E-007

講義動画のマルチモーダル検索のためのスライド画像の活用の検討
Using Slide Images for Multimodal Spoken Content Retrieval in Lecture Videos

南條 浩輝[1)]　大窄 凜[1)]　小笠原 功二[1)]
Hiroaki NANJO　Rin OSAKO　Koji OGASAWARA

1　はじめに

講義動画アーカイブの検索（音声ドキュメント検索 [1]）の研究を行っている．これまでは主として動画中の音声情報を頼りに動画検索が行われていた．これに対して本研究では，動画中の音声情報に加えて，映像・画像情報も用いたマルチモーダル検索の実現を目指している．今回，講義スライド中の文字や画像の利用法について検討したので，本稿ではそれについて報告する．

2　音声ドキュメント検索とその関連研究

コロナ禍を経てオンライン講義動画は教育現場や自宅学習の場において一般的となった．大学の講義をアーカイブした講義動画では 60 分を超える長さのものも多く，具体的に知りたい箇所を素早く見つけることは，多くの視聴者にとって大きな課題となっている．

動画を検索したり，動画の特定のチャプターを検索したりするためには，ハッシュタグに代表されるようないくつかの検索用のキーワード（索引）をつけることが行われる．音声ドキュメント検索は，音声認識を用いて書き起こしテキストやキーワードを付与し，それに基づいて検索を行うものである．

2.1　音声認識

音声認識には誤りが存在し，認識誤りが音声ドキュメント検索に悪影響を及ぼすことが知られている．そのため音声認識精度を向上させる，誤り訂正を行う，などの研究が行われている．本報告では，音声認識精度の向上自体には主眼は置かない．

2.2　ドキュメント拡張

一般的に情報検索において，検索用のテキスト付与は重要である．ユーザの検索要求に合致する文書であっても，検索要求中の語や言い回しが元の文書に含まれていないと検索が難しい問題がある．そこで，もともとの文書（音声ドキュメント）には含まれないが，検索には有効そうな語やテキストを音声ドキュメントに付与するドキュメント拡張の研究もおこなわれている．本研究では，OCR（光学文字認識）と VLM を利用したイメージキャプショニングを利用して，ドキュメント拡張の検討を行う．

OCR を用いたドキュメント拡張では，OCR 誤り訂正が重要である．本研究では，レイアウト解析に基づく訂正や整形は行わないものの，動画から得られる他のモダリティ情報，具体的には音声情報を用いたマルチモーダル処理に基づく訂正を試みる．

VLM を用いたイメージキャプショニングを利用したドキュメント拡張では，そもそもどのような図表に対して有効な説明文が得られるかが明らかではないため，これを明らかにする．

1)　滋賀大学

3　テストセット

3.1　講義データセット

本研究では滋賀大の e-ラーニング教材を用いる．これらの動画では，スライドを中心に解説が進められている．スライドには文字情報，グラフ，図表，画像などが配置されている．動画内の音声の認識には，「Microsoft Stream」の文字起こし機能を利用する．これによりタイムスタンプと音声認識結果（トランスクリプト）を得ることができ，これを対象に検索したり，チャプタースキップしたりできる．トランスクリプトはアップロード（更新）することもできる．トランスクリプトの訂正やドキュメント拡張することでより高精度な講義動画の検索を実現できると考えられる．

3.2　音声認識と誤り訂正

テストデータとしてランダムに動画 10 本を選びその動画の 5 スライドずつ，合計 50 スライドを使用した．テストデータに対応する正解データは人手で作成した．台本を丁寧に読み上げている音声であるため，単語正解精度（%WAcc）は 99%程度と非常に高かった．LLM で修正する際には，“与えられるテキストが音声認識で得られたテキストであるので，誤りを修正すること．余計なことを付け加えないように” という旨のプロンプトを与えた．「ヒストグラムの…空間を」を「ヒストグラムの…区間を」のように文脈から訂正できた例も見られた．もともとの精度が高かったため，全体としての傾向は捉えられず，今後は精度の低い音声認識結果の訂正を行っていく予定である．

4　OCR を用いたスライド説明文の抽出

4.1　LLM を用いた認識誤り訂正

EasyOCR[2] を用いてスライドから説明文の抽出を行った．この際，スライド内の文字の位置情報（例: テキストボックスの座標など）は考慮せず，自動で取得できる順にデータを接続した．

次に，LLM を用いて OCR の結果を訂正させた．LLM で修正する際には，“与えられるテキストが OCR で得られたテキストであるので，誤りを修正すること．余計なことを付け加えないように” という旨のプロンプトを与えた．

最後に，LLM に音声認識結果と OCR の結果を与え，音声認識結果を参照して OCR の結果を訂正し，出力するようにという指示を与えた．

人手で用意した正解を用いて，OCR 結果，修正結果の WER を算出した．結果を表 1 に示す．GPT3.5 では訂正がうまくいかないが，GPT4o を用いて訂正させることで正しく訂正が進んでいることがわかる．音声認識結果も同時に参照させることでさらに正しい訂正ができることがわかった．

表 1　OCR 結果の訂正

	WAcc
ocr のみ	0.601
GPT4o で訂正 w/o ASR text	0.652
GPT4o で訂正 w/ ASR text	0.674
GPT3.5 で訂正 w/o ASR text	0.586
GPT3.5 で訂正 w/ ASR text	0.511

表 2　OCR 結果と正解との意味的類似度

	Cos_sim
ocr のみ	0.423
GPT4o で訂正 w/o ASR text	0.966
GPT4o で訂正 w/ ASR text	0.914
GPT3.5 で訂正 w/o ASR text	0.777
GPT3.5 で訂正 w/ ASR text	0.467

4.2　OCR の修正結果の意味的類似性

次に OCR 結果と正解の OCR テキストの意味的類似度を測った．本研究では日本語用 Sentence-BERT モデル [3] を使用した．全 token の embedding を平均値 pooling し，それらのコサイン類似度を算出した．

表 2 に OCR の結果と正解テキストとの意味的類似度を示す．GPT4o で訂正することでかなり意味の近い表現になっていることがわかる．音声認識結果を参照させた時は，させなかった時よりも意味の類似度が低下している．音声認識由来の正解にはない単語が混入したことが示唆される．

最後に，音声認識結果と OCR の結果の類似度も調査した．およそ 0.5〜0.6 程度であり，音声認識結果と OCR の結果は意味的に異なることから，ドキュメント拡張として有望であることが示唆された．

5　VLM を用いた画像からの説明文の抽出

次にスライドの OCR では得られない情報，具体的にはイラストからのドキュメント拡張を考える．ここでは VLM を用いてイラストの説明文を生成させることを検討した．これは講義で講師が行っている説明およびスライドに書いてある文字，以外の説明を生成させることで，ドキュメント拡張を目指すものである．

5.1　VLM による図表の分類

はじめにイラストの種類が理解できていないと正しい説明文は生成できないと考え，VLM を用いてイラストの大分類と細分類を行った．

大分類では，グラフ，表，グラフや表を用いた解説図，その他，のカテゴリに分類した．GPT4o による分類結果を表 3 に示す．Acuraccy，Precision, Recall ともおよそ 80%程度であることがわかった．

次に，それぞれで正しく分類されたイラストをさらに細分類した．グラフは，折れ線グラフ，複数のグラフ，曲線グラフ，散布図，ヒストグラム，箱ひげ図，ヒートマップ，棒グラフに細分類させた．Precision 0.90, Recall 0.92 とグラフの中では細分類できている．表は，クロス集計表，ワイドデータ，複数の表，リスト形式，マトリックス表に分類させたところ，Precision 0.15, Recall 0.38 とほとんど分類できないことがわかった．グラフや表を用いた解説図は，グラフを用いた解説図，表を用いた解説図，グラフと表を用いた解説図に分類させた．分類精度は Precision 0.78, Recall 0.96 であった．その他の図は，タイムライン・時系列系統，関係性・相関系統，WEB や UI の画面，補助的図解系統，構造系統，データ可視化系統，プロセス系統に分類させた．Precision 0.71, Recall 0.72 であった．

表 3　VLM（GPT4o）によるイラストの大分類精度

カテゴリ	Precision	Recall	F1-Score	Support
グラフ	0.71	0.80	0.75	40
表	0.73	0.93	0.82	29
グラフや表を用いた解説図	0.65	0.88	0.75	114
その他	0.99	0.72	0.83	190
Accuracy	0.79			
Macro Avg	0.77	0.83	0.79	373

表については細分類できておらず，図表の説明において表の説明は苦手と考えられる．

5.2　VLM による図表の説明

大分類が成功したイラストに対して，実際に説明をさせ，ドキュメント拡張ができるかを検証した．説明文が意味的に正しいかと情報が詳細かの 2 つの観点から主観評価を行った．

意味的に正しいかの観点からは約 75%から 88%の説明文が正しい説明であった．一方，情報が詳細かの観点からは，およそ 10%の文が詳細という結果になった．この結果は，生成された文は「これは○○グラフです」のように，間違ってはいないがほとんど意味をなさないような説明であることを示唆している．今後はプロンプトを工夫し，正しく意味のある説明文の生成を行っていきたい．

6　おわりに

講義の音声ドキュメント検索のために，講義スライド中の文字や画像の利用法を検討した．OCR および VLM を用いてドキュメント拡張の方法を検討した．OCR を行い，結果を音声認識結果を参照させながら LLM で訂正させることで，ドキュメント拡張が行える可能性を示した．VLM による説明文生成では，意味のある説明文を生成することに課題があることがわかった．

謝辞　本研究は科研費（23K11216）の補助を受けて行われた．

参考文献

[1] 秋葉友良. 音声ドキュメント検索: マルチメディアデータを対象とした音声言語情報検索 (＜特集＞『検索』のゆくえ). **情報の科学と技術**, 63(1):21–27, 2013.

[2] Easyocr. https://github.com/JaidedAI/EasyOCR（2025.6.13 アクセス）.

[3] sentence-bert-base-ja-mean-tokens-v2. https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens （2025.6.13 アクセス）.