JSiSE Research Report
vol.37,no.2(2022-7)

講義アーカイブにおける関心領域の自動推定手法の提案

Yang Yuhui[*1], 太田 光一[*1], 長谷川 忍[*1]
*1 北陸先端科学技術大学院大学

A Proposal for Automatic Region of Interest Detection of Lecture Archives

Yuhui Yang[*1], Koichi Ota[*1], Shinobu Hasegawa[*1]
*1 Japan Advanced Institute of Science and Technology

This research proposes an automatic region of interest (ROI) detection architecture with a deep neural network for predicting the learners' ROI from the lecture archives to generate content to fit smaller screens like smart devices. To achieve this goal, we applied an Encoder-Decoder architecture that combines U-Net and Resnet with lecturer's action features as input to build a deep neural network model for predicting ROI. Through the experiment, the agreement between the ROI labels and the predicted regions was evaluated by Dice loss using and improved from 0.9 in a single image as a baseline to 0.4 in Openpose and temporal features.

キーワード: 講義アーカイブ，関心領域(ROI)，アイトラッキング，Optical Flow，OpenPose，深層学習

1. はじめに

タブレット端末やスマートフォンを利用して「いつでも，どこでも，誰とでも」学習できるスマートラーニングは，新型コロナウイルスの感染拡大による対面講義の制限下において，重要な役割を果たしている．対面講義を固定カメラ・マイクなどで収録する講義アーカイブは，制作における時間的コストの観点からスマートラーニングに対する有力なコンテンツの一つである．北陸先端科学技術大学院大学（以下，本学）においても 2006 年度から年間 1,000 コマ以上の対面講義を講義アーカイブとして収録し，学習者のための補完的な学習環境として提供している(1)．こうした講義アーカイブには講義中の講師の行動が把握できるように，ホワイトボードやスクリーンを含む講義室前方が含まれている．一方で，スマートデバイスの画面サイズの制約により，学習者が講義中に関心のある領域(Region of Interest: ROI)を切り替えつつ注視しながら学習することは容易ではない．

本研究の目的は，講義アーカイブにおける講師の動作の特徴量から学習者の ROI を推定することである．従来の研究では講師の位置を ROI として定義しているものがほとんどである(2-5)．しかしながら，Zhang らによると，講義アーカイブを視聴中の学習者の注意は，講師の位置に限定されず，ホワイトボードやスライドなど，様々な領域を切り替えて視聴している(6)．

本稿では，まず学習者の ROI を，講義アーカイブの学習過程におけるアイトラッキングデータに基づいて作成した小規模なデータセットについて紹介する．1 秒間に分割されたセグメントに対して得られたアイトラッキングデータに対して，K-means によるクラスタリングとスムージングを行った結果，16,039 個の ROI ラベルを収集した．次に，セグメント中の講師の動作を複数のアプローチで特徴マップとして抽出し, U-Net と Resnet を組み合わせた Encode-Decoder アーキテクチャを適用することで，ROI ラベルを予測する深層学習モデルを構築した．

予備的な評価実験を通して，ROI ラベルと予測領域の一致度を，各特徴マップを用いて Dice loss で評価したところ，Openpose と時間特徴を利用することで一定の改善が見られた．現時点での精度はまだ十分であるとは言えないが，講義アーカイブ中の講師の動作から学習者の ROI を適切に予測できれば，ROI 部分のみを切り出してスマートデバイスの画面に表示することが

可能となり，スマートな学習環境の利用に貢献することが期待される．

2. 関連研究

2.1 講義アーカイブにおけるアテンションマップ

Zhang らは，アイトラッキングとデータ可視化技術を応用し，アテンションマップと呼ばれる，学習者の視覚的な注意の分布を可視化したものを提案した[6]．学習者のアテンションマップと講師の行動の関係を分析した結果，講師のジェスチャーや視線などの行動が，学習者の視線を効果的に誘導できることを示した．また，学習者の視線座標は，講師，ホワイトボードの書き込み，プレゼンテーションのスライドに集中していた．しかしながらこの研究では，講義アーカイブから ROI を予測するモデルは含まれていない．

2.2 講義アーカイブにおけるアクティブカメラ制御

Zobel らは，Active Camera Control を利用して移動体を追跡しながら録画する手法を提案した[7]．また，Mukhopadhyay らは，講義内容の記録時に IR ビーコンを用いて講師を追跡している[8]．ただし，これらの自動制御カメラは，講師がカメラのアングルから外れた場合に追跡が困難となる課題がある．また，講師以外の ROI は想定されていない．さらに，このような特殊なカメラを各講義室に設置するのはコストがかかることに加えて，すでに録画されている講義アーカイブには適用することができない．

2.3 講義アーカイブにおける仮想カメラ制御

Sun らは，講義室内の複数の固定カメラ映像を合成した高解像度パノラマ映像から，講師を ROI として自動抽出し，人間の操作を模した仮想カメラ制御を行うシステムを開発した[2]．Dickson らは，講師のみがアーカイブ内を移動していると仮定し，高解像度映像から講師領域を切り出したコンテンツを自動生成した[3]．Yokoi らは，高解像度の講義映像から ROI を追跡し，疑似カメラパン期間を検出し，仮想カメラワークを算出した[4]．Mavlankar らは，対話型オンライン講義視聴システム ClassX を開発した[5]．ClassX では，人間のカメラオペレータを模倣したトラッキングモードと，ホワイトボードなどの特定の領域を予め指定するプリセットモードの 2 種類の ROI 設定が可能である．

このように，講師の位置を ROI として推定する手法が様々な研究で提案されている．高解像度の講義アーカイブから ROI を切り出すことで，ユーザとデバイスのインタラクションを最小限に抑えることができ，小さな画面に対応し，低帯域幅に適したコンテンツを生成することが可能である．しかしながら，2.1 節で述べた通り，講義アーカイブにおいて学習者の注意を引く ROI は講師領域に限られるわけではない．そこで本研究では，学習者の視線情報に基づいてより実践的な ROI を定義し，講義アーカイブの特徴量からその ROI を推定する方法を提案することを目指す．

3. データ収集

3.1 対象講義

本研究では，本学の学習管理システムである JAIST-LMS 上で配信された 2019 年度講義である遠隔教育システム工学の講義アーカイブを対象とした．講義アーカイブは，講義室における対面講義を固定された天井カメラ・マイクにより講義室前方を収録したものであり，講師の話やスライド，ホワイトボードを利用した知識伝達を中心とした 1 回約 100 分のものであった．なお，音声データには教室内の雑音が含まれており，学習者による視聴には問題ないものの，音声認識には適さないものであった．録画ファイルは MP4 形式で保存されており，解像度は 1920x1080，フレームレートは 30fps であった．また，スライドコンテンツは，図 1 に示すように，アーカイブの左下隅に統合されていた．学習者は PC やスマートデバイスから学内ネットワークを通じてアクセスすることができる．

図 1 オリジナルの講義アーカイブ

3.2 アイトラッキングデータの収集

学習者の ROI を定義するために，サンプリングレート 30Hz，精度 0.5-1°，動作距離 45-75cm で画面上の注視点情報を収集できるアイトラッカーである The Eye Tribe をモニタ下部に設置して利用した．注視点座標は，モニタ座標系で与えられる(x, y)の組で表現される．参加者がモニタ外に焦点を合わせた場合，注視点座標は左下隅(0, 0)に設定される．データ収集時の疲労を考慮して，1 回の講義アーカイブを 90 秒単位で 48 個に分割して利用した．

4 名の大学院生が，それぞれ異なる 12 個の動画を視聴し，注視点が検出されない等の異常値を除去した結果，128,461 点の注視点座標データを収集した．

3.3 ROI ラベリング

1 秒間 30 フレームを 1 セグメントとした時に得られた注視点座標群を，K-means を用いて 3 つのカテゴリにクラスタリングした．これは，生データを ROI，ノイズ，画面外に分類するためである．図 2 に示すように，最も多くの注視点座標を含むクラスタに矩形を設定した．これは，セグメント内で参加者の注視時間が最も長いクラスタと言える．これにより，生データからジッタやノイズを除去した．また，同じサイズを保つために，矩形の重心に対して 500×300 ピクセルの領域を ROI ラベルとしてデータベースに保存した．このようにして，最終的に 16,039 個の ROI ラベルをアノテーションした．

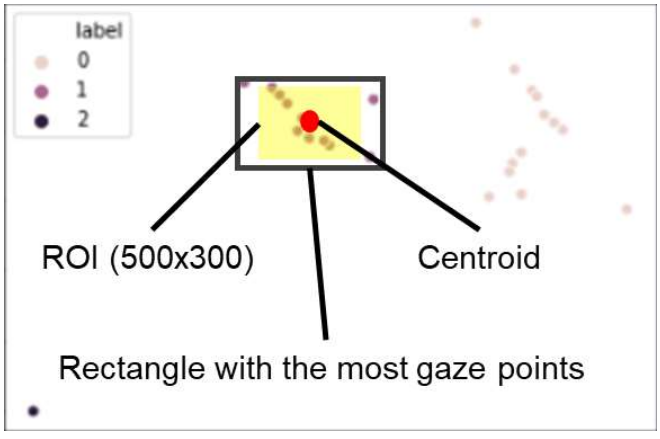

図 2 ROI ラベリング

図 3 にラベル付き ROI の例を示す．上の元画像は，講師が話し始めたタイミングであり，ROI ラベルは講師の周辺となっている．一方，下の画像では，スクリーン上でスライドの説明をしているが，左下のスライド領域が ROI となった．このように，講師の周辺以外の領域にも ROI が設定された．

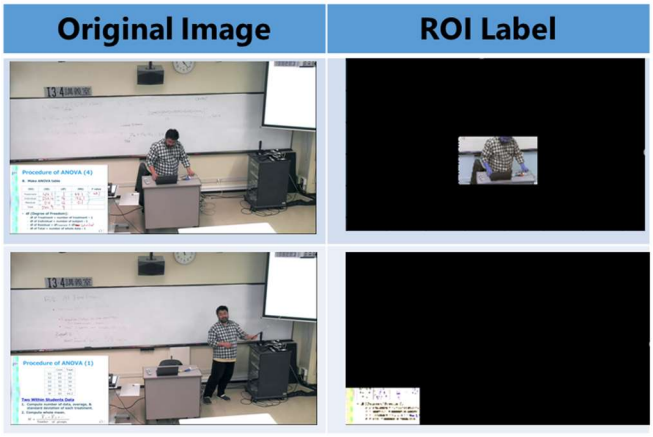

図 3 ROI ラベリングの例

4. 提案手法

4.1 特徴マップ

講義アーカイブの元の映像には冗長な情報が多く含まれている．学習者の ROI を推定するためには背景情報よりも講師の活動が重要であるという仮説のもと，以下の方法で 1 秒間のセグメント動画の特徴を表現する画像である特徴マップを生成した．

4.1.1 フレーム差分特徴

映像中の冗長な特徴を減らし，時間的に変化する特徴を反映させるために，図 4 に示すようにセグメント動画の最初と最後のフレームの差分を利用する．これにより，映像から背景情報を除去することができる．同様に，左下のスライド領域の変化も抽出する．

図 4 フレーム差分特徴

4.1.2 Optical Flow 特徴

Optical Flow とは，視聴者とシーンとの相対的な動きによって生じる，各セグメント内の物体，表面，エッジの見かけ上の動きのパターンである[10]．画像の明るさが一定で動きが小さいとき，近傍の空間的・時間的な勾配を用いて Optical Flow ベクトルを計算する．講義中はほとんど講師しか動いていないため，OpenCV の cv2.calcOpticalFlowPyrLK [11]を 1 秒間の動画に適用することで Optical Flow の特徴を抽出し，図 5 の緑チャネルとして保存した．

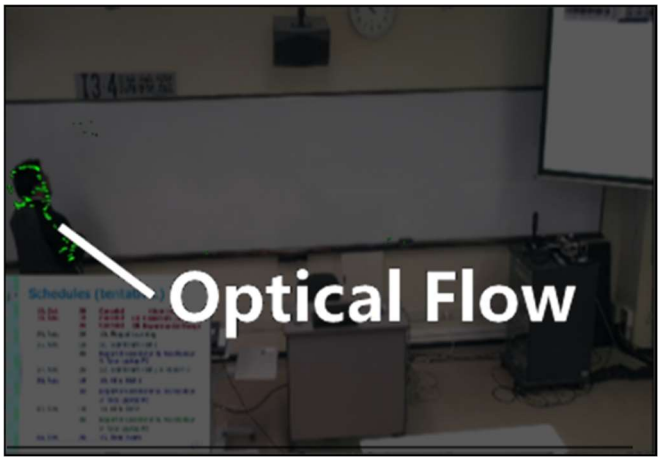

図 5 Optical Flow 特徴

4.1.3 OpenPose 特徴

前節で述べた Optical Flow 特徴は，映像の特徴点に依存する．そのため，講師の着衣の違いやスライドの変更などにより、結果が不安定になる場合がある．そこで，OpenPose 特徴では，画像中の人間の 2 次元姿勢をリアルタイムに検出する OpenPose[12]を採用した．Openpose はまず，各フレームにおいて講師の身体のキーポイントと関節からなるポーズグラフを生成する．セグメント動画に対して生成したポーズグラフから Optical Flow を生成し，図 6 のように追加した．これにより，講師の着衣や動作にロバストな特徴マップが構成される．

4.1.4 時間的特徴

上記の特徴量では時間情報が失われているため，講師の活動方向を特定することができない．そこで，時間情報を保持するために，図 7 で示すように，青チャンネルでの Optical flow と OpenPose のフレームごとの変化を格納した．色域は 0-255 で，1 秒間のセグメント動画は 30fps であるため，フレームインデックス i の RGB カラーは(8*(i-1), x, y)とした．これにより青色の変化で時間的な前後を視覚化することができる．

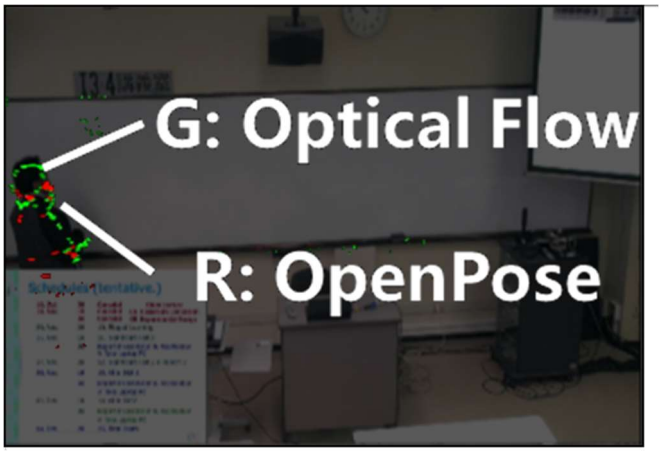

図 6 OpenPose 特徴

図 7 時間的特徴

4.2 予測モデル

4.1 節で生成した特徴マップから ROI を予測するために，図 8 に示すように，U-Net の Encoder-Decoder モデルを用いた深層学習アプローチ[13]を採用した．

U-Net は FCN（fully convolution network）の一つで，画像の意味的なセグメンテーション（オブジェクトの位置）を推定するためによく利用されるネットワークである．U-net は一般的な画像分類タスクとは異なり，特定の特徴を持つ領域を出力することができるため，今回の ROI 推定タスクに適していると考えられる．

U-Net のエンコーダは，入力画像を複数回畳み込み，画像の特徴を抽出する．エンコーダとしては，He らが提案した ResNet-34 エンコーダを採用した[14]．ResNet

は深い多層を用いて入力と出力の間の残差表現を学習する．これにより，層数を増やしながら精度を向上させることが期待される．

U-Net のデコーダは，エンコーダで抽出された特徴を，デコンボリューションと呼ばれる逆畳み込みを行い，入力画像と同じ大きさの確率マップを出力する．最初は広く，次にそのユニットやコネクションが中央に向かって絞られ，そしてまた広がっていくアーキテクチャにより，学習済みモデルの汎化性能の向上が期待される．

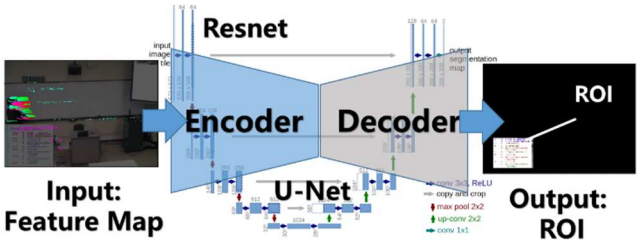

図 8 U-Net アーキテクチャ

5. 予備実験

5.1 実験設定

本予備実験では，提案する ROI 推定アルゴリズムの品質を，以下の異なる特徴マップに対して検証することを目的とする．

1. 1 枚目の画像のみ（ベースライン）
2. フレーム差分特徴
3. Optical Flow 特徴
4. Optical Flow + OpenPose 特徴
5. 時間的特徴

本実験では，これらの特徴マップを以下の設定で学習させ，その品質を図 9 に示す Dice loss により比較した．Dice loss は画像分割のための損失関数であり，ラベルと予測領域がどれだけ重なっているかを表す．

- フレームワーク: pytorch
- 前処理：リサイズ，正規化
- モデル：Unet+ResNet-34
- データ数：16,039
- トレーニング／テスト：0.8/0.2（ランダム）
- Learning Rate：0.001
- Epoch 数：30
- Activation：sigmoid

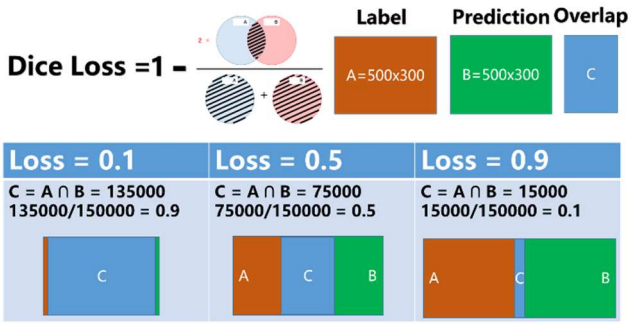

図 9 Dice Loss

5.2 結果

表 1 に各特徴マップにおけるテストデータの平均 Dice loss 結果を示す．

表 1 特徴マップ毎のテストデータの平均 Dice loss

特徴マップ	平均 Dice Loss
1. ベースライン	0.92
2. フレーム差分特徴	0.82
3. Optical Flow 特徴	0.43
4. Optical Flow+OpenPose 特徴	**0.41**
5. 時間的特徴	**0.41**

1 枚目画像によるベースライン実験では，平均 Dice loss が 0.92 となり，学習に失敗した．ROI 推定に必要な講師の動作の特徴を静止画から抽出することは難しいと言える．

フレーム差分特徴では，平均 Dice loss が 0.82 とベースラインよりは改善したが，2 フレーム間のわずかな輝度差により，スライドが変化していないにも関わらず不安定な差分が発生する問題があった．

Optical Flow 特徴では，平均 Dice loss が 0.43 となり，ある程度推定に利用できる可能性が示された．ただしこの特徴は，コーナーポイントから生成される特徴マップから ROI が予測できることを想定しているが，コーナーポイントの検出は学習サンプルに依存する．例えば，同じ動作であっても異なる服装で行った場合，異なる結果となる可能性がある．また，講師が背景やホワイトボードと同系色の服を着ているとコーナーポイントが正しく検出できなかったり，スライド内の文章や図が不要に検出されたりする問題もあった．

Optical Flow に OpenPose を追加した特徴の平均 Dice Loss は 0.41 であり，若干改善された．これは，講師のコーナーポイント検出に比べ，ポーズグラフが比較的安定していたことが一因と考えられる．

また，時間的特徴による平均 Dice Loss は 0.41 であった．学習損失は他の条件よりも良好であったが，テスト損失は過学習する傾向にあり，実験条件の再考やデータオーギュメンテーションの必要性が示唆された．

5.3 考察

図 10 に，4. Optical Flow + OpenPose 特徴におけるラベルと予測結果の例を示す．全体として，予測結果は講師やスライドの周りに集中しており，特に講師が特定の位置に指示するようなケースは改善の余地がある．例えば，最も損失が大きいケースでは，講師は右上のスクリーン領域を指しているが，学習者は左下の組み込み PC の画面に焦点を合わせている．このような講師の特別なポーズは，適切に推定できる必要がある．また，ホワイトボードや講師，PC 画面以外の領域はほとんどの場合は注視点の対象とならないため，これらを自動的に対象外とする処理を追加することも効果的であると考えられる．

ベースライン法を他の手法と比較すると，学習者の ROI は静的な背景情報よりも講師の活動に対して敏感であることがわかる．さらに，Optical Flow による講師の動きの追跡は，ROI と強く関連していると考えられる．しかしながら，時間的特徴を追加しても結果の改善にはほとんどつながらなかった．これは，データ数の観点からセグメント動画を 1 秒間とした結果，時間的特徴が十分に活かされない設定になったことが一因であると考えられる．今後はより大規模かつ一定時間のデータセットを準備し，対応する手法を適用する必要がある．

また，提案手法では，OpenPose で得られたポーズグラフの Optical Flow 特徴を全て R チャンネルで入力した．その結果，4 の特徴マップから講師の体の部位の情報が失われた．OpenPose は体の部位ごとに色が異なるポーズグラフを出力できるため，前述のように講師が特定の部位を指していることを認識するために，姿勢と体の部位の情報を保持した特徴マップを構築することも考えられる．

なお，今回のデータセットでは，予備的な実験として 1 回の講義の学習者の注視点座標データから ROI ラベルを生成した．そのため，講義スタイル，講師，講義室のレイアウト，学習者が異なると結果が異なる可能性がある．また，ROI ラベルの固定サイズが比較的

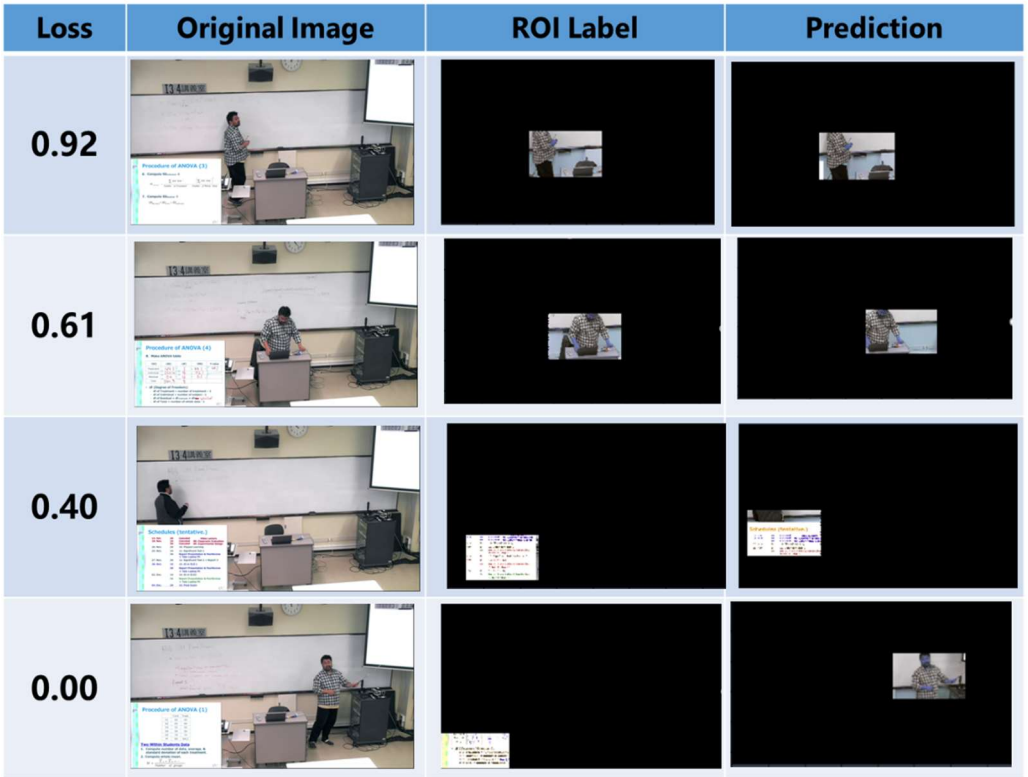

図 10 実験結果

小さかったため，スライド領域をカバーするようにROI ラベルのサイズを拡張することが現実的であると考えられる．

6. おわりに

スマートデバイスのような狭小画面に適した学習用コンテンツを低コストで生成するために，講義アーカイブからROIを自動予測するための深層学習アーキテクチャを提案した．この目的を達成するために， 本学の高精細・固定カメラアングルで収録された講義アーカイブを対象に，学習者のアイトラッキングデータを収集・クラスタリングすることで，1 秒間のセグメント動画単位で 16,039 個の ROI ラベルのデータセットを収集した．次に，フレーム差分，Optical flow，OpenPose，時間的特徴といった複数の特徴マップを設計し，セグメント動画から講師の動作を抽出した．そして，U-Net と ResNet を用いて，特徴マップから ROI を推定するエンコーダデコーダモデルを提案した．予備的な実験を通して，テストデータにおいて，Optical flow+OpenPose と時間特徴量の特徴マップに対する平均 Dice Loss は 0.41 であった．損失の観点からは精度に改善の余地が見られるが，ROI 予測によるスマートデバイス向けコンテンツの自動生成に対して一定のポテンシャルがあることがわかった．

今後の課題としては，特徴量の拡張が考えられる．現在は講師の動作が中心であったが，ホワイトボードやスライドの変化率，スライド上のマウスの動き，講師の指示ポーズなど，さらなる特徴量の抽出が考えられる．また，今回は特徴量マップとして一定期間の変化量を表す静止画を利用したが，今後は，Recurrent Neural Networks のような時系列データを扱える手法も検討すべきである．さらに，今回提案したモデルで予測したROIは1秒間のセグメント動画に対する予測であるため，そのまま適用すると結果が不安定になる．動画のクリッピングをスムーズに遷移させることは，今後の重要な課題である．また，より汎用的な予測モデルを構築するためには，異なる講義テーマや講師などに関する視線データをより多く収集することが必要不可欠である．

参 考 文 献

(1) S. Hasegawa et al., Case studies for self-directed learning environment using lecture archives, Proc. of The Sixth IASTED International Conference on Web-based Education, pp. 299-304, (2007)

(2) X. Sun, et al., Region of interest extraction and virtual camera control based on panoramic video capturing, IEEE Transactions on Multimedia, 7(5), pp.981-990, (2005)

(3) P. E. Dickson, et al., Automatic creation of indexed presentations from classroom lectures, Proceedings of the 13th annual conference on Innovation and technology in computer science education, pp.12-16, (2008)

(4) T. Yokoi and H. Fujiyoshi, Virtual camerawork for generating lecture video from high resolution images, IEEE International Conference on Multimedia and Expo, (2005).

(5) A. Mavlankar, et al., An interactive region-of-interest video streaming system for online lecture viewing, 18th International Packet Video Workshop, pp. 64-71, (2010)

(6) J. Zhang, et al., The effects of video instructor's body language on students' distribution of visual attention: an eye-tracking study, Proceedings of the 32nd International BCS Human Computer Interaction Conference 32, pp.1–5, (2018)

(7) M. Zobel, et al., Entropy based camera control for visual object tracking, Proceedings of International Conference on Image Processing, 3, pp.901-904, (2002)

(8) S. Mukhopadhyay and B. Smith, Passive capture and structuring of lectures, Proceedings of the seventh ACM international conference on Multimedia (Part 1), pp.477–487, (1999)

(9) The Eye Tribe, (2022/3/25 確認) https://theeyetribe.com/theeyetribe.com/about/index.html,

(10) B.D. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision, Proceedings of the 7th international joint conference on Artificial intelligence, 2, pp.674-679, (1981)

(11) OpenCV, calcOpticalFlowPyrLK(), (2022/3/25 確認). https://docs.opencv.org/3.3.1/

(12) Z. Cao, et al., OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, IEEE Transactions on Pattern Analysis and Machine Intelligence, 43, pp.172-186, (2021)

(13) O. Ronneberger, et al., U-net: Convolutional networks for biomedical image segmentation, International Conference on Medical image computing and computer-assisted intervention, pp.234–241, (2015)

(14) K. He, et al., Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, pp.770–778, (2016)