1Q-04

# フレーム間差分を用いた画像セグメンテーション改善法の検討

河村 聡太† 本多 泰理† 中村 周吾† 佐野 崇†

東洋大学 情報連携学部 情報連携学科†

## 1. はじめに

動画の各フレームからの物体セグメンテーションは、監視カメラからの不審者の検出や、自動運転中の他車両や歩行者などの検出といった応用のある重要な技術である。静止画とは異なり、セグメンテーション対象の画像に対して、時間的に前後する画像を用いることで、より効率的なセグメンテーションが可能であると考えられる。もっとも素朴な方法として、セグメンテーション対象フレームと、1 時刻離れた画像フレームとのピクセル強度差(差分画像)を用いる方法が考えられる。このような方法は[1]で提案されていたが、近年のディープラーニングベースの手法には取り入れられていなかった。

本研究では、ディープラーニングを用いたセグメンテーションモデルである U-net [2]をベースとして、対象画像と差分画像を入力とするモデルを作成し、差分画像がセグメンテーションの性能に与える効果を検証した。

## 2. 関連研究

画像のセグメンテーションを行うディープラーニングモデルとして U-net がある[2]。U-net は畳み込み演算とプーリングを繰り返して特徴マップの縮小を行った後、畳み込み演算とアップサンプリングを繰り返して画像の生成を行う。入力画像に対し、正しくセグメンテーションされた画像を教師信号とし、全体を誤差逆伝播法で最適化することで、end-to-end の学習が可能である。この手法は静止画のセグメンテーションモデルとして広く用いられている。

動画のセグメンテーションモデルとして、Two-Stream Fully-Convolutional Network(Two-Stream FCN)がある[3]。Two-Stream FCN は U-ne と類似した特徴マップの縮小部分と画像生成部分を持つが、2 つの画像を入力として受け取り、1 つの画像を出力するネットワーク構造となっている。2 つの入力画像のうち、1 つはセグメンテーション対象の画像、もう 1 つは動画から推定されたオブジェクトの optical flow である。このモデル構成は提案手法と類似するが、提案手法では、optical flow の代わりに差分画像を用いることが異なる。

## 3. 手法

本研究では、U-net をベースとした手法を用いて差分画像の有用性を検証する。比較対象となる U-net は、128x128 ピクセルの RGB 画像を受け取り、畳み込みとプーリングによって 2x2 ピクセルの 512 チャンネルの特徴量にまでダウンサンプリングしたのち、128x128 のグレースケール画像にアップサンプリングを行い出力とする。この出力が、教師として与えるセグメンテーションを再現するよう訓練を行う。

差分画像を用いる場合は次のように U-net を拡張する。まず入力画像として、セグメンテーションを行いたい画像(元画像)の他に、その画像の 1 時刻前のフレームの画像(前フレーム画像)と、それらのピクセル強度差をとった画像(差分画像)の、合わせて 3 つを用いる。これら 3 つの同時入力画像にあわせて、U-net のダウンサンプリング部分を 3 つに拡張する。ダウンサンプリングを行った結果、2x2 ピクセルの 512 チャンネルの特徴量が 3 つ作られる。これらの配列は平坦化され、結合されて 6144 要素の 1 次元配列に変換される。この 1 次元配列は、2 層の全結合層によって、大きさ 4096 の 1 次元配列に変換される。この 1 次元配列を再度 2x2 の 1024 チャンネルの特徴量に成形し、U-net と同様にアップサンプリングを行い、128x128 のセグメンテーション画像の生成を行う。

さらに、比較のために、元画像と差分画像の 2 つを入力とするモデルと、元画像と前フレーム画像の 2 つを入力とするモデルも作成した。これ

Image segmentation using differences between frames

† Sota Kawamura, Hirotada Honda, Shugo Nakamura, Takashi Sano

† Department of Information Networking for Innovation And Design (INIAD), Toyo University

らはどちらも2つのダウンサンプリング部を持ち、2 つの特徴量を平坦化、結合を行ってから全結合層で変換を行い、アップサンプリングを行ってセグメンテーション画像の生成を行った。

## 4. 実験と結果

表 1: U-net と 3 通りの提案手法の平均損失関数の値。3 入力は元画像、差分画像、前フレーム画像を入力とする

| U-net | 3 入力 | 元+差分 | 元+前フレーム |
|---|---|---|---|
| 0.0376 | 0.0371 | 0.0449 | 0.1031 |

データとして DAVIS 2017[4]を用いた。差分画像がセグメンテーションに効果的であるのは、カメラが固定されており、セグメンテーションすべき対象物が移動している場合であると予想できる。そのため、このうち、定点カメラによって撮影されていると思われる 10 の動画を実験に用いた。

実験に使用したモデルは、元画像のみを入力とする U-net、元画像、差分画像、前フレーム画像を入力とする 3 入力モデル、元画像と差分画像を入力とする 2 入力モデル、元画像と前フレーム画像を入力とする 2 入力モデル、の 4 種類である。

すべてのモデルにおいて、損失関数として平均自乗誤差を用いた。最適化アルゴリズムは Adam を用い、学習率は 0.005 とした。また、ミニバッチサイズは 32、訓練エポック数は 50 とした。

表 1 は、各モデルにおいて 10 回の異なる初期条件から学習を行った後、テストデータに対して平均自乗誤差を計算し、その平均値を記載したものである。U-net に対して、提案手法の 3 入力モデルはより小さな誤差の値を示した。また、2 通りの 2 入力モデルは、どちらも 3 入力モデルよりも大きな誤差の値を示した。

2 つの 2 入力モデルを比較すると、差分画像を含むモデルの方がより小さな損失関数を与えた。

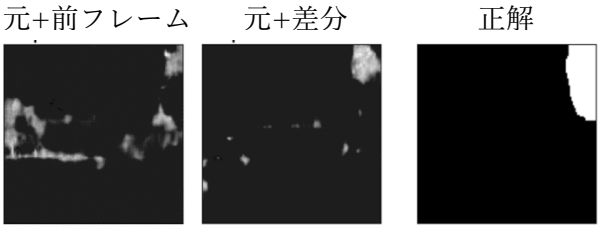


図 1: 2 入力モデルの予測結果と正解の例

図 1 に、この 2 つのモデルの予測結果の例を示した。正解画像と比較すると、差分画像を与えた場合のほうがより正しくセグメンテーションできている。

## 5. 考察とまとめ

本研究では動画中のある時刻の画像セグメンテーション法の検討を行った。1 時刻前の画像との差分を利用することでより精度の良いセグメンテーションが可能ではないかと考えた。U-net に類似した、ダウンサンプリングとアップサンプリングを行うニューラルネットワークを用いて検証を行った結果、元画像の他に差分画像と前フレーム画像を入力としたモデルが最も性能が良いことが分かった。さらに、元画像と差分画像を入力とするモデルと、元画像と前フレーム画像を入力とするモデルの比較を行ったところ、元画像と差分画像を入力としたモデルの方が性能がよく、セグメンテーションにおける差分画像の重要性が示唆された。

本研究では最終的な損失関数の大きさの比較を行ったが、差分画像は訓練速度を早める可能性があるため、その検証も今後は行いたい。また、類似手法として、差分画像の代わりに optical flow を用いた Two-Stream FCN がある[3]。Two-Stream FCN と提案手法との比較も行いたい。

## 謝辞

本研究は東洋大学重点研究推進プログラムの助成を受けた。

## 参考文献

[1] Jain,R., Martin, W.N., Aggarwal, J.K.," Segmentation through the detection of changes due to motion," Computer Graphics and Image Processing, Volume 11, Issue 1, pp. 13-34, 1979.

[2] Ronneberger,O., Fischer,P., Brox,T., "U-Net: Convolutional Networks for Biomedical Image Segmentation," MICCAI 2015, 2015.

[3] Jain, S.D., Xiong,B., Grauman, K.," FusionSeg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos," CVPR 2017, 2017.

[4] Pont-Tuset, J. et.al., "The 2017 DAVIS Challenge on Video Object Segmentation," arXiv:1704.00675, 2017.