚数	収集方法 1	92	98	93	96
	収集方法 2	519	468	480	471

表 2: シャピロウィルク検定による p 値

		物体位置			
		A	B	C	D
Beacon	1	2.32E-06	5.67E-05	7.34E-06	2.00E-07
	2	1.89E-10	0.09	2.29E-05	8.58E-05
	3	3.15E-07	8.73E-08	3.97E-05	8.71E-05
	4	1.75E-10	1.13E-05	7.69E-08	1.07E-07

表 3: 歪度

		物体位置			
		A	B	C	D
Beacon	1	-0.568	-0.613	-0.706	-0.705
	2	-0.901	-0.203	-0.678	-0.509
	3	-0.694	-0.892	-0.592	-0.400
	4	-0.947	-0.751	-1.004	-0.828

表 4: 尖度

		物体位置			
		A	B	C	D
Beacon	1	0.913	1.037	0.807	0.499
	2	1.200	0.217	0.740	0.347
	3	1.046	0.947	0.517	-0.162
	4	1.509	1.12	2.446	2.109

表 5: 標準偏差

		場所			
		A	B	C	D
Beacon	1	5.763	5.267	5.599	5.548
	2	5.292	5.998	5.705	5.156
	3	5.212	5.959	5.687	5.287
	4	5.531	4.999	5.198	6.494

表 6: RSSI の平均

		撮影場所			
		A	B	C	D
Beacon	1	-64.641	-71.708	-70.143	-67.589
	2	-69.648	-61.872	-64.400	-68.454
	3	-68.976	-66.914	-62.552	-68.460
	4	-67.250	-69.091	-67.587	-61.521

4 画像認識モデル

本稿の先行研究 [1] では，ビーコンの RSSI とカメラで撮影した写真を用いてマルチモーダル学習を行い，インスタンス認識を行うために，表 7 に示す 3 つのモデルを提案した．実験の結果，最も高い精度を得られたのは mCNN-c であった．しかし，学習にかかる時間や安定性については mCNN-f が優れていた．mCNN-f は式 1 のように記述され，確率的勾配降下法によりニューラルネットワーク f のパラメーター θ を決定する．mCNN-f モデルを図 4 に示す．x は画像，s は RSSI，$\|$ はテンソルの連結を表す．

$$f_\theta(x,s) = f'_{\theta_0}\left(f''_{\theta_1}(x) \| f'''_{\theta_2}(s)\right) \quad (1)$$

5 実験

データセットに対して前処理を行い，各手法で収集したデータセットの分類精度を以下にまとめる．

5.1 前処理

画像は 0 から 1 の範囲を取るよう 255 で割る．Kudou ら [14] は，ビーコンの RSSI は端末によって異なりやすいと述べている．そのため，0 から 1 の範囲を取るよう

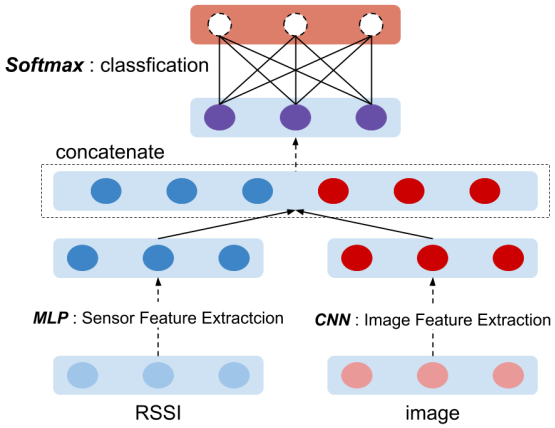

図 4: mCNN-f モデル　各ドメインごとに特徴を抽出し，それをもとに分類を行う

表 7: 環境センシングを活用した 3 つのマルチモーダルなインスタンス認識モデル

モデル名	概要
mCNN-w	各センサに対して重みをつけ初期融合するモデル
mCNN-c	画像のチャンネルをセンサ値で拡張し CNN を行うモデル
mCNN-f	画像には CNN をセンサ値には MLP を用いて後期融合するモデル

前処理を行う．まず，同時に観測された RSSI をその中で最も小さい値で差を求め，次にその差をその中で最も大きな値で割る．

5.2 実験設定

実験はデータフレーム 1 で学習する mCNN-f モデル，データフレーム 2 の画像のみから学習する CNN，データフレーム 2 で学習する mCNN-f モデルを用いる．データフレーム 2 の動画からフレーム分割した画像はデータフレーム 1 のために撮影した画像に比べ，枚数の面から，より網羅的に集められている．そのため，データフレーム 1 の画像のみから学習する CNN は今回実験の対象としない．mCNN-f と CNN のハイパーパラメータは入力層から出力層の 2 つ前まで同じになるよう設定する．各モデルでミニバッチ学習を行い，毎回学習後にテストデータを用いて分類精度を調べる．これを 7 回繰り返し，その平均を取る．データフレーム 2 の画像の枚数はデータフレーム 1 のおおよそ 5 倍ある．おおよその訓練回数を合わせるため，データフレーム 1 は 500 回，データフレーム 2 は 100 回行う．これにより，以下の 2 点について調べる．

- 各条件におけるモデルの分類精度
- データオーギュメンテーションを行ったデータセットを用いた mCNN-f の妥当性

5.3 実験結果

データフレーム 1 を学習する mCNN-f，データフレーム 2 の画像を学習を行う CNN，データフレーム 2 を学習する mCNN-f の比較を行い，その結果を図 5 から 7 に示す．

データフレーム 1 を学習した mCNN-f については，最初はばらつきがあるものの，学習回数は 200 回前後で収束しており，分類精度は 90%前後である．次にデータフレーム 2 の画像を学習した CNN は，学習回数 30 回前後で精度の平均が 80%前後で横ばいになっており，誤差範囲が 5%前後ある．最後にデータフレーム 2 を学習した mCNN-f について，学習回数は 40 回前後で 90%の分類精度があり，誤差範囲は 1%前後である．

5.4 考察

画像のみで分類を試みるモデルに比べ，データフレーム 1 を学習した mCNN-f とデータフレーム 2 を学習した mCNN-f はより適合したモデルであると言える．また，データフレーム 1 の収集に必要な時間は各物体に対して 10 分前後なのでおおよそ 40 分であり，データフレーム 2 の収集に必要な時間は各物体に対して 1 分なので 4 分である．データフレーム 1 は各 90 枚前後であったが，これを各 400 枚前後集めるにはおおよそ 160 分前後は必要となる．データフレーム 1 はデータ数の少ないデータセットであるが，これを学習した mCNN-f はデータフレーム 2 で学習した mCNN-f と同じくらい分類精度が高い．これは撮影位置や保持姿勢に影響があったとも考えられる．そのため，収集方法 1 でのデータセット作成がより良い選択肢に考えられるが，収集にかけられる時間の問題から収集方法 2 は実際に運用する面で妥当であると考えられる．

6 まとめ

本稿では RSSI を対象としたデータオーギュメンテーションにより，データ収集コストの削減を目的とした実験を行った．評価実験にはマルチモーダルな画像認識モデルを用いた．各モデルの分類精度を比較すると，データフレーム 2 で学習した mCNN-f の分類精度が 90%となり，従来の CNN の精度を上回る結果を示した．また，データの収集に要する時間をおよそ 50 分の 1 に短縮できることがわかった．今後は実用化に向けた実証実験及びシステム構築を行う．

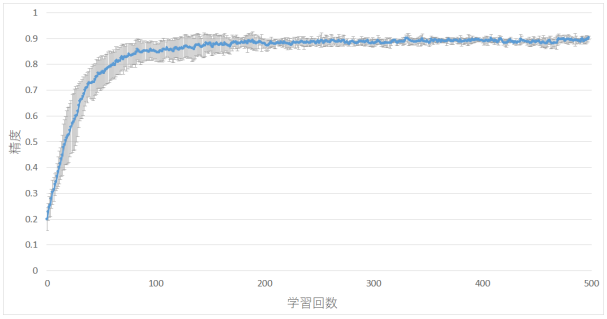

図 5: データフレーム 1 で学習する mCNN-f での各回における精度の平均と誤差の推移

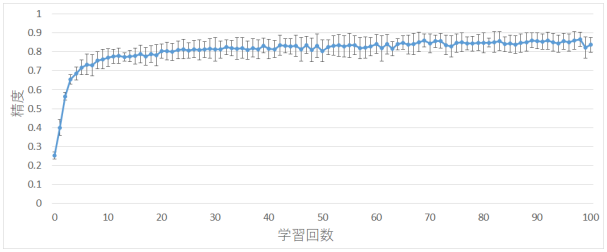

図 6: データフレーム 2 の画像を使って学習する CNN の各回における精度の平均と誤差の推移

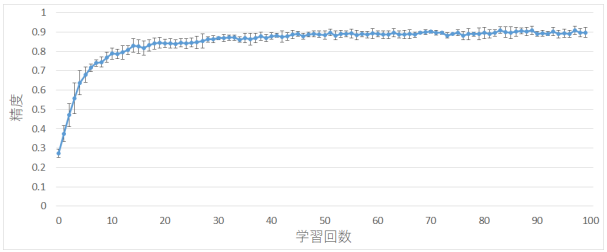

図 7: データフレーム 2 で学習する mCNN-f での各回における精度の平均と誤差の推移

参考文献

[1] 橋本 和幸，堀川 三好，岡本 東，環境センシングを活用したマルチモーダル学習

[2] Baltrusaitis T., Ahuja C., and Morency L.P.: Multimodal machine learning: A survey and taxonomy., PAMI(2018)

[3] Ngiam J., Khosla A, Kim M., Nam J., Lee H., and A. Y. Ng, “Multimodal Deep Learning,” ICML(2011)

[4] Kahou S. E., Bouthillier X., Lamblin P., et al.: EmoNets: Multimodal deep learning approaches for emotion recognition in video, Multimodal User Interfaces(2015)

[5] Gao H., Mao J., Zhou J., et al.: Are you talking to a machine? dataset and methods for multilingual image question answering, NIPS(2015)

[6] Malinowski M., Rohrbach M., and Fritz M.: Ask your neurons: A neural-based approach to answering questions about images, ICCV(2015)

[7] Vinyals O., Toshev A., Bengio S., and et al.: Show and Tell: A Neural Image Caption Generator, ICML(2014)

[8] Zhong Z., Zheng L., Kang G., et al.: Random Erasing Data Augmentation *arXiv preprint arXiv 1708.04896*:, 2016.

[9] DeVries T., Taylor G. W.: Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv 1708.04552*, 2017.

[10] Kingma D. P., Rezende D. J. , Mohamed S., et al.: Semi-Supervised Learning with Deep Generative Models, NIPS(2014)

[11] Um T. T., Pfister F. M. J., Pichler D., et al.: Data augmentation of wearable sensor data for parkinson’ s disease monitoring using convolutional neural networks. *arXiv preprint arXiv:1706.00527*,2017.

[12] Guennec A.L., Malinowski S., and Tavenard R.: Data Augmentation for Time Series Classification using Convolutional Neural Networks. In ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data(2016)

[13] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. arXiv preprint arXiv:1611.01891, 2016.

[14] Kudo D., Horikawa M. , Furudate T., et al. Indoor Positioning Method Using Proximity Bluetooth Low-Energy Beacon. Proceedings of the 17th Asia Pacific Industrial Engineering and Management Systems Conference. 2016.