IIEEJ ショートペーパー

SiamMaskを用いた動画修復

坪田颯生† 岡部誠†(正会員) 工藤隆朗††
由良俊樹††† 本間祐作†††

†静岡大学， ††株式会社 IMAGICA GROUP アドバンストリサーチグループ，
†††株式会社 IMAGICA Lab. 技術研究開発部

Video Completion Using SiamMask

Satsuki TSUBOTA†, Makoto OKABE†(*Member*), Takaaki KUDO††,
Toshiki YURA†††, Yusaku HOMMA†††

†Shizuoka University, ††IMAGICA GROUP Inc., †††IMAGICA Lab. Inc.

〈あらまし〉 本研究では, いずれも既存手法である SiamMask と動画修復技術を用いることで, 動画内の物体やロゴ, 文字, ノイズなどの消去をなるべく人手をかけずに行うシステムを開発した. ユーザは動画の最初のフレームにおいて, 目的物体をバウンディングボックスで囲むことで指定する. バウンディングボックスは SiamMask への入力となる. SiamMask は目的物体を追跡しつつ, 各フレームにおいて大雑把なマスクを生成する. 生成されたマスクを入力として動画修復技術を用いることで修復結果が生成される. 本研究が目指すのは, バウンディングボックスを描いたら即座に修復結果が生成されるようなシステムだが, 現状では SiamMask によって生成される大雑把なマスクが必ずしも完璧に目的物体を捉えない場合もあり, そのような場合はユーザによる手作業での修正 (フレームごとの画像編集作業) が必要となる. また本手法の応用として, 入力動画と修復結果の差分に基づいて目的物体の高精度なマスクを生成する手法も提案する.

キーワード：動画修復, マスク生成, SiamMask, 物体追跡

<**Summary**> In this project, we are developing a system to remove objects, logos, annotations, and noises from videos with as little human intervention as possible by using SiamMask and a video completion method, both of which are existing methods. The user specifies a target object by drawing a bounding box around it in the first frame of the video; this bounding box is taken as input by SiamMask. SiamMask then tracks the target object and produces its mask in each frame. The resulting masks are then taken as input by a video completion method, which produces the final video completion result. The goal of this project is that, after drawing the bounding box, the user immediately obtains the video completion result. However, the mask produced by our current method is not always perfect. When imperfections arise, the user still has to manually modify the mask using an image editing software. As an application of this method, we also propose a method to generate a highly accurate mask of the target object based on the difference between the input video and the restoration result.

Keywords: video completion, mask generation, SiamMask, object tracking

1. はじめに

映像制作の現場において, 動画内の物体やロゴ, 文字, ノイズなどを消去することは, 多くの需要があると同時に大変な労力を要する作業である. この作業を行うには, まず, 消去したい物体を指定するためのマスクを作成する必要がある. 次に, 動画修復技術を適用することで, 指定した物体を動画から消去する. 適切に作成されたマスクが与えられれば, 近年開発された動画修復技術[1)~3)]を用いることで, 物体が消去されたということがわからないような自然な動画を短時間で出力できる. しかし, 物体を指定するためのマスクはすべてのフレームにおいて手作業で描く必要があり, 作業に多くの時間を要する. そこで我々は, 動画修復のためのマスクを容易に短時間で作成し, 動画修復を効率的に行えるようなシステ

ムを開発している．さらに，元の動画と修復結果の差分に基づき，物体の高精度なマスクを生成する実験も行ったので報告する．

画像や動画から物体を正確にセグメンテーションするために数多くの手法が提案されてきた．その多くはユーザの入力無しに画像から物体を自動的に抽出する技術[4]や，それらを動画へ拡張した技術[5]などであるが，我々はあらかじめ学習済みの物体ではなく，ユーザによってその場で指定された物体のマスクを作成することを想定しているのでこれらの技術を適用することはできない．そこで我々は使い勝手が良く，計算効率も良い SiamMask[6]に注目した．ユーザが動画の最初のフレームで物体をバウンディングボックスで囲むと，SiamMask はそれ以降のフレームでその物体を追跡しつつ，各フレームにおいてマスクを作成する．

我々の目的は，ユーザが物体をバウンディングボックスで囲んだら，その物体が消去された動画を即座に出力することである．つまり，我々の目的は物体のマスクを正確に作成することではなく，動画修復をより正確により効率良く行うことである．動画修復に用いられるマスクは目的物体の形に完璧に一致する必要はないが，目的物体を完全に覆うような形でなければならない．そこで我々は，SiamMask で得られるマスクを膨張させて一回り大きくなるようにする．しかし，マスクを膨張させても常に完璧に目的物体を覆えるわけではないため，その場合はユーザーが手作業で追加のマスクを描いて修正する必要がある．

2. 提 案 手 法

例として，**図 1** に提案手法を用いて動画からローラーブレードに乗っている人物を消去する際の作業工程の概要を示す．まず，ユーザは動画の最初のフレームで人物を囲むようなバウンディングボックスを描く (図 1-Input の赤枠)．SiamMask はこのバウンディングボックスを入力として，各フレームにおけるマスクを生成する (図 1-SM の赤いピクセル)．このマスクの生成の処理には数秒を要する．SiamMask により生成されるマスクは目的物体の大部分を覆うが，覆い切れない部分も多く，すなわち，目的物体がマスクをはみ出してしまうことがある．例えば，6 フレーム目では手や足など，人物の領域の端に当たる部分を覆い切れていない (図 1-SM の黄矢印)．我々はそれらの部分をマスクが覆うようにするために，各フレームのマスクを一回り大きくするような膨張操作 (モルフォロジー変換) を加えた (図 1-Dilated)．しかしこれでも目的物体を覆い切れていない箇所がいくつかある．例えば，10 フレーム目や 33 フレーム目では手が覆われていない (図 1-Dilated の緑矢印)．このような場合，ユーザは GIMP[7]などの画像編集ツールを使って手作業でマスクを修正する必要がある (図 1-Modified)．最終的に修正されたマスクは動画修復技術[3]へ入力され，動画からローラーブレードに乗っている人物を消去することができる (図 1-Result)．この動画修復の処理には数秒を要する．

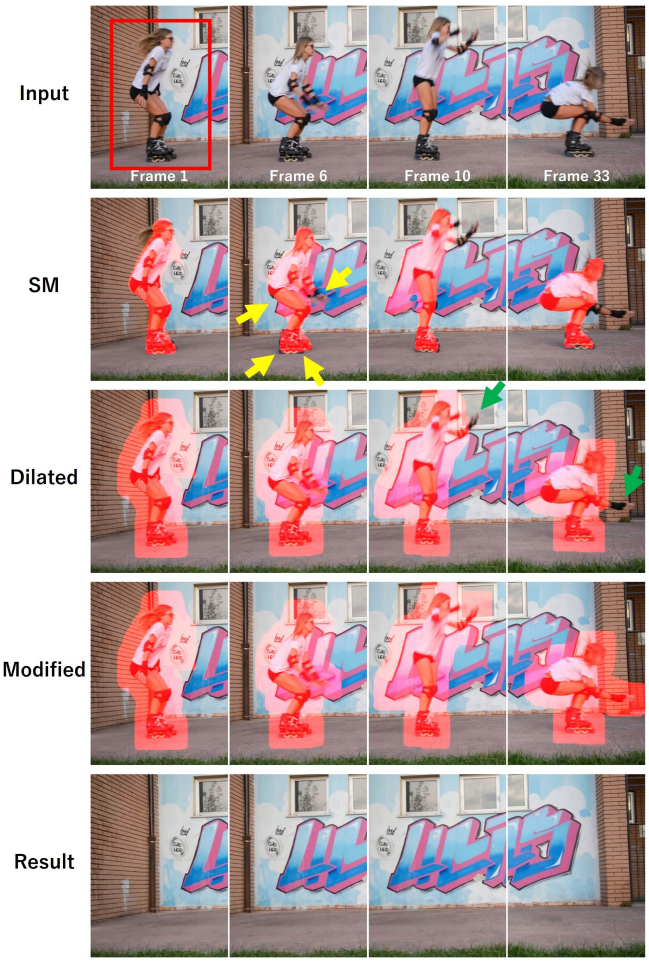


図 1 提案手法の概要
Fig. 1 The overview of the proposed method

3. 結果と考察

3 種類の動画を対象として提案手法を適用し，生成した動画について 12 名の学生を被験者として実験を行った．

3.1 実験結果

図 2 に提案手法による動画修復結果と背景差分によって生成されたマスクを示す．なお，マスク生成については 3.2 節で後述する．

“soccerball” の動画はフレーム数が 48 である．その 12 フレーム目を図 2 の 1 行目に示す．この動画については，本研究の目的を達成することができた．すなわち，ユーザが動画の最初のフレームにおいてバウンディングボックスでボールを囲めば，即座にボールが消去された動画を得ることができた．

“bmx-trees” の動画はフレーム数が 80 である．その 31 フレーム目を図 2 の 2 行目に示す．動画の最初のフレームでバウンディングボックスを描くと，SiamMask と膨張操作によって 72 枚のフレームで適切なマスクを得たが，残り 8 枚のフレームでは正確なマスクが得られなかった．それらのフレームのマスクを GIMP を用いて手作業で修正することで満足のいく修復結果を得ることができた．

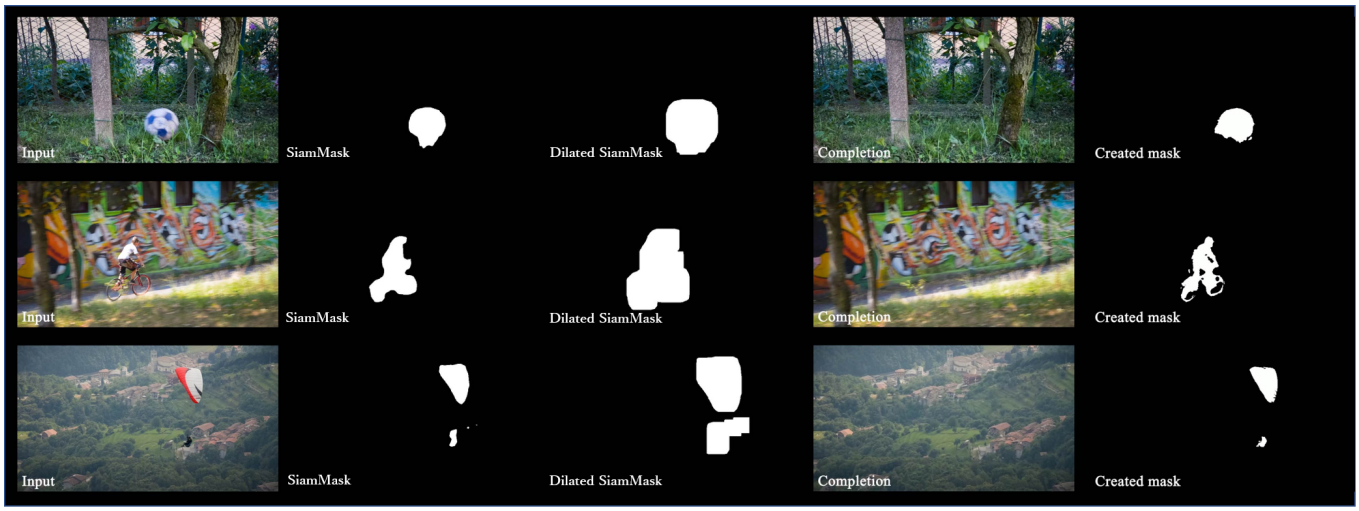


図 2　提案手法による動画修復結果と背景差分によって生成されたマスク
Fig. 2　Video completion results by the proposed method and the mask generated by the background subtraction

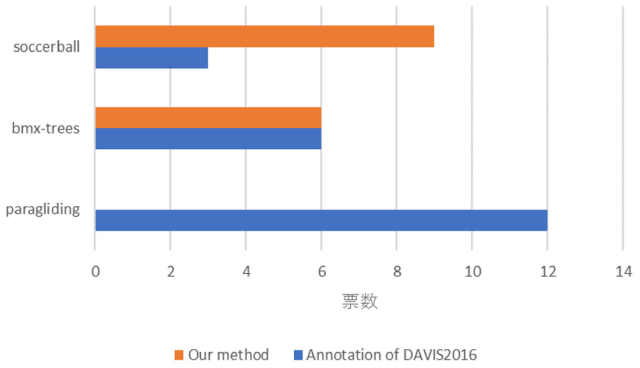


図 3　定性的な評価の結果
Fig. 3　Results of the qualitative evaluation

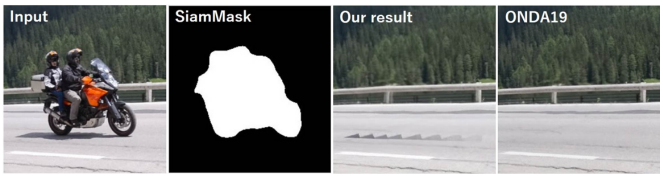


図 4　提案手法では影の消去ができない
Fig. 4　Shadow elimination is not possible with the proposed method

“paragliding” の動画はフレーム数が 70 である．その 31 フレーム目を図 2 の 3 行目に示す．この動画では SiamMask と膨張操作によって 61 枚のフレームで適切なマスクを得て，残り 9 枚のフレームでは GIMP での手作業の修正により適切なマスクを得た．

生成した動画に対し，定性的な評価を行った．被験者は 12 名の学生であり，比較対象として DAVIS2016[8]のアノテーションであるマスクに膨張操作をして動画修復技術により動画修復を行ったものを用意した．各動画について入力動画と動画修復結果を被験者に見せ，提案手法と比較対象のうち自然に物体を消去できていると思った方に投票してもらい，投票した理由も答えてもらった．

結果は**図 3** のようになった．ここでオレンジの棒グラフは提案手法を選択した人数，青の棒グラフは既存手法を選択した人数を示す．“soccerball” と “bmx-trees” に関しては投票結果にばらつきがあり，被験者の大半の意見として本手法と比較対象の結果にほとんど違いは見られないとのことであった．比較対象が全フレームにおいて手作業で作られた正確なマスクであるのに対し，本手法では “soccerball” については手作業無し，“bmx-trees” については 8 枚のフレームでの手作業により同程度のクオリティの修復動画を生成したことになる．これは動画修復の作業において，提案手法を用いることで従来よりも大幅な負担軽減になっていると言える．“paragliding” では被験者全員が DAVIS2016 のアノテーションによる修復動画に投票し，その理由として本手法による修復動画には物体の一部が少し残っているということを挙げた．“paragliding” の動画にはパラシュート部分と人間の間にロープがある．DAVIS2016 のアノテーションではこのロープ部分もマスクされているが，一方，SiamMask はこのロープのような細い物体のマスクを作る能力がない．そのためロープが消去されず動画に残ってしまった．

今後の目標は，SiamMask を改良することで，さらに適切なマスクを生成できるようなシステムを開発し，ユーザの手作業での修正といった負担を減らしていくことである．今回提案した手法では，図 1 の手や，図 2 の 3 行目のパラグライダーのような小さい物体や細い物体のマスクを適切に生成することは難しいことがわかった．また，図 2 の 2 行目の木の後ろから急に現れる自転車のように，動画内で突然消えたり出現したりする物体を捉えてマスクを生成することも困難であることがわかった．これは，SiamMask がマスクの推定をする際に，直前のフレームで生成されたマスクの情報を使うためだと考えられる．これらの問題を今後解決していきたいと思う．

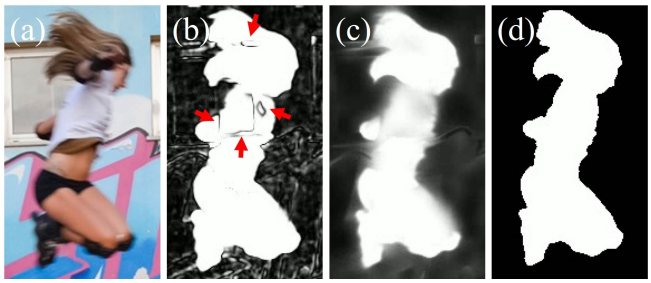


図 5　背景差分によるマスクの生成プロセス
Fig. 5　The process of generating a mask by background subtraction

さらに別の問題として, 提案手法では物体の下にできる影を取り除くことも難しい. 例として**図 4** を示す. 目的の物体はオートバイで, その下にくっきりとした影があるが (図 4-Input), SiamMask では正確に捉えることができない (図 4-SiamMask). これは SiamMask が, 画像分類のためのデータセットを事前学習したネットワークをバックボーンに使用しているからではないかと推察する. つまり, 画像分類のためには影を考慮する必要がない. 影の部分をほとんど捉えていないため, 膨張操作を行っても影を消すことはできなかった (図 4-Our result). 影を捉え, 既存手法[3]で得られたような結果 (図 4-ONDA19) を達成することも今後の目標の 1 つである.

3.2　背景差分によるマスク生成への応用

動画修復に使うマスクは上述のように目的物体の形に完璧に一致するマスクである必要はないが, そのようなマスクを作成することは, 映像制作の現場における別の重要な問題の 1 つである. そこで我々は, 元の動画と修復結果の差分に基づき, SiamMask では得られなかったような高精度なマスクを生成できるのではないかと考えた. これが上手くいけば, ユーザは動画修復に必要な程度のラフなマスクを作成するだけで高精度なマスクを手に入れることができる.

背景差分により SiamMask より高精度なマスクを生成するプロセスを**図 5** に示す. まず, 図 1-Input と図 1-Result の差分によって図 5(b) のような画像が得られる. 人物の領域にマスクができているが, ところどころに黒い穴が開いていて (図 5(b) の赤矢印), このままでは高精度なマスクと言えない. 図 5(b) を観察すると, ここに見られる穴は主に背景, つまり, 図 1-Result に見られる窓枠や壁の模様が原因であることがわかる. つまり, 背景差分におけるノイズは背景画像のエッジに起因する傾向が強い. そこで, cross bilateral filter[9]を適用し, 入力動画 (図 5(a)) に見られるエッジは保存し, それ以外のエッジを消去するように平滑化することで, 図 5(c) を得る. 最後に閾値によって 2 値化した画像が図 5(d) である. 以上のプロセスにより, **図 6** のようなマスクを得ることができる.

図 2 の 5 列目は, この手法により生成したマスクである. これは, 図 2 の 2 列目の SiamMask によるマスクよりも高精度なマスクであることがわかる. 動画修復に必要な程度のラフなマスクは高精度なマスクに比べて圧倒的に簡単に作成できるので, 本手法によって高精度なマスクを必要とするユーザの負担をかなり軽減できているのではないかと期待している. 将来はユーザテストや現場で働くデジタルアーティストへのインタビューを実施することによって実際の効果を確認したい.

図 6　背景差分によるマスクの生成結果
Fig. 6　Masks generated by background subtraction

4.　おわりに

本稿ではいずれも既存手法である SiamMask と動画修復技術を用いることで, 素早く簡単に動画修復を行えるようなシステムを提案した. 実験の結果, 提案手法を用いるとバウンディングボックスを描くだけで動画から物体を完全に消去できる場合があることがわかった. また手作業での修正が必要な場合も, 編集すべきフレーム数を大幅に削減できることがわかった. さらにユーザ調査の結果, 提案手法の動画修復結果は, 正確に作られた DAVIS2016[8]のマスクによる結果と比較しても, 遜色ないクオリティであることが確認できた.

しかし, ロープのように細い物体や動画内に突然現れる物体, そして物体の影などに対しては SiamMask では適切にマスクを作ることができず高精度な動画修復を行うことができなかった. 今後はこれらの問題を解決していきたいと思う.

参考文献

1) A. Bokov, D. Vatolin:“100+ Times Faster Video Completion by Optical-Flow-Guided Variational Refinement”, Proc. of IEEE International Conference on Image Processing 2018, pp. 2122-2126 (2018).
2) R. Murase, Y. Zhang, T. Okatani:“Video-Rate Video Inpainting”, Proc. of IEEE Workshop on Applications of Computer Vision 2019, pp. 1553-1561 (2019).
3) M. Okabe, K. Noda, Y. Dobashi, K. Anjyo:“Interactive Video Completion”, IEEE Computer Graphics and Applications (CG&A), Vol. 40, pp. 127-139 (2019).
4) R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, K.He, Detectron, `https://github.com/facebookresearch/detectron` (2018).
5) P. Voigtlaender, M. Krause, A. Ošep, J. Luiten, B. B. G. Sekar, A. Geiger, B. Leibe:“MOTS: Multi-Object Tracking and Segmentation”, Computer Vision and Pattern Recognition (2019).
6) Q. Wang, L. Zhang, L. Bertinetto, W. Hu, P. H. Torr:“Fast

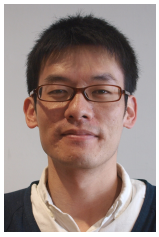
Online Object Tracking and Segmentation: A Unifying Approach", Proc. of Computer Vision and Pattern Recognition 2019 (2019).
7) The GIMP Development Team, Gimp, https://www.gimp.org
8) F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, A. Sorkine-Hornung:" A benchmark Dataset and Evaluation Methodology for Video Object Segmentation", Computer Vision and Pattern Recognition 2017 (2017).
9) S. Paris, F. Durand:"A Fast Approximation of the Bilateral Filter Using a Signal Processing Approach", Proc. of European Conference on Computer Vision 2006, pp. 568-580 (2006).

(2020 年 12 月 17 日 受付)
(2021 年 4 月 17 日 再受付)

坪 田 颯 生
2020 年 静岡大学工学部数理システム工学科卒業. 現在, 静岡大学大学院総合科学技術研究科修士課程に在学. 動画中の人や物体のマスク生成に関する研究に取り組んでいる.

岡 部 誠 （正会員）
2008 年 東京大学大学院情報理工学系研究科博士課程修了. 博士 (情報理工学). 2008 年マックスプランク研究所ポストドクター. 2010 年電気通信大学助教. 2016 年静岡大学助教. 2017 年同准教授. 動画データの分析と映像製作ツールに関する研究に取り組んでいる.

工 藤 隆 朗
2005 年 東京農工大学工学府電気電子工学専攻修了 (修士). 同年, 株式会社 IMAGICA に入社. 主に映像圧縮技術の改善, パッケージメディア生産ラインの構築に従事. 2021 年より株式会社フォトロンにて先端映像技術の研究に取り組んでいる.



由 良 俊 樹
2006 年 上智大学理工学研究科修士課程修了. 同年, 株式会社 IMAGICA に入社. 主に映像作品の色管理や画像処理技術の研究に従事. 2021 年より株式会社フォトロンにて先端映像技術の研究に取り組んでいる.

本 間 祐 作
2018 年 首都大学東京大学院システムデザイン研究科経営システムデザイン学域修士課程修了. 同年, 株式会社 IMAGICA に入社. 2021 年より株式会社フォトロンにてクラウド関連の事業や先端映像技術の研究に取り組んでいる.