Vol.2021-MUS-130 No.1
Vol.2021-EC-59 No.1
2021/3/16

登壇発表を対象としたライブ中継のための マルチカメラ自動スイッチングシステムの提案

比佐 翔太[1]　竹川 佳成[1]　松村 耕平[2]　平田 圭二[1]　五十嵐 健夫[3]

概要：本研究では登壇発表を対象としたライブ中継のための自動スイッチングシステムの設計と実装を目的とする．近年，Covid-19 の影響に伴い，Zoom などを用いた学術発表やセミナーのライブ中継が一般的になりつつある．また，単一のカメラによる固定アングルでの中継は，視聴者の集中力が途切れやすかったり，離脱率が高い傾向にあるといわれており，マルチカメラによるスイッチングはプロフェッショナルなライブ中継において一般的に利用されている．適切なスイッチングには経験や知識が必要とされると同時に，人的リソースが求められる．本研究では，音声認識・文字認識・画像処理を用いて登壇発表中の各種イベントを自動認識する機能，Endo らが提案したスイッチングの状態遷移モデルにもとづきマルチカメラから最適なカメラ映像を選択する照合器をもつ自動スイッチングシステムを提案する．各イベントの認識モデルおよび照合器の妥当性を検証する評価実験を実施し，現時点の精度や課題を明らかにした．

1. 背景

近年，ZOOM，ニコニコ動画，YouTube などを利用した学術発表のライブ中継が盛んである．平井によると，学術発表のライブ中継には次のようなメリットがある[*1]．まず，当日参加できない研究者や普段研究会に参加しない研究者ではない人に対しても研究成果をアピールできるメリットがある．次に，アーカイブ映像により，新規の研究者や当日参加できなかった研究者も発表内容を知ることができるというメリットがある．近年，Covid-19 の影響に伴い，Zoom などを用いた学術発表やセミナーのライブ中継が一般的になりつつある．

学術発表やセミナーのライブ中継では，マルチカメラで同時撮影しその中から最適なカメラ映像を 1 つ選択し配信するというライブ中継がプロフェッショナルな現場では多用されている．例えば，専門業者に中継を委託した IPSJ-ONE ではスライド、ワイプ、発表者、発表全体という 4 つのカメラビューで撮影が行われていた [2]．マルチカメラで同時撮影しその中から最適なカメラ映像を 1 つ選択し配信する理由としては，視聴者に適切な情報を伝えるためであること，視聴者の集中力を保つためであること，離脱率の軽減させるためであることが挙げられる．適切なスイッチングには経験や知識が必要とされると同時に，人的リソースが求められる．

そこで本研究では，登壇発表を対象としたライブ中継のための自動スイッチングシステムの設計と実装を目的とする．本研究では，音声認識・文字認識・画像処理を用いて登壇発表中の各種イベントを自動認識する機能と Endo らが提案したスイッチングの状態遷移モデルにもとづきマルチカメラから最適なカメラ映像を選択する照合器をもつ自動スイッチングシステムを提案する．

2. 関連研究

スイッチングの自動化に関しては様々な研究がなされている．例えば，対話映像のスイッチングを自動化した M.Leake らの研究 [3] や，スタジオ撮影におけるスイッチングを自動化した J.Daemen らの研究 [1] がある．また，自動スイッチングの研究対象は多岐に渡り，ダンスの撮影におけるスイッチングを自動化した土田らの研究 [6] や，遠隔ピアノレッスンにおけるスイッチングを自動化した松井らの研究 [4] などがある．これらの研究は，対象とする映像により手法やモデルが扱う特徴量が異なっている．本研究では、登壇発表を対象とする点で異なる．

登壇発表と類似した映像を扱った研究として，講義を対象とした研究がある．中村らは、音声情報とスライド内のテキストのマッチングにより講義中継のスイッチングを自動化した [7]．また，先山らは，講義をモデル化し，送信映像の選択ルールを作成することでスイッチングを自動化した [5]．本研究の対象である登壇発表においては，スラ

[1] 公立はこだて未来大学
[2] 立命館大学
[3] 東京大学

[*1] http://www.sigmus.jp/?page_id=966

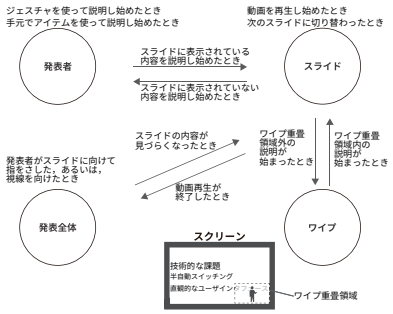

図 1 登壇発表におけるスイッチングの状態遷移モデル

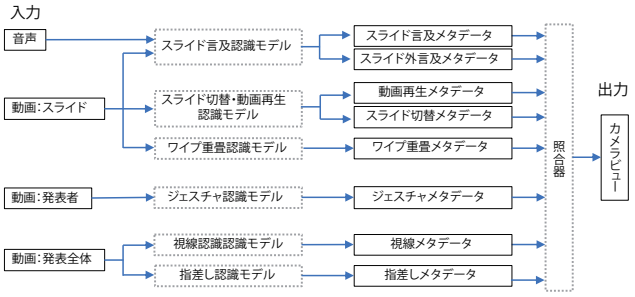

図 2 システムのワークフロー

イド内にキーワードのみを描写し口頭で説明したり，ジェスチャを用いて説明をする場合も多い．そのため発表内容を理解したうえでのスイッチングが必要である．したがって，本研究では音声認識や画像認識を用いて発表内容を考慮してスイッチングを自動化する．

3. 提案システム

本研究では，図 1 に示す Endo らが提案したスイッチング状態遷移モデルに準じて，スイッチングを自動化する．Endo らは登壇発表のライブ中継を対象としたイベントベースのスイッチングインターフェースを提案した [2]．その際，プロの意見に基づいてスイッチングを分析し，スライドの切り替えや，スクリーンへの指差しといったイベントに基づきスイッチングしていることを明らかにした．スイッチングの条件は，遷移前のカメラビューに依存しない条件（図 1 中の円の上部）と，遷移前のカメラビューに依存する条件（図 1 中の矢印付近）の 2 種類があった． 本研究では，スイッチングの条件となるイベントをシステムで認識しスイッチングを自動化する．

3.1 提案システムの概要

提案システムは，登壇発表の動画に対し自動でスイッチングを行うシステムである．スイッチングの条件となるイベントを認識することで，最適なカメラビューを決定する．扱うカメラビューはスライド，ワイプ，発表者，発表全体の 4 つである．これらのカメラビューは Endo らの研究と同様のカメラビューである．図 2 に提案システムのワークフローを示す．入力は，発表者の音声と 3 つのカメラビュー（スライド，発表者，発表全体）の動画である．提案システムは入力された動画に対して，対応するモデルを適用することで，それぞれのスイッチング条件をメタデータとして取得する．その後，それらのメタデータを照合器に入力し，最終的に 1 つの最適なカメラビューを出力する．なお，各イベントの認識モデルについて本研究では，ワイプ重畳モデル，ジェスチャ認識モデル，視線認識モデル，指差し認識モデルについては未実装である．発表経験の少ない，発表者においても頻発するイベント（スライド切替，動画再生，スライドについて話す，スライド外について話す）を認識させるために，本研究では，スライド言及認識モデル，スライド切替・動画再生認識モデル，ワイプ重畳モデルを実装した．

スライド言及・スライド外言及認識モデル

発表者映像とスライド映像間のスイッチングは発表者がスライドの内容について話しているかどうかが条件となる．本論文では，この事象をスライド言及と記述する．また，スライド外の内容について話すことをスライド外言及と記述する．スライド言及・スライド外言及認識モデルは音声と，動画（スライド）を入力とし，各フレームに対し発言なし，スライド言及，スライド外言及の何れかのクラスに分類する 3 値分類モデルである．提案システムでは，スライド言及と認識された出力結果をスライド言及メタデータ，スライド外言及と認識された出力結果をスライド外言及メタデータとして用いる．図 3 に示すように，スライド言及・スライド外言及認識モデルでは発表者の発言内容とスライド内のテキストの一致によりスライド言及とスライド外言及を認識する．具体的な処理について述べる．まず，入力された音声に対し音声認識（google speech to text）を適用する．音声認識により，動画中の音声区間が推定され，音声区間ごとの発言内容がテキストとして出力される．その後，出力されたテキストに対し，単語分割ライブラリ（nagisa）を適用し，それぞれの音声区間に含まれる単語群を取得する．一方，入力された動画（スライド）に対しては，音声区間に対応するスライド（各音声区間の中間の時刻のスライド）を抽出し，それぞれのスライドに対し文字認識（Cloud Vision API）を適用する．その後，文字認識により得られたテキストに対し，単語分割ライブラリ（nagisa）を適用し，それぞれのスライドに含まれる単語群を取得する．最後に，音声認識と文字認識によって得られた単語群に同一単語が 1 つでも含まれている場合，その音声区間はスライド言及であるとする．また，音声認識と文字認識によって得られた単語群に同一単語が 1 つも含まれない場合，その音声区間はスライド外言及であるとする．

3.2 スライド切替・動画再生認識モデル

スライド切替・動画再生認識モデルは，Endo らの状態

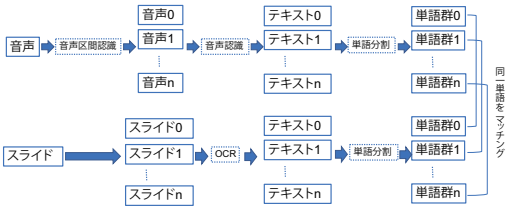

図 3 スライド言及・スライド外言及認識モデルのワークフロー

遷移図における動画を再生し始めたとき，次のスライドに切り替わったとき，動画再生が終了したときの３つのイベントに関わるメタデータを得るために導入した．スライド切替・動画再生認識モデルの入力は動画（スライド）であり，各フレームを変化なし，スライド切替，動画再生の何れかのクラスに分類する３値分類モデルである．提案システムでは，スライド切替と認識された出力結果を動画再生メタデータ，動画再生と認識された出力結果を動画再生メタデータとして用いる．スライド切替・動画再生認識モデルはスライド映像における画素値の変化によりこのイベントを認識する．具体的な処理について述べる．はじめに，入力された動画（スライド）のすべてのフレームに対しメディアンフィルタを適用することでノイズを除去する．次に，すべてのフレームを順次読み込み，1 つ前のフレームとの画素値の差分を取得する．画素地に差分が生じている画素の割合が，閾値を超えている場合にスライド切替とする．また，スライドが連続で変化している場合を動画再生とする．

3.3 照合器

照合器は，認識されたイベントを入力とし最適なカメラビューを決定する．基本的な構造としては，Endo らの状態遷移モデルを条件分岐で表したプログラムである．しかし，スイッチングの条件となるイベントが複数同時に起こった場合の対応と出力の安定化のために，追加のルールを定義した．以下の 3 点が追加のルールである

(1) 現在の状態に依存しない条件を依存する条件よりも優先する．
(2) スライド，全体，発表者，ワイプの順に優先度をつけ，いくつかのイベントが同時発生した際には、この優先度に従いスイッチする．
(3) スイッチングによりカメラビューが切り替わった場合，一定時間はスイッチングを行わない．スイッチングを行わない時間は切り替わった後のカメラビューにより異なり，それぞれ発表者は 6 秒，スライドは 17 秒 ワイプは 27 秒，発表全体は 10 秒である．

(1) のルールは，アイテムの使用や動画の再生など，情報量が多いと考えられるイベントは現在の状態に依存しない条件であるため導入した．(2) のルールは，Endo らの論文におけるプロのスイッチングした動画の平均時間の順序に基づいて決定した．平均時間は，スライド，全体，発表者，ワイプの順に長かった．また，それらの時間に基づいて定義したルールが (3) のルールである．

4. 評価実験

Interaction2019 の 1 件分の発表を撮影した動画を対象とし，前章で提案した各イベントの認識モデルおよび照合器の妥当性を検証した．

4.1 正解データの構築

正解データは Endo らの状態遷移モデルに基づいてスイッチングが行われた動画（以降，正解動画と記述）と各イベントのアノテーションデータ（以降，正解アノテーションと記述）である．正解データは本論文の第一著者と有識者 1 名で相互に確認しつつ構築した．具体的な正解データの構築手順は以下の通りである．

(1) 動画に対し，Endo らの状態遷移モデルにおけるそれぞれのイベントについて手動でアノテーションする．
(2) アノテーションされたデータをもとに Endo らの状態遷移モデルに基づきスイッチングされるよう動画を編集する．
(3) 動画中で感覚的に不自然な箇所や視聴者にとって理解しづらいと感じた部分について調整する．

4.2 照合器の妥当性

提案する照合器が出力するカメラビューの妥当性を検証するために，手動で作成した正解アノテーションを照合器に入力し，照合器が出力したカメラビューと正解動画の一致率を算出した．その結果，Accuracy が 0.41，Precision が 0.41，Recall が 0.41，f 値が 0.41 となった．

f 値が 0.41 だったことから，照合器の出力は正解動画とは異なる部分が多数あることが明らかになった．要因の 1 つとして，スイッチングされたタイミングが正解動画と大きく異なるためである．この問題に対して，発言やスライド切り替えのタイミングを考慮してスイッチングすることで正解動画に近い映像を出力できると考える．

4.3 各イベントの認識モデルの妥当性

各イベントの自動認識モデル（スライド切替・動画再生認識モデル，スライド言及・スライド認識モデル）の妥当性を検証するために，各モデルに対応するカメラビューの動画を入力し，モデルが生成した出力結果と正解アノテーションの一致率を算出した．

各モデルの精度の一覧を表 1 に示す．

スライド切替・動画再生認識モデルについては，Accuracy が 0.88，Precision が 0.88，Recall が 0.88，f 値が 0.88 となった．スライド言及・スライド外言及認識モデルについては，Accuracy が 0.76，Precision が 0.76，Recall が 0.76，f 値が

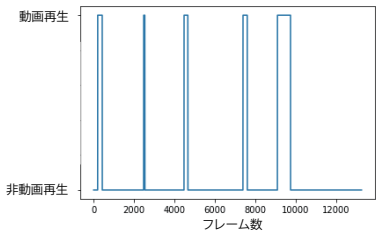

図 4 動画再生の正解アノテーション

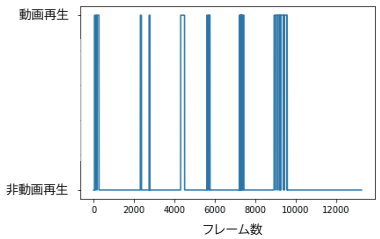

図 5 動画再生認識モデルの出力

0.76 であった．

以下，各モデルについて考察する．

スライド切替・動画再生認識モデル

スライド切替・動画再生認識モデルについては，Accuracy が 0.88，Precision が 0.88，Recall が 0.88，f 値が 0.88 であることから，正解アノテーションとほぼ一致する出力ができているといえる．しかし，図 4 および図 5 に示すように，動画再生の正解アノテーションとスライド切替・動画再生認識モデルで認識した動画再生を比較すると，スライド切替・動画再生認識モデルの出力は断片的に動画を認識していることがわかる．モデルからまとまった出力を得られるための方法として，入力動画のフレームレートを低くすることが考えられる．

スライド言及・スライド外言及認識モデル

スライド言及・スライド外言及認識モデルについては，Accuracy が 0.76，Precision が 0.76，Recall が 0.76，f 値が 0.76 であることから正解アノテーションとほぼ一致する出力ができているといえる．しかし，コンフュージョンマトリックスを分析すると，図 6 に示すようにスライド外言及の認識精度が低いことがわかる．したがって，発言箇所は高い精度で認識できているが，その発言がスライドについて言及しているかどうかを認識するアルゴリズムを改善する必要が有る．具体的には，単語だけではなく，意味的要素を考慮して発言内容とスライド内テキストを比較することで精度向上が期待できると考える．

5. まとめ

本研究では，登壇発表のライブ中継のための自動スイッチングシステムを提案した．提案システムは，音声認識・文字認識・画像処理を用いて登壇発表で生じるイベントを認識し，Endo らのスイッチング状態遷移モデルをもとに

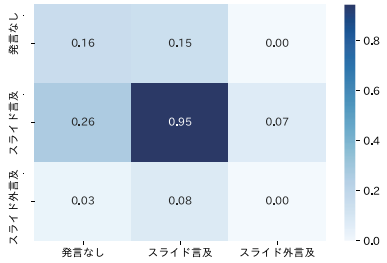

図 6 スライド言及・スライド外言及認識モデルのコンフュージョンマトリックス

構築した照合器をもとに 4 つの映像をスイッチングする．実装した照合器や各イベントの認識モデルの出力結果と正解データを比較することで，認識モデルの妥当性を評価した．その結果，照合器においては，照合器の精度が低く，改善の余地があることがわかった．また，各イベントの認識モデルについては，入力動画のフレームレートや認識方法の改善があることがわかった．

今後は，照合器や各イベントの認識モデルの改善，図 2 で未着手であるジェスチャ認識モデル，視線・指指認識モデルの構築に取り組む．また，評価実験にて，第一著者と有識者との合議にもとづき一意の正解データを構築したが，視聴者が見たいと思う最適なカメラ映像は一意ではないため，正解データの構築方法についても再度検討する必要がある．

謝辞

本研究は JST CREST JPMJCR17A1 の支援を受けたものである．

参考文献

[1] Daemen, J., Herder, J., Koch, C., Ladwig, P., Wiche, R. and Wilgen, K.: Semi-Automatic Camera and Switcher Control for Live Broadcast, *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, TVX '16, New York, NY, USA, Association for Computing Machinery, p. 129–134 (online), DOI: 10.1145/2932206.2933559 (2016).

[2] Endo, S., Takegawa, Y., Funaki, A., Matsumura, K., Hirata, K. and Igarashi, T.: Construction of a Switching Support System for Live Broadcast of Oral Presentation, *Journal of Information Processing*, p. to appear (2021).

[3] Leake, M., Davis, A., Truong, A. and Agrawala, M.: Computational Video Editing for Dialogue-Driven Scenes, *ACM Trans. Graph.*, Vol. 36, No. 4 (online), DOI: 10.1145/3072959.3073653 (2017).

[4] Matsui, R., Takegawa, Y. and Hirata, K.: Remote Piano Lesson System Considering Camera Switching, *Proceedings of International Computer Music Conference*, ICMA, pp. 1–7 (2019).

[5] Sakiyama, T., Ohno, N., Mukunoki, M. and Ikeda, K.: Video Stream Selection According to Lecture Context in Remote Lecture, *The Transactions of the Institute of Electronics,Information and Communication Engineers.*, Vol. 00084, No. 00002, pp. 248–257 (online), available from ⟨https://ci.nii.ac.jp/naid/110003184078/en/⟩ (2001).

[6] Tsuchida, S., Fukayama, S. and Goto, M.: Automatic System

表 1 各モデルの精度の一覧

評価対象	Accuracy	Precision	Recall	f 値
スライド切替・動画再生認識モデル	0.88	0.88	0.88	0.88
スライド言及・スライド外言及認識認識モデル	0.76	0.76	0.76	0.76

for Editing Dance Videos Recorded Using Multiple Cameras, *Advances in Computer Entertainment Technology* (Cheok, A. D., Inami, M. and Romão, T., eds.), Cham, Springer International Publishing, pp. 671–688 (2018).

[7] 中村亮太，井上亮文，　市村哲，岡田謙一，　松下温：誘目性の高い講義コンテンツを作成する自動編集システム，情報処理学会論文誌，Vol. 47, No. 1, pp. 172–180 (2006).