リアルタイムな動画内物体認識システム

長沢 瑛史[1]　　岡部 誠[1]

概要：我々は撮影中の動画の中から対象物体を探し出すためのインタラクティブなシステムを提案する。ユーザは対象物体が写っている画像をクエリとして与える。撮影中の動画にクエリで与えた物体が写ると、それをシステムはリアルタイムに認識してユーザに知らせることができる。提案手法は 2 つの ResNet50 から成り、一方はクエリ画像から特徴量を抽出し、もう一方は撮影中の動画の各フレームから特徴量を抽出する。両者を掛け合わせた特徴量は多層パーセプトロンによって、対象物体が写っているかどうかを表す値に変換される。対象物体は予め学習済みである必要はなく、かつ、提案手法はあらゆる対象物体の入力に対応できるようにしたい。そこで、DAVIS 2016 データセットを用いた画像合成により大量のデータセットを作成して学習を行った。車載カメラで撮影した動画を用いて道路標識の認識に関する実験を行い、その精度や実際の使い勝手について調査し、報告する。

キーワード：物体認識、深層学習、物探し

1. はじめに

人間の物体認識能力は高い一方で、人間は対象物体を簡単に見落としてしまうという特性も有している。スーパーマーケットの商品棚に欲しい商品が見つからず、店員に訊いたら見ていた棚にちゃんと目的の商品が陳列されていたとか、また、注意していたつもりでも自転車や自動車の運転中に道路標識を見逃してしまったという経験は誰しも持っている。このような見落としに伴う時間の浪費や運転中の危険性の誘発は豊かな生活を送る上で無視できない問題である。そこで我々は、人間に代わってコンピュータが対象物体を認識することで、対象物体を見落とさないようにサポートしてくれるようなシステムを構築したいと考えている。

コンピュータによる物体の認識技術は既に多数存在している。近年の物体認識技術は機械学習によって高精度に行えるようになったが、これらの手法の多くはコンピュータが予め学習してあるものを認識する[2,4]。MNIST[14]や CIFAR-10[15]といったデータセットは機械学習による物体認識のデータセットとして広く知られているものだが、これらを用いて学習したモデルはここに含まれた文字や物体の識別を高精度に行うことはできても、このデータセット含まれなかったものに関しては識別を行えず、そのような文字や物体を識別させたい場合はそれらの画像をデータセットに含めて学習をし直す必要がある。すなわちこのようなモデルでは、ユーザが新たに認識したい対象物があってもリアルタイムに対応することができない。

また失くし物を探すという点においては画像認識技術を用いない手段も考えられる。例えば、発信機等目印を予め物体に付けておき受信機でその発信源を特定する方法があり、代表的なものに Apple の AirTag[7]がある。この手法は視界に入っていない物体でも探し出せる点において画像認識を用いる手法よりも優れていると言えるが、一方で事前にその発信機が物体に付けてあることが要求されるため、陳列された商品のように予め目印を付けることのできないものを探し出すのは不可能である上、個人の所有物のような目印をつけることができるものであっても、探し出す可能性のある全ての物体に目印をつけることは現実的ではない。その点において画像から探し出す手法は、視界に入るという制約の上で全ての物体について考えることができる。

本研究では撮影中の動画にクエリで与えた物体が写った際にシステムが即座に認識し、ユーザに知らせる手法を提案する。提案手法は 2 つの ResNet50[3]から成る。一方はクエリ画像から特徴量を抽出し、もう一方は動画の各フレームから特徴量を抽出する。双方を掛け合わせた特徴量は多層パーセプトロンによって、対象物体が写っているかどうかを表す数値に変換される。人間がそうであるように、対象物体についてモデルは学習済みである必要はなく、また明るさや形などの異なる様々な対象物体の入力に対し同一物体であるか否かを判定できるようにしたい。そこで DAVIS 2016 データセット[9]から画像を合成することで大量のデータセットを生成し、これを用いて学習を行った。車載カメラで撮影した動画を用いて道路標識の識別に関する実験を行

1 静岡大学工学部数理システム工学科

い、その精度や実際の使い勝手について調査し、報告する。

2. 関連研究

動画や画像に写っている物体を認識するにあたり、様々なアプローチが存在している。

動画内でユーザの指定した物体を追跡する手法として、Video Object Segmentation (VOS)という技術が近年盛んに研究されている[1,10,11,12,13]。Wang らが 2019 年に発表した SiamMask[1]は、動画の最初のフレームで追跡対象の物体を矩形で囲むと、動画内の全フレームにおいて対象物体を追跡しマスクを生成する。Oh らが 2019 年に発表した STM[10]は、動画の最初のフレームで物体のマスクを与える必要があるが、現時点で最も高い精度でマスクを生成できる手法の 1 つである。しかし、これらの手法は動画中に連続的に表れる物体の追跡を対象としており、同じ物体が写っている別の画像をクエリとすることができない。

２つの画像にそれぞれ写っている同一物体の同じ箇所を結ぶ特徴点マッチングを用いることで同一物体を識別する手法も考えられている。ニューラルネットワークが登場する以前は画像内の各点の特徴を周囲の点などから定めた後[5,6]、アフィン変換等で空間的関係性を考慮して似た点同士をマッチングするといった手法が用いられていた。ニューラルネットワークが登場してからはより高度な特徴量の抽出を行えるようになり、特徴点の抽出及びマッチングの過程もニューラルネットワークが担うようになった。Wiles らが 2021 年に発表した CoAM[2]はニューラルネットワークを用いた最新の特徴点マッチング手法の 1 つで、天候や季節、時刻、写された角度など条件が全く違う建造物やモニュメントの画像に対してもマッチングを行うことができる頑健さを誇るが、学習も建物や彫刻の画像データを用いて行っているため、この学習データでは小さな標識や商品などに対してそのままマッチングを行うことは困難であり、学習データを多く集めるのが難しい場合に用いるのは不向きな手法であると言える。

Revaud らのオクスフォード及びパリの建造物を題材とし画像の検索技術を提供した[4]。クエリとして建物の画像を一枚入力すると、画像データの中でクエリに類似した画像（同一の建物が写された画像であると期待できるもの）が上位に来るように並べ替えることができるというものである。この手法も様々な角度・天気で撮影された画像データを柔軟に識別し高精度に並び替えることができる一方で、画像データは事前にニューラルネットワークに学習させておく必要があり、学習させていない画像を検索結果に反映させることはできなかった。

Yagi らの Go-Finder[8]は物探しに特化した支援システムである。手に持った物体をカメラで撮影してどこに収納したのかを記録しておき、ユーザはその記録から自分が持ったものを最後どこに置いたかを思い出せるようにするというものである。この手法では、探せるのは自分が手に取った物体についてのみであるため、一度も手に取ったことのない商品や看板・標識を探したり認識したりすることには用いることができなかった。

3. 提案手法

提案手法へのユーザの入力は、認識させたい対象物体が写っているクエリ画像と動画の 2 つである。クエリで与えた対象物体が動画に写ると、提案手法はリアルタイムに認識してその旨をユーザに知らせる。提案手法の概要を図 1 に示す。提案手法は 2 つの ResNet50 と、その後に追従する多層パーセプトロンから成る。168×168 ピクセルの RGB 画像２枚の画像 I_1, I_2 を入力とする。画像はそれぞれ ResNet50 によって 2048 次元の特徴ベクトル f_1, f_2 に変換され、これらを掛け合わせることで 2048 次元の特徴ベクトルが 1 つ得られる。この特徴ベクトルを 10 層の全結合層に通すことで最終的に２次元のベクトルが出力される。損失関数に多クラス交叉エントロピーを用いているため、画像 I_1 の物体が I_2 に写っていたら (0,1) 、違う物体が写っていたら (1,0) を出力するように学習を行う。

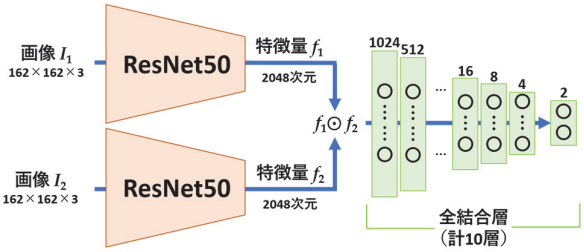

図 1. 提案モデルの概要図

4. DAVIS 2016 を用いたデータセットの作成

ある物体が写ったクエリ画像に対し、それと同じ物体が写った画像（以後正例と称する）及び違う物体が写った画像（以後負例と称する）を考える。正例が入力された時は 1 を、負例が入力された時は 0 を出力するように 3 章のモデル（図 1）を学習することで提案手法を実現する。精度の良いシステムを実現するには正例と負例を大量に生成して学習を行う必要がある。負例の大量生成は比較的容易であるものの、正例の大量生成には工夫が必要であり、ここが本論文の最も大きな技術的貢献となる。

正例はクエリ画像と同じ物体が写った画像であるが、その写り方はクエリ画像と同じであるとは限られない。クエリ画像の物体の向きや光の当たり方が変わっていたり、多少変形していたりしてもよく、精度の良い手法を開発するためにはそういったバリエーションが豊富な正例を大量生成することが必要である。このようなデータセットを作成するにあたって我々は DAVIS 2016 データセットに着目した。DAVIS 2016 データセットは 51 個の動画から成るデー

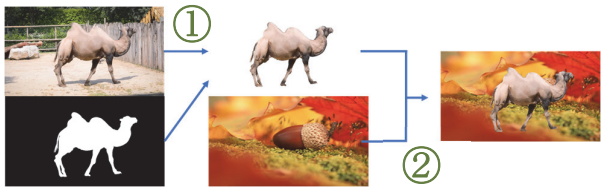

図 2. DAVIS 2016 から正例画像を生成した例。①DAVIS 2016 のラクダが写っている動画のフレームから対応するマスクを用いてラクダのみを切り取っている。②DAVIS 2016 と無関係な画像（ドングリ）の上に切り取ったラクダの画像を合成して正例画像を作成している。

タセットである。各動画には単一の物体が写っており、その物体のマスクも与えられている。写った物体は動画内において向きや光の当たり方、形など見え方が変化している。そこで、各動画の異なる 2 つのフレームをランダムに抽出し、片方をクエリ画像、もう片方を正例として用いることで、大量の正例集合を得ることができる。

但し、各動画のフレームをそのまま正例として用いると、対象物体の背景が常に同一となってしまう。我々は背景を無視し、対象物体にのみ着目した認識をさせたい。そこで、背景を様々に変更したバージョンを作成し、正例に更なるバリエーションを持たせる。DAVIS 2016 データセットで提供されている物体のマスクを用いて、ランダムに選ばれた画像と合成することで、背景の異なる正例を大量に生成する（図 2）。

データセットの生成についてより詳細に説明する。まず DAVIS 2016 データセットに含まれる 51 本の動画から 2 つの動画 A, B を選び、それぞれの動画から 2 枚ずつ計 4 枚の画像を生成する。まず、動画から切り取った物体を回転（0 度から 359 度まで 1 度ずつ、360 通り）し、また 50 % の確率で左右反転させ、一辺が 162 ピクセルの正方形となるようにリサイズした後、無地（白）の画像の上に合成する（図 3-A_1 と B_1）。続いて、動画から切り取った物体を回転・反転させた後、一辺が 81 ピクセル以上 162 ピクセル以下の正方形にリサイズしたものを、約 24000 枚の画像から無作為に選んだ背景画像のランダムな位置（但し物体が背景画像からはみ出さないよう）に合成する（図 3-A_2 と B_2）。これらを (A_1, A_2), (A_1, B_2), (B_1, A_2), (B_1, B_2) の 4 通りの組み合わせで (I_1, I_2) に入力した時の出力ラベルはそれぞれ 1, 0, 0, 1 となる。

5. 実験結果と考察

5.1 学習の様子

前節で述べた、2 つの動画から 4 組の画像を生成することをエポック毎に 2500 回行い、全部で 1 万組の画像データを生成し、それを学習に用いた。学習率は 1.0×10^{-5} で固定とした。また検証データとして一時停止を含む標識を撮影し

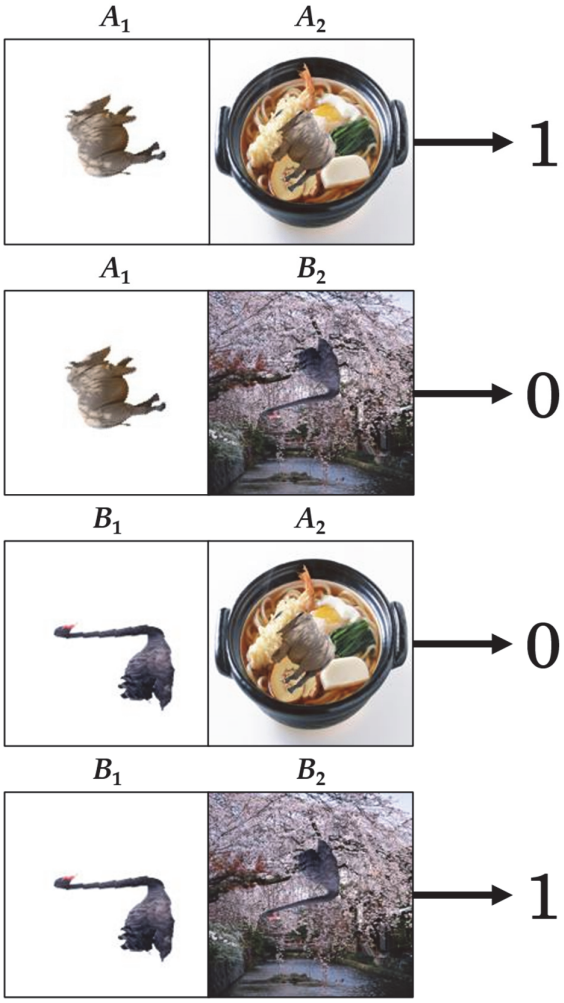

図 3. 入力の組み合わせとそのラベルの例。サイの動画から A_1, A_2 の画像を、コクチョウの動画から B_1, B_2 の画像を作成した。(A_1, A_2) 及び (B_1, B_2) はそれぞれクエリ画像に対し正例の組であるため 1 とラベリングされる一方、(A_1, B_2) 及び (B_1, A_2) ではクエリ画像に対する負例の組であるため、0 とラベリングされる。

表 1. 学習中の各エポックにおける学習データ及び検証データのロスと精度

エポック	ロス（学習）	精度（学習）	ロス（検証）	精度（検証）
1	0.6972	0.4985	0.6930	0.6298
200	0.6428	0.6572	0.7864	0.2137
400	0.5668	0.7178	0.3424	0.9313
600	0.4856	0.7744	0.4268	0.8588
800	0.3999	0.8214	0.3454	0.8817
1000	0.3013	0.8692	0.4568	0.8092

た 262 フレームの動画を用意した。この動画と一時停止の標識の画像（図 4）を入力し、モデルが一時停止の標識が写っている間に 1 を、それ以外では 0 を出力できるかどうか

図 4. 一時停止

を見た。

学習中の、学習データ及び検証データのロス及び分類の精度を示したものが表 1 である。

学習データは正例の画像の組と負例の画像の組が 5000 組ずつ生成される為、初期状態においての精度はおよそ 0.5 となる。また検証データについては、全 262 の内一時停止の標識が写っているフレームが 97 フレーム（全体の約 37 %）存在しているため、初期の精度は 0.37 もしくは 0.63 付近となる。

学習データを見ると、エポックの増加に伴いロスの減少及び精度の上昇が見られ、順調に学習が進行していることがわかる。1000 エポック目においては学習データについておよそ 87 % の精度での分類を実現している。

5.2 モデルの評価

最初に 1000 エポック学習させたモデルについて評価する。ここではその評価に、二値分類におけるモデルの性能を評価する指標の 1 つである Area Under the Curve (AUC) の値を用いる。AUC は最小値を 0、最大値を 1 とする指標で、その値が大きいほどモデルの性能は高いものと評価される。モデルがランダムに値を出力した場合の AUC の値はおよそ 0.5 となる。またこの AUC の算出に用いられる ROC 曲線も示す。

検証データとして用いた、標識を撮影した動画と一時停止の標識の画像を入力した結果における ROC 曲線は図 5 のようになった。AUC の値は 0.871 である。True Positive Rate (TPR) が 0.75 付近、False Positive Rate (FPR) が 0.00 の座標に点があるが、これは実際に標識が写っているフレームの約 75 %でモデルも標識が写っていると識別し、一方で標識が写っていないにも関わらず写っているとモデルが判定したフレームが全く存在しなかったことを示している。

続いて動画中に写っていない物体を入力した場合についても検証する。最初につづら折りありの標識（図 6）を入力した結果を示す。正解のラベルは常に 0 であるわけだが、このような場合については ROC 曲線を描くことができず、そのため AUC の値も算出できない。ここでモデルの出力をグラフに示す（図 7）。一部で反応してしまっているが、殆どのフレームにおいて出力はほぼ 0 (0.05) である。この値を上回る出力をしたフレームは 262 フレーム中 8 フレーム（約 3 %）のみであった。人間に注意を促すという目的を考慮すると、識別において大きな支障があるとは言えないだろう。

一方で、一時停止に似ている標識を入力すると期待に添わない結果となることもあった。例として進入禁止の標識（図 8）を入力した際の出力を図 9 に示す。また比較の為一時停止の標識を入力した結果も同時に示す。一時停止の標識と類似した出力をしており、このモデルでは色が似ているものについては誤検出してしまう可能性があることがわ

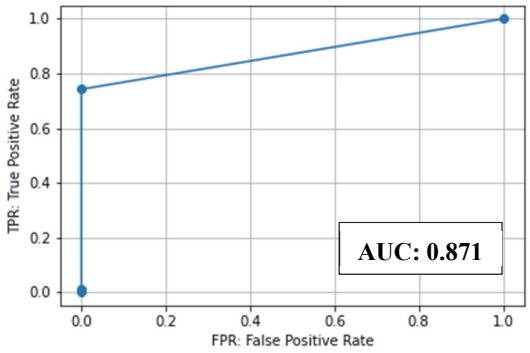

図 5. 一時停止の標識を入力画像とした時の ROC 曲線

図 6. つづら折りあり

図 8. 進入禁止

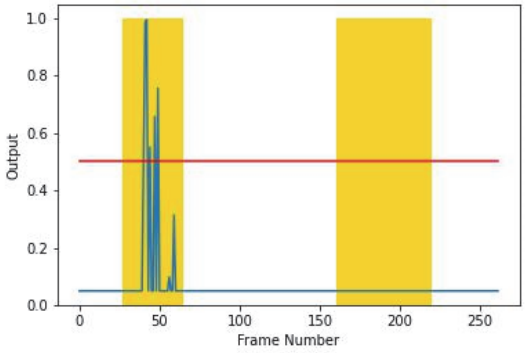

図 7. つづら折りありの標識を入力画像とした時のモデルの出力結果。横軸は動画のフレームを表し、縦軸はモデルの出力の値である。黄色のエリアは一時停止の標識が写っているフレームを示している。

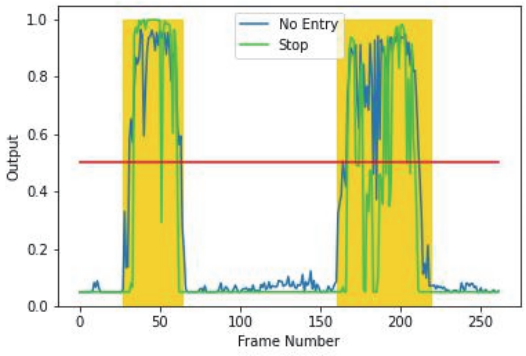

図 9. 進入禁止の標識を入力した時の出力（青）と一時停止の標識を入力した時の出力（緑）の比較

かる。一方で一時停止の入力では出力がほぼ 1.0 になっている箇所が複数認められるのに対し、進入禁止ではそのような値となることはなかった。モデルが色以外の手がかりから、一時停止の標識のほうがより写っている蓋然性が高いと考えていることが推察できる。

6. まとめ

本研究はニューラルネットワークに DAVIS 2016 データセットから作成した学習データを用いることで、学習データに含まれない物体の画像を識別できるようにし、リアルタイムに動画内から画像識別を行えるようになることを示した。今後は、精度向上は勿論のこと、単に画像を識別するだけでなく、位置の特定なども行っていきたい。

参考文献

[1] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, Philip H.S. Torr, “Fast Online Object Tracking and Segmentation: A Unifying Approach”, in Proc. CVPR 2019.

[2] Olivia Wiles, Sébastien Ehrhardt, Andrew Zisserman, “Co-Attention for Conditioned Image Matching”, in Proc. CVPR 2021.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition”, in Proc. CVPR 2015.

[4] Jerome Revaud, Jon Almazán, Rafael Sampaio de Rezende, César Roberto de Souza, “Learning with Average Precision: Training Image Retrieval with a Listwise Loss”, in Proc. CVPR 2019.

[5] D. Lowe, “Distinctive image features from scale-invariant keypoints”, International Journal of Computer Vision, Volume 60, Issue 2, pp. 91–110, 2004.

[6] Connelly Barnes, Eli Shechtman, Adam Finkelstein, Dan B Goldma,. “PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing”, in Proc. SIGGRAPH 2009.

[7] AirTag – Apple (日本), https://www.apple.com/jp/airtag/, 2022 年 2 月 7 日閲覧.

[8] Takuma Yagi, Takumi Nishiyasu, Kunimasa Kawasaki, Moe Matsuki, Yoichi Sato, “GO-Finder: A Registration-Free Wearable System for Assisting Users in Finding Lost Objects via Hand-Held Object Discovery”, in Proc. IUI 2021.

[9] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, “A benchmark dataset and evaluation methodology for video object segmentation,” in Proc. CVPR 2016.

[10] S.W. Oh, J.Y. Lee, N. Xu, S.J. Kim, “Video Object Segmentation using Space-Time Memory Networks”, in Proc. ICCV 2019.

[11] Z. Yang, Y. Wei, and Y. Yang,” Collaborative Video Object Segmentation by Foreground-Background Integration”, in Proc. ECCV 2020.

[12] J. Luiten, P. Voigtlaender, and B. Leibe,” PReMVOS: Proposal-generation, Refinement and Merging for Video Object Segmentation”, in Proc. CVPRW 2018.

[13] N.xu, L.Yang, Y.Fan, J.Yang, D.Yue, Y.Liang, B.Price, S.Cohen, and T.Huang, “Youtube-vos: Sequence-to-sequence video object segmentation”, in Proc. ECCV 2018.

[14] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.

[15] Alex Krizhevsky, “Learning Multiple Layers of Features from Tiny Images”, Master's thesis, University of Tront, 2009.