イントラフレームのみを利用した高速な動画像要約手法の検討

小島颯，田中雄一
大阪大学大学院 工学研究科

2026/08/28 信号処理研究会@伊香保

大阪大学 The University of Osaka

Signal & Information Processing

背景

- 日々多くの動画像が撮影・流通
  YouTube上では累計200億本,
  1日あたり2,000万本がアップロード（2025/03時点）
  固定回線の下りトラフィックの39%を動画アプリが占める
- 限られた通信資源の中で効率的に内容を把握する手段が必要

動画像要約タスク

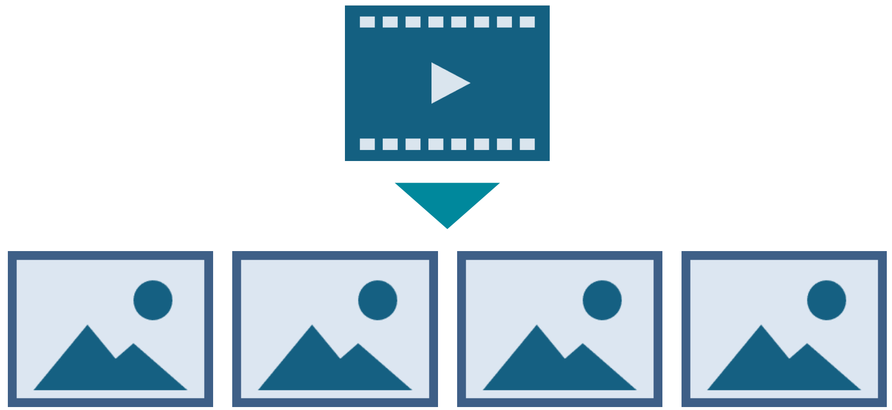
- 長時間の動画像から代表的なフレーム・動画像を抽出するタスク
- 1本あたりの処理コストが実用上の制約となる
  処理対象の動画像数が膨大のため

	Application	% DS Vol
1	Video	39%
2	Social Media	18%
3	Television	11%
4	File Sharing	9%
5	Device Gaming	7%
6	General Web Apps	6%
7	Communication	2%
8	VPN	2%
9	Audio	0.7%
10	Conferencing	0.3%
11	Cloud Gaming	0.07%
12	IoT	0.03%
13	Peer To Peer	0.03%
14	Other Apps	5%

固定回線の利用割合
www.applogicnetworks.com

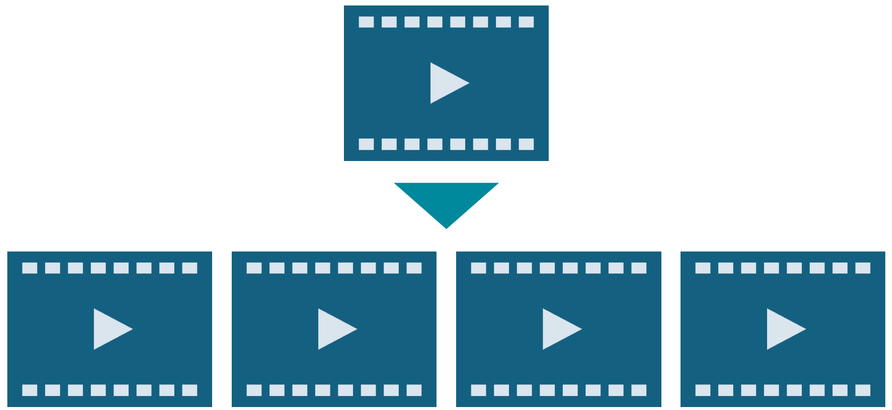
動画像要約タスク

重要フレーム抽出



- 動画像から重要なフレームを抽出する
- 応用例：
  サムネイル自動生成
  動画像のキーインデックス作成

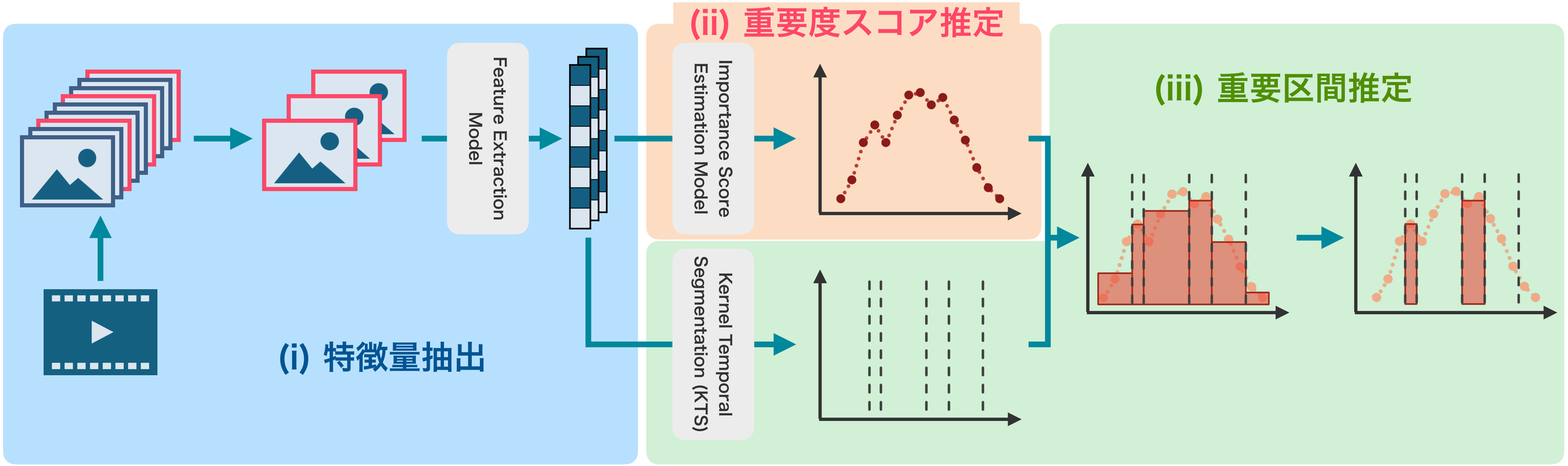
重要区間抽出



- 動画像から重要な区間を抽出する
- 応用例：
  スポーツのハイライト作成
  監視カメラ映像のダイジェスト

本研究では重要区間抽出を扱う

動画像要約のパイプライン



(i) 特徴量抽出　　：全フレームを復号し，等間隔にサンプリングしたフレームから特徴量を抽出

(ii) 重要度スコア推定：特徴量系列を入力として，各フレームの重要度スコアを推定

(iii) 重要区間推定　：特徴量系列を用いて区間を分割し，
重要度スコアの合計が最大となる区間の組を選択

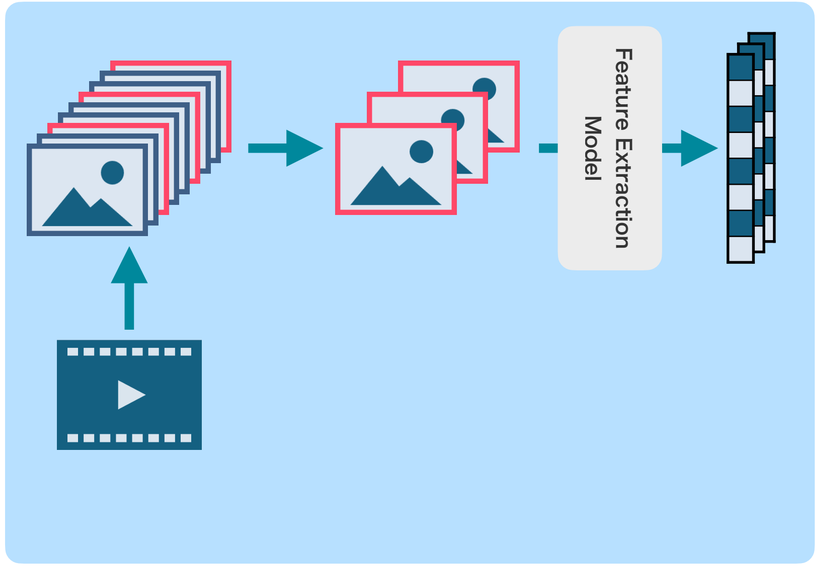
動画像要約のパイプライン

(i) 特徴量抽出

- 動画像の全体を復号してフレームを抽出
- 全フレームを等間隔にサンプリングする

  既存手法の多くは2fpsでサンプリング[Zhang+, ECCV 2016]
- サンプリングされた各フレームに対して特徴量抽出を実行

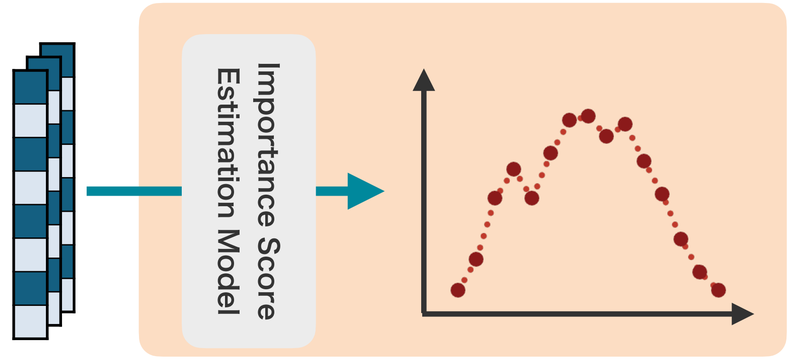
  ImageNet[Deng+, CVPR 2009]で事前学習されたGoogLeNet[Szegedy+, CVPR 2015]を利用するのが一般的



動画像要約の既存手法

(ii) 重要度スコア推定

- 特徴量系列から各フレームの重要度を推定
- 事前知識に基づく手法とデータ駆動型手法に大別



事前知識に基づく手法

- 動きや顔などの手がかりから各時刻の注目度を求める手法[Ma+, ACM MM 2002]
- フレーム特徴量行列の特異値分解に基づく手法[Gong+, CVPR 2000]

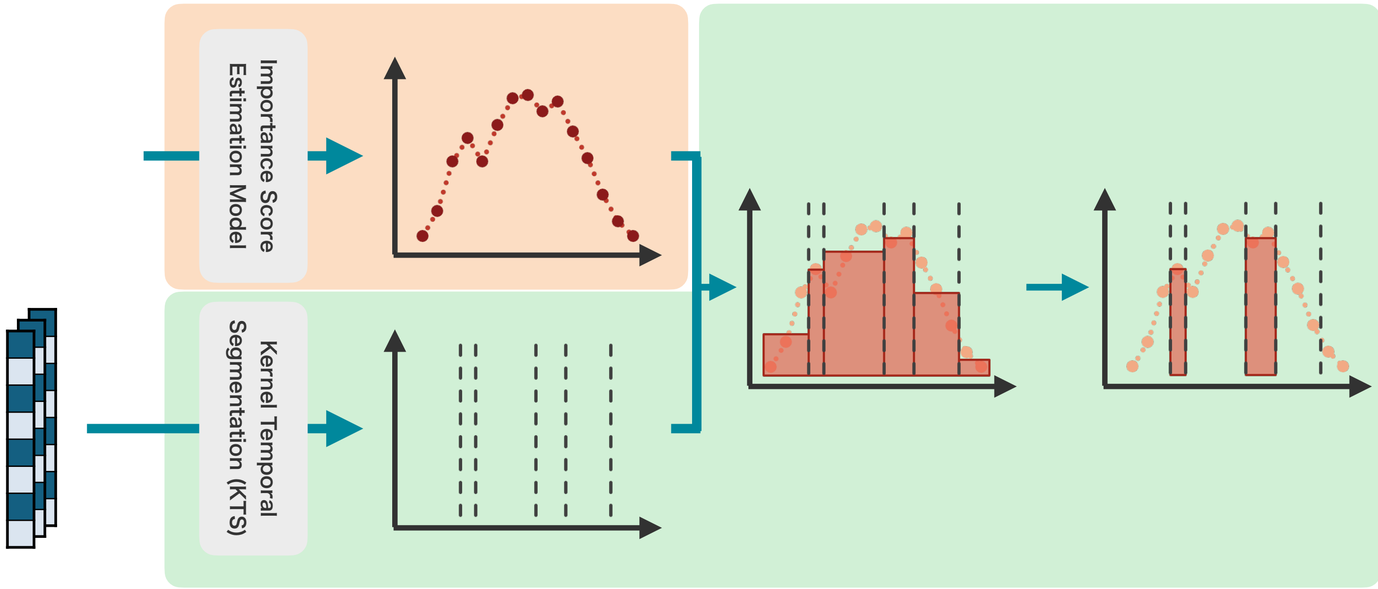
データ駆動型手法

- フレーム間の多様性に基づく手法[Zhou+, AAAI 2018]
- データセットの真値を学習に利用する手法[Fajtl+, ACCV 2018 W], [Zhu+, IEEE TIP 2021]
  ,[Apostolidis+, IEEE ISM 2021]

動画像要約のパイプライン

(iii) 重要区間推定

- 特徴量系列を用いて動画を意味的な区間に分割
  典型的にはKernel Temporal Segmentation （KTS）が利用される
- 重要度スコアをそれぞれの区間に割り当て
  区間内に含まれるフレームの重要度スコアの平均を区間のスコアとする
- 設定された要約の長さのもとでスコアが最大となる組み合わせを選択



既存手法の課題

実行時間における課題：

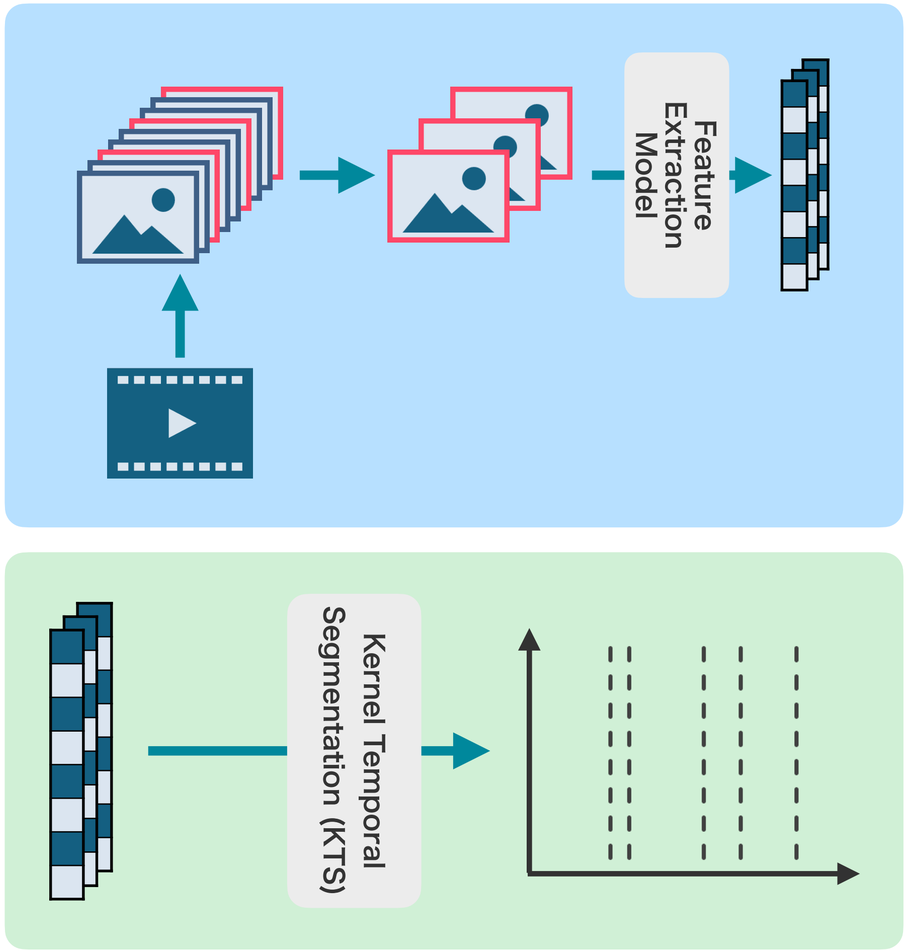
(i) 特徴量抽出

- 固定間隔のサンプリングは
  全フレームの復号を必要とする

(iii)重要区間推定

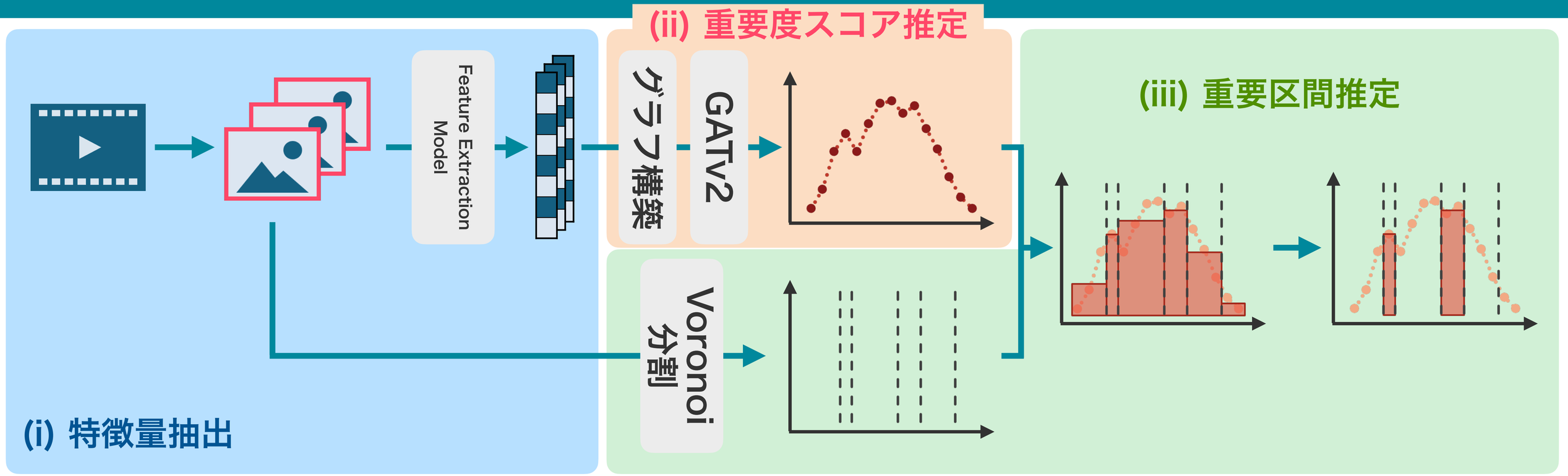
- KTSは特徴量系列を必要とする
- KTSは系列長の増加に対して急速に増大

→既存手法は動画像の全フレームの
復号が前提



リアルタイム処理や大規模データの一括処理では
これらの処理がボトルネックに

提案手法: Overview

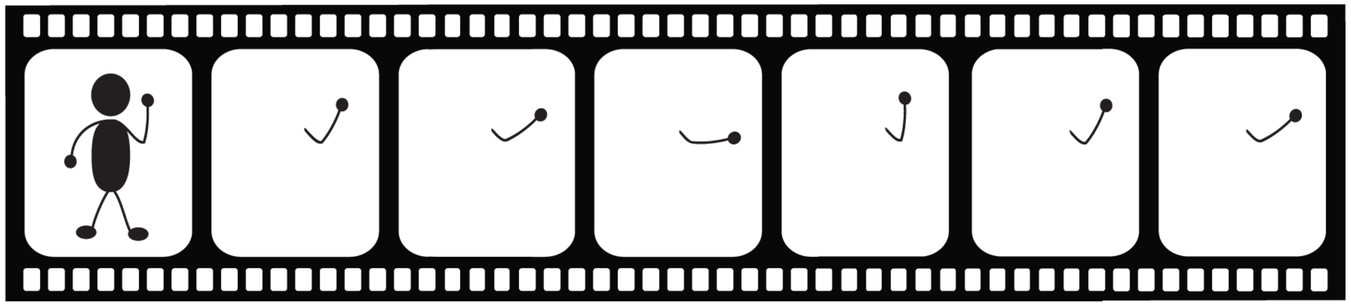


	提案手法	既存手法
特徴量抽出	Iフレームのみ抽出→特徴量計算	全フレームを復号→特徴量計算
重要度スコア推定	GATv2	(手法によって異なる)
重要区間推定	Voronoi 分割→0/1ナップサック	KTSを用いた区間抽出→0/1ナップサック

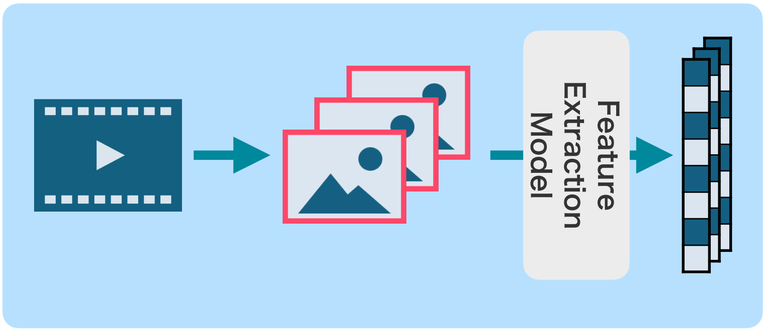
提案手法

(i) 特徴量抽出

- 動画像のイントラ（I）フレームのみから特徴量を抽出
  他のフレームに依存しないため，Iフレームは単独で復号可能
- Iフレームのみを特徴量抽出モデルに入力
  ImageNetで事前学習されたMobileNetV3-Small[Howard et al., ICCV 2019]を使用



www.bhphotovideo.com

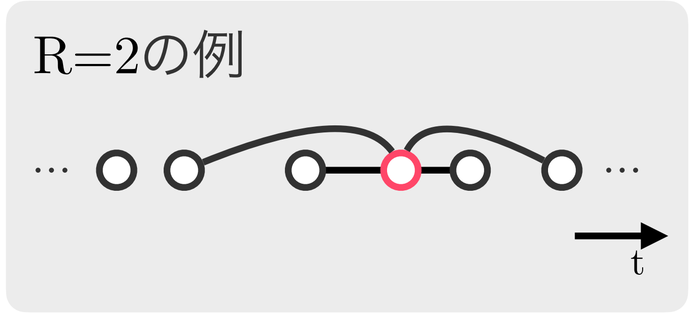


提案手法

(ii) 重要度スコア推定

※Iフレームは時間的に不均一に存在

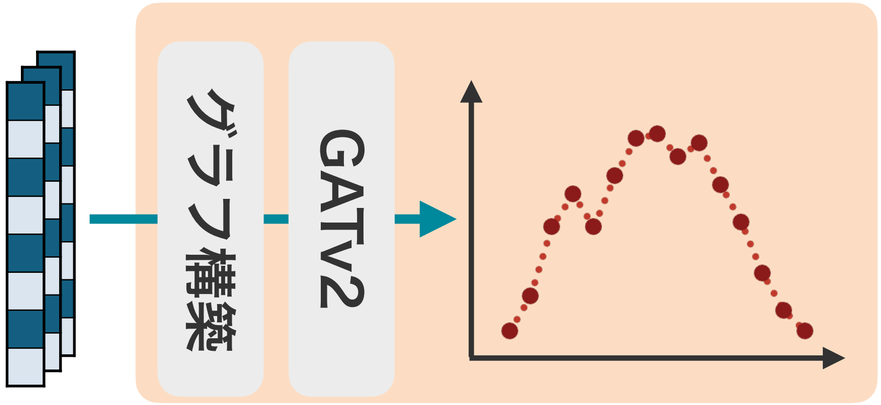
- 各Iフレームを頂点とするグラフを構築
  前後R個の頂点と接続
- 頂点の特徴量は(i)で計算した視覚特徴量と時刻情報



$$\left[\mathbf{z}_t; \left[\frac{t-1}{T-1}, \frac{\Delta\tau_t}{\max_{1\le t'\le T}\Delta\tau_{t'}}\right]^\top\right]$$

視覚特徴量　インデックス　直前のIフレームからの時間差

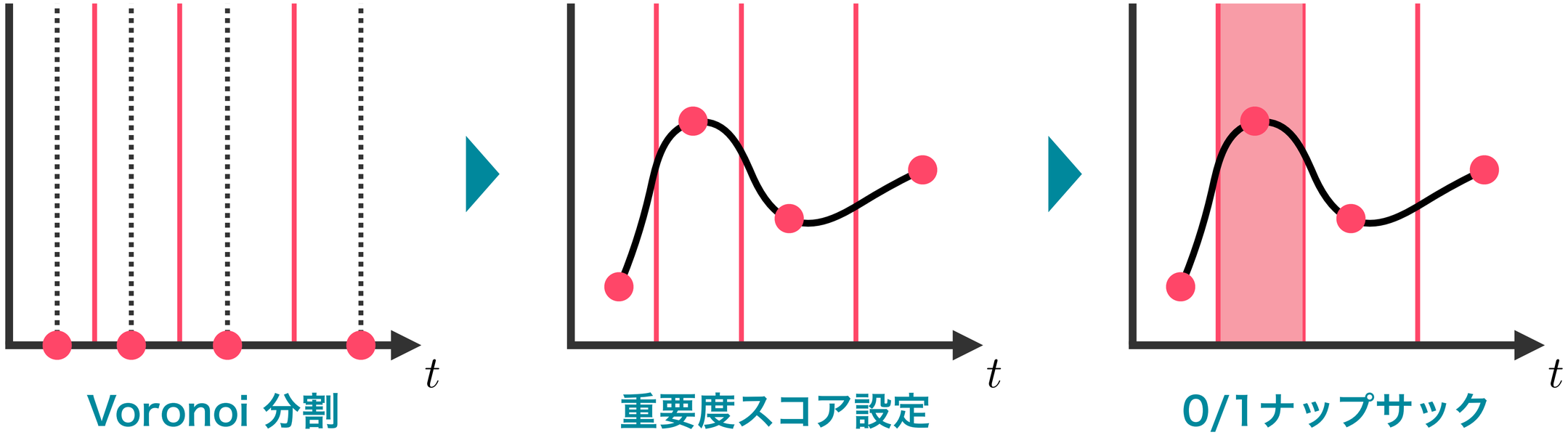
- GATv2を利用し，頂点間の各辺の重みと重要度スコアを学習



提案手法

(iii) 動画分割・重要区間推定

- 隣接するIフレームの中点を境界として動画像を分割（Voronoi 分割）
- 区間の重要度スコアとしてIフレームのスコアを設定
- 要約長が元動画像の 15% 以下という制約のもとで,
  重要度の合計が最大となる区間の組を 0/1 ナップサックにより選択



実験

データセット

SumMeデータセット[Gygli+, ECCV 2014]

YouTube上の25本の動画

各動画に対して15名以上のアノテータ

比較手法

- DR-DSN[Zhou+, AAAI 2018]
- VASNet[Fajtl+, ACCV 2018 W]
- DSNet-AB[Zhu+, IEEE TIP 2021]
- DSNet-AF[Zhu+, IEEE TIP 2021]
- PGL-SUM[Apostolidis+, IEEE ISM 2021]

実験内容

- 要約性能の比較

  真値とのF1スコアによって評価
- 実行時間の比較

  前処理（特徴量抽出）

  推論

  後処理（区間分割）

結果：F1スコアおよび実行時間の比較

※F1は5試行の平均と標準偏差

Method	F1	Time[s]
DR-DSN	0.40 ± 0.01	2.61
VASNet	0.38 ± 0.02	2.61
DSNet-AB	0.41 ± 0.01	2.61
DSNet-AF	0.43 ± 0.01	2.61
PGL-SUM	0.43 ± 0.02	2.61
PathGAT-IFrame	0.38 ± 0.02	**0.11**

- **要約性能**：既存手法と比較してF1で最大0.05低下
- **実行時間**：全手法と比較して20倍以上高速化

結果：各処理における実行時間の比較

Method	Preprocessing[s]	Estimation [s]	Postprocessing [s]	Total [s]
DR-DSN	2.49	<0.01	0.12	2.61
VASNet	2.49	<0.01	0.12	2.61
DSNet-AB	2.49	<0.01	0.12	2.61
DSNet-AF	2.49	<0.01	0.12	2.61
PGL-SUM	2.49	<0.01	0.12	2.61
PathGAT-IFrame	**0.11**	<0.01	<**0.01**	**0.11**

- 推論時間は全て0.01秒以下
- 既存手法のほとんどの処理が特徴量抽出と区間分割に占める
- 提案手法はパイプライン全体の実行時間を削減

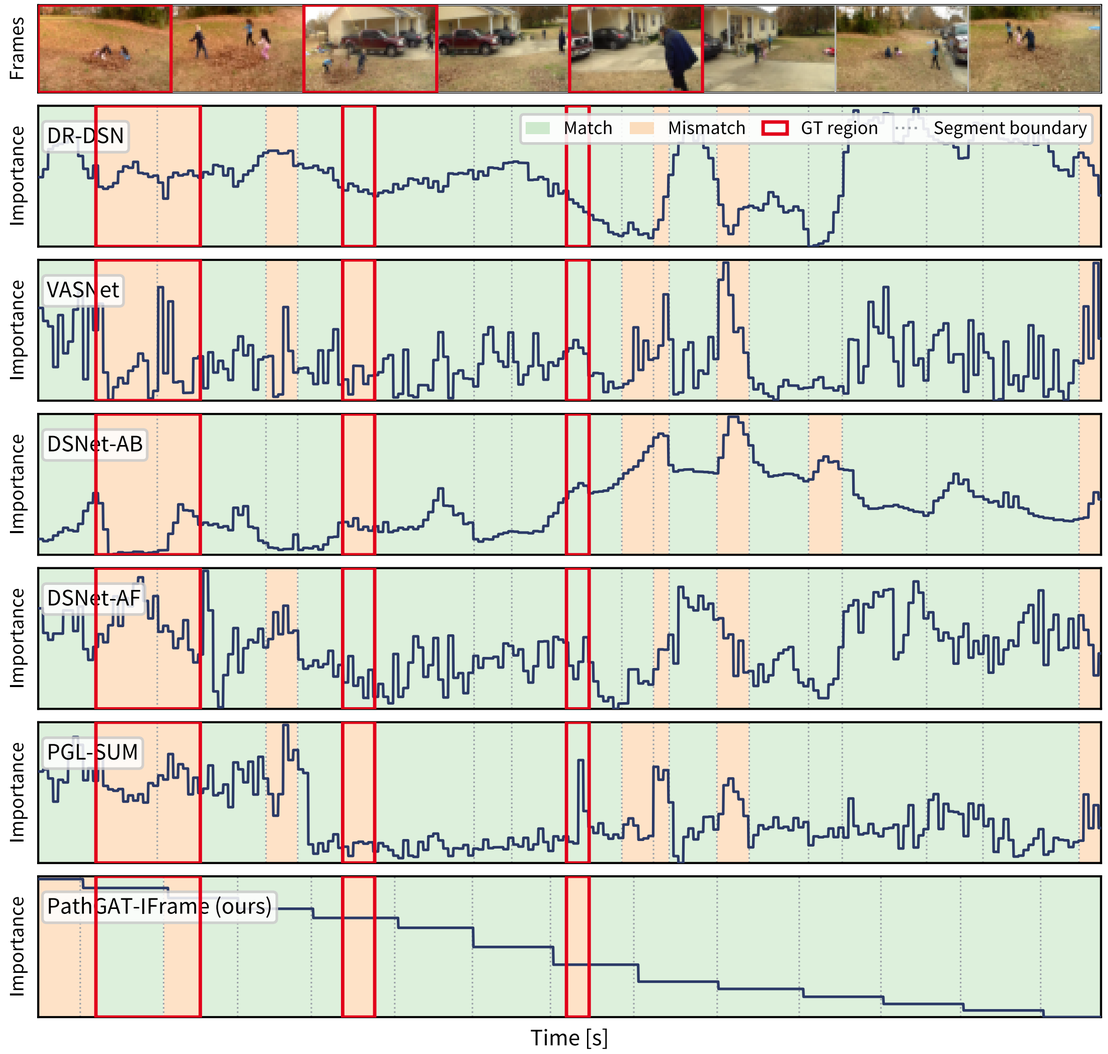
実験

結果：可視化結果

- 3個の重要区間が真値
  1個目：子供が落ち葉に飛び込む
  2個目：女性が登場
  3個目：女性がアップで映る

実験

結果：可視化結果（うまくいった例）



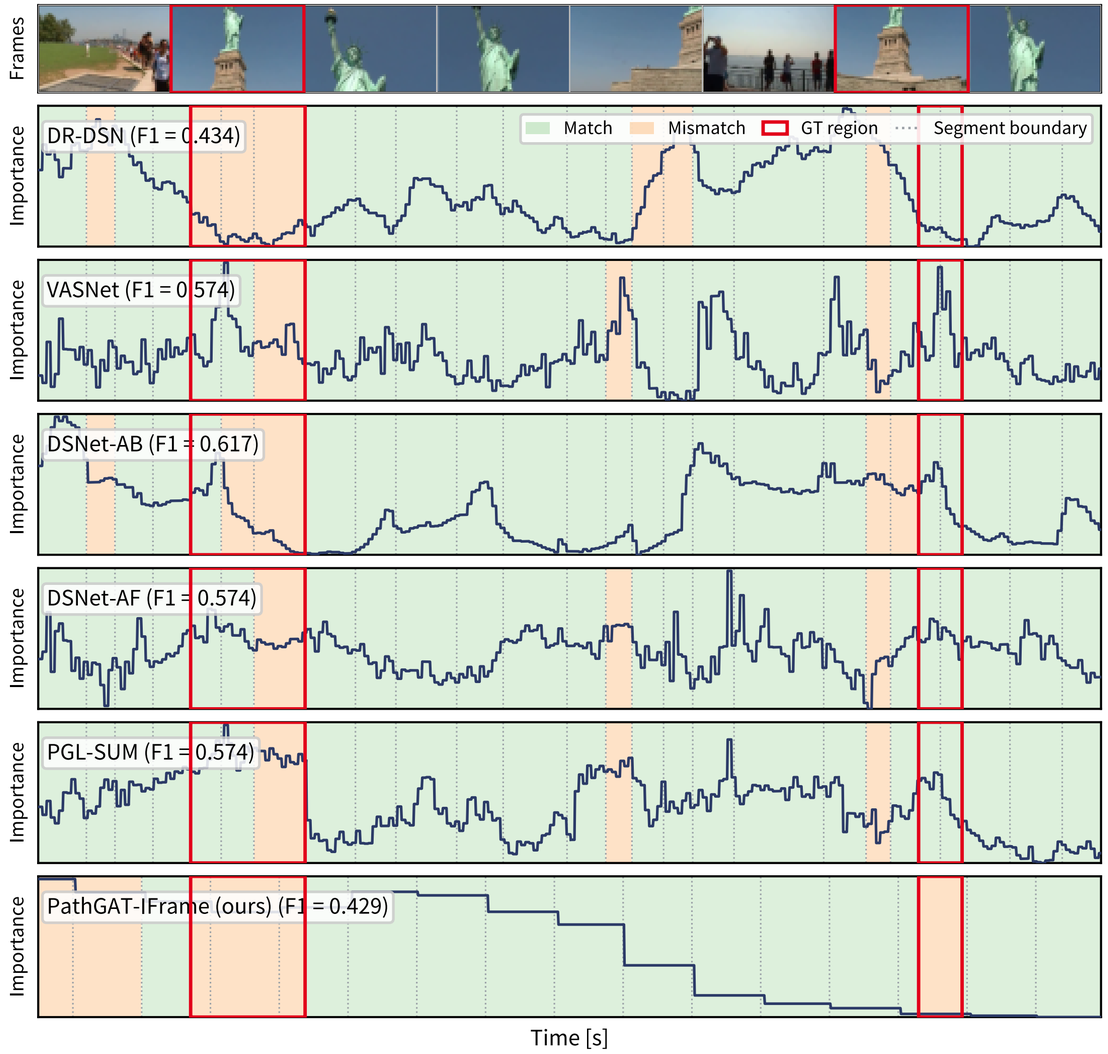
- 提案手法のみ最初のシーンを検出
- 2個目のシーンは既存の2手法のみ検出
- 最後のシーンは提案手法のみ未検出

→既存手法と同等の検出精度といえる

1個目：子供が落ち葉に飛び込む
2個目：女性が登場
3個目：女性がアップで映る

実験

結果：可視化結果（うまくいかなかった例）



- 重要範囲を2件とも見逃し
- KTSを用いた分割と比べて
  Iフレームを使った分割は数が少ない
  →Iフレームの数が推定精度を下げている？

まとめ

研究概要

イントラフレームのみを利用した**高速な動画像要約手法**を提案

実験

- SumMe データセットで既存手法 5種と予測性能・実行時間を比較
- F1 スコアは 0.38 ± 0.02 で，既存手法との差は最大 0.05
- 動画像 1 本あたりの実行時間は1/20以下

今後の展望

- 区間の選択まで含めた学習による重要度スコア推定の改善
- イントラフレームが少ない動画像に対する要約性能の改善
- 大規模な動画像要約データセットの整備