2ZA-07

VR 講義システムにおける 2D 講義動画の自動生成

服部 広大† 長尾 確‡
名古屋大学 大学院情報学研究科†‡

1. はじめに

VR を用いて講義を三次元的に記録し，2D 動画化して配信するオンライン講義システムが開発されている．一般の講義動画は講義の様子を 2D 情報として記録していたのに対し，本システムでは VR を用いて講義を 3D 情報として記録し，それをもとに 2D 講義動画を作成するため，あとからカメラワークを決めるなどの柔軟な動画作成が可能である．

講義動画は定点の映像よりも動きがある映像のほうが望ましい．しかし，現在は手動でカメラワークを設定する必要があり，その作業が手間となる．

そこで本研究では，機械学習を用いてカメラワークを生成し，生成したカメラワークをもとに講義動画を自動で作成することを目的とする．機械学習によるカメラワークの生成は，講義に用いたスライド画像，VR で記録される講師の動き，講義中のポインターの動きから，時系列情報を考慮して適切なカメラを選ぶ分類問題として解く．

2. VR 講義システム

本研究で用いる VR 講義システムは VR 発表練習システム [1]をもとに開発されたシステムである．本システムでは通常の講義と同じように PowerPoint とポインターを用いた講義を行うことが可能である．講師は VR ヘッドセット，コントローラ，両足にトラッカーを装着して，頭と両手両足をトラッキングした状態で講義を行う．VR コンテンツ中の動きを記録する VRec [1]を用いて，講師とポインターの動きを記録し，スライドはキャプチャして動画として記録する．

本システムで収録されたコンテンツは VR 内で視聴することも可能であるが，2D 動画に変換して広く配信することも可能である．現実の様子をカメラで記録する従来の講義動画と異なり，講義を VRec で三次元的に記録しておくことにより，収録後にカメラを自由に設定することが可能である．さらに動画化の際にカメラの切り替わりを指定してカメラワークを設定することが可能である．カメラワークを設定することで定点カメラに比べて変化のある講義動画を作成することが可能である．

選択可能なカメラの例として，図 1 左に示したスライドと講師を映した一般的なカメラ，右に示したスライドにズームしたカメラがある．その他にも，講師にズームしたカメラや自由にカメラ位置を指定することも可能で，ニーズに応じた 2D 講義動画の作成が可能である．

図 1 選択可能なカメラワークの例

しかし，現在のシステムでは，適切なカメラワークを設定するために，収録後に再度講義を見直す必要がある．さらに，適切なカメラを選ぶためには慣れが必要などの問題点がある．

3. カメラワークの自動生成

カメラワーク編集の要因として，説明しているスライドの内容，講師のジェスチャーなどが考えられる．それらの要素を説明変数とし，双方向の時系列情報を考慮するために BiLSTM を用いて，適切なカメラ番号をフレームごとに出力するモデルを作成した．分類するカメラは図 1 右に示したスライドズームとそれ以外のカメラである．

3.1. 入力データ

収録した講義データを 1 秒ごとに分割してモデルへの入力とする．スライド画像を入力とするために，学習済みの ResNet-152 による特徴抽出を行った．スライド画像特徴量は 2048 次元で，他の特徴量と比較して次元数が大きいため，次元数を下げるために全結合層を追加して 12 次元に変換する．講師のモーションデータは，VR2ML [1]を用いて座標，速度，加速度，回転角，角速度，角加速度といった特徴量に変換して入力とする．ポインティングの情報はスライドをポインティングしているかどうかの真偽値に変換して入力とする．上記の 3 つの特徴量を結合して BiLSTM に入力する．

Automatic Generation of 2D Lecture Videos in VR Lecture Systems
†HATTORI, Kodai(hattori@nagao.nuie.nagoya-u.ac.jp)
‡NAGAO, Katashi(nagao@nuie.nagoya-u.ac.jp)
†‡Graduate School of Informatics, Nagoya University

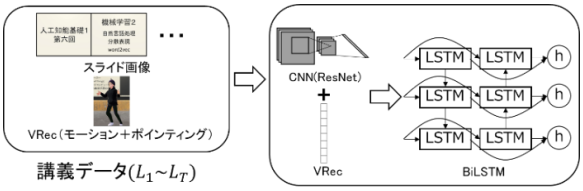

図 2 モデルの構造

3.2. ResNet の転移学習

ResNet は ImageNet で学習済みのモデルが公開されているが，今回の入力画像はスライド画像であり，一般的な画像分類における画像とは特徴が異なると考えられる．そのため，スライド画像の特徴量を抽出するために転移学習(fine-tuning)を行う必要があると考えられる．

転移学習のために，学習データとして研究発表で用いられたスライド 16967 枚を用意した．スライドズームを選択する要因として，スライド内の表や画像などの割合，文字の大きさなどが考えられる．それらを含んだ特徴を抽出するため，スライド内の画像割合，文字のジャンプ率を出力とするモデルを作成した．

多ラベル分類として解くためにラベル付けを行った．スライドに図が含まれるかどうかは重要な特徴であるため，画像割合が 0 とそれ以外で分割した．0 以外のラベルは，データ数が 5 分割されるように度数分布表を作成した．ジャンプ率は 6 分割されるように度数分布表を作成した．度数分布表により，画像割合とジャンプ率それぞれに 6 つのラベルを付けて多ラベル分類を行った．

4. カメラワーク分類モデルの学習結果

3 節のとおりにモデルを作成し，学習を行った．学習データとして，これまでに VR 講義システムで収録された，複数人による 1 回 20 分程度の講義データを計 46 回分用意した．

モデルの比較として，転移学習した ResNet の中間層 2048 次元を特徴量抽出に用いたモデル，転移学習の出力層を用いたモデル，転移学習なしの中間層を用いたモデルを用意した．10 分割交差検証により，モデルの評価を行った（表 1）．

表 1 特徴量抽出によるモデルの比較

	正解率	適合率	再現率	F 値	AUC
転移学習あり（中間層）	0.742	0.732	0.731	0.725	0.725
転移学習あり（出力層）	0.697	0.706	0.674	0.681	0.688
転移学習なし（中間層）	0.721	0.736	0.679	0.698	0.709

一般的にカメラが頻繁に切り替わりすぎる動画は好ましくない．今回のモデルの出力にはノイズが含まれるため，1 秒程度で切り替わるカメラワークが生成された．それらを除去するために，モデルの出力をフィルタリングして，短いカメラワークを削除する．表 1 で性能がよかった転移学習の中間層を用いたモデルにおいて，削除するカメラワークの長さごとに比較して評価を行った（表 2）．

表 2 フィルタサイズによる比較

フィルタサイズ	正解率	適合率	再現率	F 値	AUC
1 秒以下	0.757	0.745	0.747	0.739	0.739
2 秒以下	0.763	0.750	0.753	0.744	0.744
3 秒以下	0.763	0.750	0.749	0.743	0.743
4 秒以下	0.758	0.747	0.742	0.737	0.735

以上の結果から，一番精度が高いモデルは転移学習の中間層を特徴量抽出に用いたモデルから 2 秒以下のカメラワークを削除したものであり，転移学習の有効性と出力のフィルタリングサイズの最適な値がわかった．

適切なカメラワークを決めることは人手でも難しいため，カメラワークを評価するためには機械学習の精度以外の評価指標も必要となる．今後は機械学習の精度以外の観点からもモデルの評価を行っていく予定である．

提案したモデルによる自動生成はまだ十分と言えず，人手による確認や修正が必要となる．精度が低い原因として，学習データの不足が考えられる．学習データを収集するために，今後は提案モデルによってカメラワークを生成し，それを人手で修正する方法で運用を進めていき，学習データを収集する．

5. まとめ

本研究では，VR 講義システムで収録された講義データから，機械学習によりカメラワークを自動生成することで，2D 講義動画の作成を自動化する手法を提案した．複数のモデルを比較することで転移学習の有効性や適切なフィルタリングサイズを検証した．

現在の精度では人手による確認，修正が必要であるが，一からカメラワークを生成することに比べてコストは低くなった．今後運用を進めていくことで学習データを収集することが可能になり，精度向上が期待できる．

参考文献

[1] Y. Yokoyama , K. Nagao, “VR2ML: A Universal Recording and Machine Learning System for Improving Virtual Reality Experiences,” 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), 2020.