氏　　名	Trung-Nghia LE
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 2038 号
学位授与の日付	平成 30 年9月28日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Video Salient Object Segmentation Utilizing Deep Features in Spatial and Temporal Domains
論 文 審 査 委 員	主　査　教授　杉本　晃宏 准教授　CHEUNG Gene 准教授　児玉　和也 教授　佐藤　真一 国立情報学研究所 教授　佐藤　洋一 東京大学　生産技術研究所

(Form 3)

Summary of Doctoral Thesis

Trung-Nghia LE

Video Salient Object Segmentation Utilizing Deep Features in Spatial and Temporal Domains

The concept of visual saliency has been extensively studied via multiple disciplines including cognitive psychology, neuroscience, and computer vision. Especially, in computer vision, many research efforts have been devoted toward to investigate salient object segmentation methods, which aim to localize and segment salient regions/objects from images and videos. However, existing methods for videos still do not effectively exploit temporal information, which is crucial to deal with videos.

Here in this dissertation, we investigate video salient object segmentation methods by leveraging long short-term memory. Essentially, we study two complementary tasks, including salient object segmentation and semantic salient instance segmentation. For coarse segmentation task (i.e., salient object segmentation), we divide a video into small blocks and then learn spatial and temporal information from the whole video block. On the other hand, for fine segmentation task (i.e., semantic salient instance segmentation), we utilize both propagation and re-identification schemes to exploit different timescales in the temporal domain. Therefore, our work effectively utilizes spatial and temporal domains for saliency computation.

We investigate region-based and pixel-based approaches for salient object segmentation. In the first approach, we propose a method for segmenting salient objects in videos where temporal information in addition to spatial information is fully taken into account. Following recent reports on the advantage of deep features over conventional hand-crafted features, we propose the SpatioTemporal Deep (STD) feature that utilizes local and global contexts over frames. We also propose the SpatioTemporal Conditional Random Field (STCRF) to compute saliency from STD features. STCRF is our extension of CRF toward the temporal domain and formulates the relationship between neighboring regions both in a frame and over frames. STCRF leads to temporally consistent saliency maps over frames, contributing to the accurate segmentation of the boundaries of salient objects and the reduction of noise in segmentation.

In the second approach, we present a novel end-to-end 3D fully convolutional network for salient object segmentation in videos. The proposed network uses 3D filters in the spatiotemporal domain to directly learn both spatial and temporal information to obtain

3D deep features, and transfers the 3D deep features to pixel-level saliency prediction, outputting saliency voxels. In our network, we combine the refinement at each layer and deep supervision to efficiently and accurately detect salient object boundaries. The refinement module recurrently enhances to learn contextual information into the feature map. Applying deeply-supervised learning to hidden layers, on the other hand, improves details of the intermediate saliency voxel, and thus the saliency voxel is refined progressively to become finer and finer.

We verify the effectiveness of our approaches on publicly available benchmark datasets, including 10-Clips, SegTrack2, DAVIS-2016. Intensive experiments confirm that our proposed methods outperform state-of-the-art methods.

Furthermore, we investigate in-depth tasks of salient object segmentation by raise a new interesting yet challenging problem of video semantic salient instance segmentation. We here do not only segment salient regions in a video but also decompose them into instances and exploit the semantic information of instances. To address the problem, we propose a new video dataset, namely Semantic Salient Instance Video (SESIV), with corresponding evaluation measures. Our SESIV dataset consists of 84 high-quality video sequences with various densely annotated, pixel-accurate and per-frame ground-truth labels for different segmentation tasks. We believe that our novel dataset will promote new advancements on video semantic salient instance segmentation. We also provide a baseline for solving the problem, called Fork-Join Strategy (FJS). FJS is a universal two stream framework, leveraging advantages of different segmentation tasks (i.e., semantic instance segmentation for the main stream and salient object segmentation for the context stream). In FJS, we introduce a novel sequential fusion, which can deal with overlapping regions of multi-instance segmentation, to frame-wisely combine the results of the two streams. We also propose a novel recurrent instance propagation to refine instances in the temporal domain for both mask shape and semantic meaning. The identity propagation and re-identification of instances are also introduced in both short-term and long-term memories to maintain both the identity and the semantic meaning over the entire video. Experimental results demonstrated that our proposed method is capable of achieving state-of-the-art performance on the newly constructed SESIV dataset.

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏　　名　Trung-Nghia LE

Title
論文題目　Video Salient Object Segmentation Utilizing Deep Features in Spatial and Temporal Domains

博士論文は、「Video Salient Object Segmentation Utilizing Deep Features in Spatial and Temporal Domains（時空間領域を考慮した深層特徴による顕著物体領域切出し）」と題し、英文で書かれている。博士論文における研究は、映像中の顕著物体をできるだけ正確に切り出すことを目的として、領域ベース、画素ベースで深層学習によって得られた特徴を時間的、空間的に利用して顕著物体領域を切り出す手法を提案し、従来手法に対する優位性を示している。また、切り出した領域を物体ごとに分離し、各物体に意味ラベルを付与するという新たな問題を提起し、その解決手法を提案している。これまで開発されてきた手法では、時間情報の抽出は原理的に２フレームからであったのに対し、本論文で提案された手法では、時間情報の抽出に利用できるフレーム数は計算機の能力に応じてスケールするという特長を持っている。

博士論文は６章で構成されている。第１章では、映像中の顕著物体領域切出しの重要性を論じ、さらに切り出した領域を物体ごとに分離し、各物体に意味ラベルを付与することの重要性を論じることで、博士研究で取り組む問題の意義を議論している。そして、領域ベース、画素ベースで深層特徴を時空間的に利用したアプローチの意義を論じて博士研究の位置づけを明確にしている。第２章では、顕著物体領域切出しにおける従来研究の動向と問題点を整理し、さらに、ベンチマークとされている公開データセットを網羅し、整理している。引き続く３つの章が本論文の主要部分となっている。第３章では、映像中の各フレームから抽出された時間的に追跡可能な小領域から得られる局所的深層特徴、数フレームの時系列画像全体から得られる大域的深層特徴を組み合わせた時空間深層特徴を新たに導出し、それを用いて時空間領域を考慮したグラフィカルモデルのエネルギー関数を定義し、そのエネルギー関数最小化によって顕著物体領域を切り出すという手法を提案している。そして、提案手法による顕著物体領域切出しの精度を最先端の関連手法と比較し、その優位性を示すとともに、提案手法を構成する各要素が最終結果にどのように貢献しているかを実験的に検証している。第４章では、映像中の数フレームの時系列画像全体を入力として、画素ベースで深層特徴を抽出し、それ用いて顕著物体領域を切り出す end-to-end の手法を世界に先駆けて提案している。この手法は、深層特徴と顕著物体領域との関係を効果的に学習するための再帰型のネットワーク機構の導入と各層ごとに教師データを利用する学習法を特長としている。ここでも、提案手法が最先端の関連手法に対して優位であることや導入した学習法の効果などを実験的に検証している。また、第５章では、切り出した顕著物体領域を物体ごとに分離し、各物体に意味ラベルを付与する問題を新たに提起し、そのための正解値付きベンチマークデータセットを新たに構築するとともに、

映像中のフレームごとに独立に付与された意味ラベルを時空間方向に伝搬、修正し、映像全体にわたって整合した意味ラベルを付与する手法を提案している。これによって、顕著物体領域を単に切り出すだけではなく、個別の物体領域に分離し、その意味を与えるという一歩踏み込んだ映像理解が可能となった。第６章では、まとめと今後の課題、展望を示している。

出願者による約４５分の発表もこの順で説明が行われ、その後、３０分弱の質疑応答があった。審査委員からは、顕著物体領域切出し結果のちらつきを評価する指標の導入や領域ベース、画素ベースのそれぞれのアプローチの結果の長所短所などについて質問とコメントが寄せられ、それらに対し出願者は適切に回答した。

質疑応答後に審査委員会を開催し、審査委員で議論を行った。審査委員会では、出願者の博士研究が顕著物体領域切出しという問題に対して独創的であることが評価されるとともに、研究成果として、トップレベルの国際学術雑誌 IEEE Trans. on Image Processing での査読付き論文１件の採択、および、権威ある国際会議やワークショップにおいて査読付き論文６編を発表していることが確認された。以上の理由により、審査委員会全員一致で、博士論文として十分な水準にある研究であると認め、本論文が博士の学位請求論文として合格であり、学位の授与に値すると結論づけた。

（備考）

１．用紙の大きさは，日本工業規格（JIS）Ａ４縦型とする。

２．１行あたり 40 文字（英文の場合は 80 文字），１ページ当たり 40 行で作成する。

３．上マージン，下マージンは２cm，左マージンは 2.5 cmとする。

４．タイトルと本文の間は，１行空ける。

５．ページ番号は入れない。

６．出願者（申請者）が論文審査に合格し，博士号が授与された場合は，本紙を総合研究大学院大学リポジトリにおいて，インターネット公開する。

Note:

1. The sheets must be Japanese Industrial Standard (JIS) A4 vertical.
2. Each line shall have approximately 40 characters in Japanese or 80 characters in English, and each page shall have 40 lines.
3. The top, bottom, and right margins must be 2 cm and the left one must be 2.5 cm.
4. Single spacing is required between the title and the text.
5. There must be no page numbers.
6. If the applicant is conferred a doctoral degree, this paper will be published on the SOKENDAI Repository.