深層距離学習のための MOT 解析を使った教師画像自動生成

Automatically dataset generation using MOT analysis for deep metric learning

吉田 龍人[*1] 菊池 恵和[*1] 堀井 大輔[*1] 大久保 順一[*1] 高森 秀司[*1]
Ryuto Yoshida Yoshikazu Kikuchi Daisuke Horii Junichi Okubo Shuji Takamori

[*1] 八千代エンジニヤリング株式会社
Yachiyo Engineering Co.,Ltd.

Occlusion reduces the performance in Object Tracking. For keeping tracking when occlusion for a long period of time occurs, Metric Learning is usually used that measures similarity between images. On the other hand, it is difficult to train Metric Learning model because making training data takes high cost. In this paper, we propose the method for making training data automatically by MOT analysis and evaluate the model trained by this method.

1. はじめに

MOT（Multiple Object Tracking）タスクにおいてオクルージョンは物体追跡の精度を低下させる要因である．MOT とは人や車といった移動体の動画上での軌跡を物体別に取得するタスクで，各フレームに対する物体検出とフレーム間での物体追跡によって成り立つ．さらにオクルージョンとは手前の物体と奥の物体が重なり合う現象を指すが，MOT の解析においてオクルージョンが物体検出の漏れやそれに伴う追跡ミスを誘引する．MOT の物体追跡には，軌跡の座標情報を使って追跡する手法や検出領域の特徴量を用いる手法などが用いられるが，検出領域の持つ特徴量はオクルージョンの前後でも概ね同等であることから，特徴量を用いる手法は長期のオクルージョンにおける精度向上策としての期待がある．

近年では特徴量を用いて物体追跡する手法に，教師ありで構築した距離学習モデルを利用するのが一般的となっている．しかし，一般的な距離学習データセットである LFW で 13000 枚 / 1680 ID，Market-1501 で 32668 枚/1501 ID といった枚数の画像が構成されているように，教師あり距離学習用のデータセットは物体別に膨大なデータが必要となるため，モデルの作成は困難を極める．これに対して，物体ごとに ID を付与し，各 ID の位置を出力する MOT の解析結果を活用すると，ID ごとに検出された領域を切り出して保存するだけで，労せずにデータセットを作成することが可能となる．

本研究では MOT 解析を用いた自動生成データセットによって距離学習モデルの構築に取り組む．自動生成データセットには検出や追跡に失敗したことで生じるノイジーデータが多数存在するため，ノイジーデータの自動クレンジングも合わせて試行する．一連の手順によって得られたクレンジングをしていない「ノイジーだが多様性のあるデータセット」と，クレンジング後の「ノイジーではないが画一的なデータセット」のそれぞれで構築したモデルの性能を評価し，データセットの有用性を確認する．さらに物体追跡アルゴリズムとしての本モデルの適用性を確認するため，MOT 解析で得られたある人物の一連の画像に対して類似度計測を行い，定性的な評価を実施する．最後にこれらの成果を踏まえて本研究の課題を整理し，より良い距離学習モデルを構築する方法を考察する．

連絡先：吉田龍人，八千代エンジニヤリング株式会社，東京都台東区浅草橋 5-20-8 CS タワー3F 技術創発研究所，
TEL：03-5822-6843，Mail：ry-yoshida@yachiyo-eng.co.jp

2. 関連研究

距離学習の代表的な手法に，Contrastive Loss [Chopra 05]や Triplet Loss [Wang 14]のように Intra-class および Inter-class の画像の組によって学習を行う手法がある．これらはアンカーに対してポジティブ・ネガティブの画像の組み合わせを入力し，ポジティブの距離を最小化，ネガティブの距離を最大化させるように学習するものである．これらの手法には①入力データのマイニングアルゴリズムによっては学習が不安定になる，②しばしば Intra-class 内の最大距離が Inter-class の最小距離を上回るなどといった問題が発生する．

これらの手法に対して，A-Softmax(Angular Softmax)を利用した SphereFace [Liu 17]が登場して以来，Softmax 関数を活用した様々な距離学習手法が提案されている．A-Softmax とは角度距離を用いた Softmax 関数であり，適切な Angular margin を与えて学習を行うことでモデルの精度が向上すると報告されている．さらに ArcFace [Deng 18]では，(1)式で与えられる Loss を適用し，さらなる精度向上を果たした．

$$L = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(\cos(m_1\theta_{yi}+m_2)-m_3)}}{e^{s(\cos(m_1\theta_{yi}+m_2)-m_3)}+\sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}} \quad (1)$$

ここで，m_1，m_2，m_3 は SphereFace，ArcFace，CosFace [Wang, 18]のそれぞれの論文で提案された Angular margin を表している．

本研究での実験には ArcFace を用いる．実装は GitHub にて公開されている ArcFace の PyTorch を使った公式のものを利用した．margin は元の設定値に準拠してそれぞれ m_1=1.0，m_2=0.5，m_3=0 とし，特徴量抽出器には ResNet50 [He 15]を用いた．

学習の実施にあたって，距離学習においても他の深層学習タスクと同様にノイジーなデータセットが学習に悪影響を及ぼすことが問題視されている．ノイジーデータセットの影響を低減するための手法として ArcFace に改良を加えた Sub-center ArcFace [Deng 20]などの手法が既に提案されているが，本研究ではデータセットと学習結果の因果関係を直接的に評価するため，あえて元の ArcFace で学習を行った．

3. MOT を用いた自動生成教師による距離学習

3.1 実験データ概要

実験データは図-1 のような画角によって撮影された動画である. 動画は建物 2 階にあるバルコニーから高さ約 3m の三脚を使って地上を見下ろすように撮影している. 撮影日時は 2022 年の 12 月 9 日, 10 日の朝 7 時台から 18 時台であり, 日付によって撮影角度を変えたため映る範囲が異なる. 当該現場は非常に交通量が多く, 基本的に絶え間なく何らかの交通がある.

撮影に使用したカメラは Sony のビデオカメラ HDR-AS50 であり, 解像度は 1920×1080, フレームレートは 30fps である. 連続撮影時間は 35 分程度であるが, 全ての動画を連結する処理を施したものを解析に用いた.

3.2 MOT 解析概要

MOT 解析は既往論文[高森 21]の手法によって実施する. 撮影日は両日ともに晴れである. 撮影範囲は背の高いビルで囲まれていることから, 基本的には一帯が影になる時間が長いが, ごくまれに日照のある時間帯も存在し, 同一人物の持つ特徴量が日照の有無で瞬間的に変化することがある.

本システムにおける物体検出モデルは Intel の公開する person-detection-retail-0013 [Intel 23]である. モデルの詳細は明らかではないが mobilenetV2 [Sandler 18]ライクなモデルを特徴量抽出器に用いた SSD [Liu 15]であると公式ホームページにて示されている. よって精度よりも推論速度に強みがあるモデルであり, 推論結果にはノイズが多く含有していると判断される.

person-detection-retail-0013 の検出結果を確認すると, カメラ側の広範囲, 特に人の頭頂部を捉えたような画角にて, 歩行者が検出されていないことを確認した. 一方で対象データは長時間で歩行者数も膨大であったため, 数十万枚規模のデータを確保することができた. この結果を受けて, 物体検出モデルの改善は今後の課題としつつ, 以降の実験に取り組んだ.

3.3 実験

本実験では MOT 解析の解析結果を基に距離学習モデルの学習データセットを自動生成するとともに, 自動生成したデータセットで ArcFace の性能を評価する. MOT で作成されるデータセットには, 検出誤差等のノイズが多数含まれるため, ノイズをできる限り自動的に除去することが望まれるが, ノイズ除去で必要なデータを過度に削除するリスクもある. そこで本研究では元の MOT 解析で得たままのノイジーデータセットで構築したノイジーモデルと, ノイジーモデルによってクレンジングしたクレンジングモデルを構築し, これらの性能を評価する.

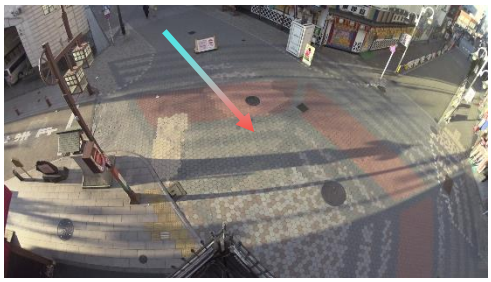
12 月 9 日の動画で MOT 解析を実施すると, 48428 ID が計上された. これらを全て切り出すとあまりにも膨大な画像となるため, 本研究では切り出す枚数を絞って以降の検証を行った. 切り出ししたデータセットを確認したところ, 各 ID を構成する画像群には以下に示す問題がしばしば確認された.

a) 検出範囲が瞬間的に過大になる.
b) 同一 ID 内に 2 人以上の異なる人物が存在する.
c) 同一人物に複数の ID が割り振られる(ID switch 現象).

これらの問題のうち, 問題 c)は出現してから消失するまでのフレーム数が短い ID で多発していることが明らかとなった. そこでノイジーデータセットにて出現から消失までのフレーム数が 250 を下回る ID を除外した.

次にノイジーデータセットによって学習したノイジーモデルを活用して, ノイジーデータセットの自動クレンジングを実施した.

(a) 学習地点 (12 月 9 日)



(b) テスト地点 (12 月 10 日)

図-1 対象現場

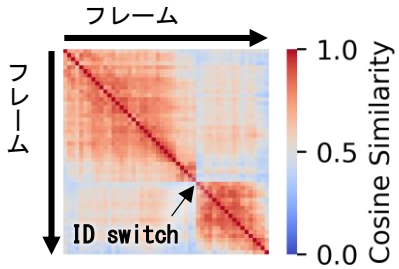


図-2 Switch 発生 ID での類似度の相関行列

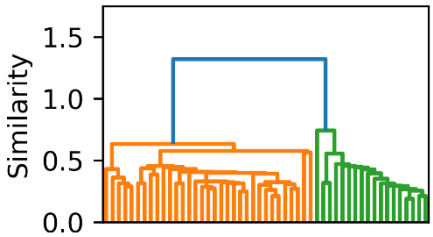


図-3 相関行列のクラスタリング結果

実施にあたって, 上述の問題点 a)および b)を引き起こす画像は, 同一 ID 内の画像同士でノイジーモデルによって推論した embedding feature のコサイン類似度を取った際に, 概ね識別できることを定性的に確認した. 実際に ID switch が発生した ID の画像群に対して, embedding feature の類似度を取得し, その結果を相関行列で表したものが図-2 である. 相関行列の要素は時系列順に並んでいるが, ID switch が発生したタイミングで相関性が劇的に変化している. この結果からノイズを極力減らしたデータセットを作成するために, 相関行列に対して階層型クラスタリング[Müllner 11]を適用した. 図-2 に示す相関行列にクラスタリングを実施した結果が図-3 である. これにより ID が変化した前後で, クラスタを二分することができた. クラスタリングで要素数が最大となったクラスタ以外に属する画像は, 最大クラスタに

属する画像と異なる見た目だと判断されるため，最大クラスタ以外を除去したクレンジングデータセットを作成した．検証に用いたノイジーデータセットは 51977 枚/873 ID であったが，クレンジング後に 40164 枚/851 ID となった．クレンジングの強弱はクラスタリングの閾値によって変えられるが，今回はデータセット内のノイズをほとんど除去するよう強めに実施した．

本実験では各モデルの学習曲線や推論精度を比較するとともに，同一 ID の類似度を測り，物体追跡での距離学習モデルの利用可能性を検討する．推論精度は CMC（Cumulative Matching Characteristics）曲線によって評価する．評価に用いるテストデータセットには 12 月 10 日の動画から MOT 解析によって生成した画像を精査したものを用いる．精査の際は，各 ID の本人が 8 割～10 割程度写った画像を Easy samples，5 割～8 割程度しか写っていない Hard samples に分け，それ以外の画像はノイズとしてテストデータから削除した．テストデータは 9383 枚（うち Easy samples は 8280 枚，Hard samples は 1103 枚）/224 ID である．CMC 曲線は Easy samples，Hard samples のそれぞれで取得し，Easy samples から 1 枚抜粋した画像を両データの Query として評価した．

4. 実験結果

4.1 学習曲線の比較

図-4 に本実験で行った 2 度の学習で得られた学習曲線を示す．いずれのモデルにおいても Loss がほぼ収束したことが確認された．収束時の Loss の大きさには乖離があり，ノイジーモデルは各 step で 6～8 あたりを推移し，クレンジングモデルは 2～4 あたりを推移した．最終的に到達した Loss の値には差異があるが，学習曲線の傾きには変化がなく，両者がほぼ平行するような形で収束する様子が観測された．

4.2 CMC 曲線による評価

図-5 にノイジーモデルとクレンジングモデルの CMC 曲線の取得結果を示す．ノイジーモデルは赤色，クレンジングモデルは青色で示しており，Easy samples に対する結果は実線，Hard samples に対する結果は破線で示している．

CMC 曲線より，ノイジーモデルの方が本テストデータにて高い精度となることが分かった．Easy samples における差は少ないが，Hard samples に対する識別精度には明瞭な差が見られた．いずれのモデルも Easy samples に対する識別精度は高く，8 割強の画像を Rank1 として正しくマッチングできた．

4.3 同一人物に対する類似度計測

MOT における物体追跡での本モデルの利用可能性を検証するため，ある ID の軌跡に対して embedding feature の類似度の相関行列を取り，その結果を評価した．対象人物は**図-1 (b)** 中に矢印で示す軌跡を歩んだ人物とした．当該人物は人目には明らかに同一人物であると判断されるものであり，モデルが全て同一と判定することを望んで選定した．出現時刻は**図-1 (b)** とほぼ同時刻で，赤色のレンガを通り抜けるあたりから日照によって画像の見た目が変化する．検証に用いたモデルは CMC 曲線で高精度と判定されたノイジーモデルである．

対象人物の出現から消失にかけて類似度の相関行列を取得した結果を**図-6** に示す．相関行列を見ると，日照が変化した際に相関行列の傾向が大幅に変化していることが分かる．日照なしからありへと推移した直後は日照なし・あり間でもやや高い類似性を示したが，時間が経過するにつれて日照なし・あり間の類似度は大きく低下した．

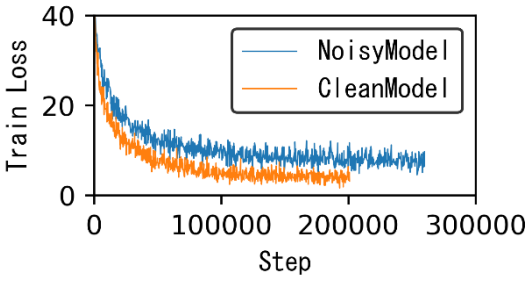


図-4 学習曲線

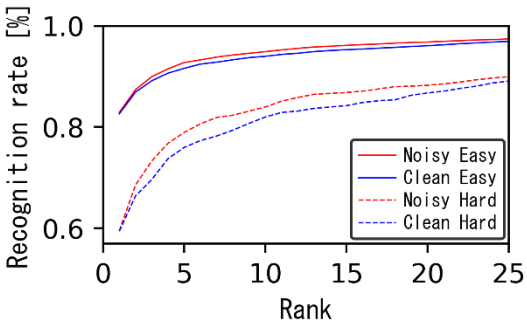


図-5 CMC 曲線

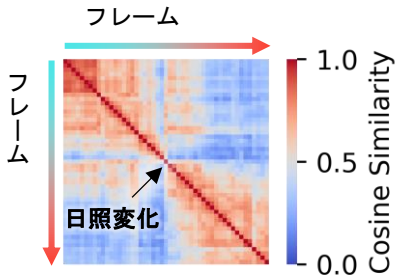


図-6 同一 ID の相関行列取得結果

また，登場直後の画像同士の類似度は 1 に近い極めて高い値を示したが，その後 0.7 程度まで類似度が下がる様子が確認された．これらに該当する画像を見ると，登場直後は人の頭から足先までが無駄なく検出されていたが，登場から時間がある程度経過すると，車止めの立て看板が影響して検出範囲がやや過大になる現象が生じていた．

4.4 考察

図-4 よりクレンジングモデル学習時の Loss はノイジーモデルよりも低い値を推移したが，学習曲線の傾きはほとんど同等であった．さらに**図-5** より推論精度はノイジーモデルの方が高く，特に Hard samples に対する精度が著しく高くなった．クレンジングモデルの Loss が低い値を示したのは学習データ自体が Easy samples の集合であり，モデルが ID の分類をしやすかったためだと判断した．さらに，クレンジングモデルは Hard samples を十分に学習していないため，モデルの汎化性が損なわれ，Hard samples に対してのみ大幅に精度が低下したのだと考えた．ノイジーモデルとクレンジングモデルで収束性に大きな違いが見られなかったのは各バッチの Loss の大きさによるものと予測した．ノイジーモデルは Loss の勾配が大きく，バッチごとのパラメータ更新量が大きいため，クレンジングモデルと同様の収束性を見せたと判断した．ノイジーデータセットは正しくラベルが付与され

た Easy samples および Hard samples と誤ってラベルが付与されたノイズの 3 つの要素に分けられるが，本研究で用いたノイジーデータセットに占めるノイズの割合が少なかったため，Hard samples のモデルを最適化する効果が，ノイズの学習を阻害する効果を上回ったのだと考えた．本研究ではクレンジングデータセットを作成する際に，ノイズのほぼ全てを除去するために高い強度のクレンジングを実施したことで，学習に必要な Hard samples を削除してしまったと推察する．クレンジング後の画像にも各 ID に平均約 50 枚の画像はあったが，ノイジーモデルで同一 ID であると識別された画像，つまり Easy samples らしい画像のみでクレンジングデータセットが構成されたため，モデルの精度向上に至らなかったのだと判断した．よって汎用的なモデル構築のためには Hard examples の存在は不可欠であると考えた．今回使用したデータセットにおいてはノイズを多少許容してでも Hard examples をデータセットに加える必要があった．

ただし本研究ではクレンジングの有無に関わらず Angular margin に同一のパラメータを用いた点に留意が必要である．Angular margin を大きくするとクラスの分離境界がより明確になるよう学習が進むため，過学習のような振る舞いを見せたクレンジングモデルを学習させる際は，margin を少なくするといった対策によって，モデルの性能が向上する可能性がある．

図-2 や**図-6** に示す同じ ID を構成する画像同士で類似度を計測した結果より，今回構築した距離学習モデルにおける物体追跡アルゴリズムとしての一定の実用可能性が示された．基本的に同一人物に対しては高い類似度を示し，ID switch の発生も見分けることが可能となった．その一方で日照の変化には弱く，対象人物が影から日向に出た瞬間に ID switch が発生したときと同じように類似度が著しく下がる状態が確認された．当該現場は基本的に一帯が影であったことから，教師データに影と日向を含んだデータが少なく，モデルが日照の変化に弱くなった可能性が高い．また今回用いたデータセットは，MOT 解析精度の都合により，基本的に同じ向きで写り続ける人物が多かったため，角度変化に対する頑健性は確かめられていない．本実験で検出漏れが多発した画面手前側に写る人物の方が，画面遠方に写る人よりもカメラと成す角度の変化量は大きく，画像の見かけも急激に変化すると想定される．よって角度変化に対する距離学習モデルの頑健性も追って検証する必要がある．

類似度を計測した結果からは，物体検出領域が広がると類似度が低下する現象も確認されたが，この現象は検出領域中の背景の情報が embedding feature に含まれたためだと考える．これを防止するためには，人以外の領域の特徴量を抽出しないモデルを構築する必要がある．具体的な対策としては，人の領域の特徴量を集中的に抽出するため空間方向に対する Attention を特徴量抽出モデルに導入することが想定される．

本研究で用いた動画の画角は，解析に用いた物体検出モデルに適していなかったため，十分な精度が得られなかった．特に物体検出が可能な範囲が限られたため，生成したデータセットにおける人の向きなどのバリエーションが限定的なものになった．他方で，本研究の MOT 解析から自動生成したデータセットで構築した距離学習モデルでも，ある程度の精度が得られたと判断する．したがって高い物体検出精度を持った MOT 解析システムを構築することで，既存の距離学習モデルの性能を最大限引き出せるようなデータセットを MOT 解析結果から自動生成することが可能であると考える．同じ人物をいかなる距離・角度からでも検出可能な物体検出モデルを用いると，データセット内の Hard samples を増やすことができ，より汎化性の高いモデル構築に繋がると判断する．

5. おわりに

5.1 まとめ

MOT 解析から距離学習データセットを自動的に作成し，そのデータセットによって学習した距離学習モデルの性能を評価した．検証では MOT 解析によって切り出しただけのデータセットと，それをクレンジングしたデータセットを用意し，それぞれで学習したモデルを比較した．その結果，データクレンジングをしないデータセットで構築したモデルの方が高精度となり，学習時の Hard samples の重要性が明らかになった．構築したモデルは汎化性等の課題も残るが，MOT 解析結果から自動生成したデータセットの有用性が確かめられた．

5.2 今後の課題

本研究の結果より，教師データの内容改善が構築されるモデルの精度に大きな影響を与えることが確認された．そこでより良いデータセットを構築するため，「MOT 解析システムにおける物体検出モデルを改善する」あるいは「今回用いた物体検出モデルが正しく動作する動画にて MOT 解析によるデータセット自動生成を行う」といった物体検出側の問題を改善するアプローチを取り，実用を想定した細かなモデルの精度検証に繋げたい．

謝辞：本研究に使用した動画は，著者が一般社団法人浅草六区エリアマネジメント協会の協力のもと撮影したものになります．AI 解析のための動画撮影および論文による解析結果の公表に快諾いただき本研究を進めることができました．ここに謝意を表します．

参考文献

[Chopra 05] S. Chopra, R. Hadsell, and Y. LeCun: Learning a similarity metric discriminatively, with application to face verification, CVPR, 2005.

[Deng 18] J. Deng, J. Guo, N. Xue et al.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition, arXiv: 1801.07698, 2018.

[Deng 20] J. Deng, J. Guo, and T. Liu: Sub-center ArcFace: Boosting Face Recognition by Large-scale Noisy Web Faces, ECCV, 2020.

[He 15] K. He, X. Zhang, and S. Ren et al.: Deep Residual Learning for Image Recognition, arXiv: 1512.03385, 2015.

[Intel 23] Intel: person-detection-retail-0013, https://docs.openvino.ai/latest/omz_models_model_person_detection_retail_0013.html, 22 Feb 2023.

[Liu 15] W. Liu, D. Anguelov, and D. Erhan et al.: SSD: Single Shot MultiBox Detector, arXiv: 1512.02325, 2015.

[Liu 17] W. Liu, Y. Wen, and Z. Yu et al.: SphereFace: Deep Hypersphere Embedding for Face Recognition, arXiv: 1704.08063, 2017.

[Müllner 11] D. Müllner: Modern hierarchical, agglomerative clustering algorithms, arXiv: 1109.2378, 2011.

[Sandler 18] M. Sandler, A. Howard, and M. Zhu et al.: MobileNetV2: Inverted Residuals and Linear Bottlenecks, arXiv: 1801.04381, 2018.

[Wang 14] J. Wang, Y. Song, and T. Leung et al.: Learning Fine-grained Image Similarity with Deep Ranking, arXiv: 1404.4661, 2014.

[Wang, 18] H. Wang, Y. Wang, and Z. Zhou et al.: CosFace: Large Margin Cosine Loss for Deep Face Recognition, arXiv: 1801.09414, 2018.

[高森 21] 髙森真紀子，大久保順一，藤井純一郎: 都市空間での人流解析における深層学習の応用，AI・データサイエンス論文集，2 巻 J2 号，pp.113-120，2021.