SwapVid: 資料動画の視聴支援のための 動画・資料ビューアの統合型インタフェース

村上大知[†1] 藤田和之[†1] 原航太郎[†2] 高嶋和毅[†1] 北村喜文[†1]

概要：スライドや論文等の資料を説明する動画（資料動画と呼ぶ）において，ユーザは動画とそれに対応する資料の元データ間の対応関係を把握することが要求され，これには大きな負荷を伴う．本研究では，この動画・資料間の相互の探索を支援するための動画・資料ビューアの統合型ユーザインタフェースである SwapVid を提案する．本インタフェースでは，動画・資料コンテンツの両者を OCR 解析し，各動画フレームと資料位置との対応関係を推定する．これに基づき，ユーザの操作に応じて動画・資料ビューアを切り替え表示することにより，動画・資料を 1 つのウィンドウ内でシームレスに扱うことを実現する．本デモでは，SwapVid インタフェースに加え，Zoom や Web ブラウザ等の既存アプリに SwapVid を適用可能なアプリケーションを紹介する．

図 1　SwapVid 概要

1. 序論

COVID-19 の流行を契機として，スライドや論文等のドキュメントを動画形式で見る機会が急増している．このような動画を本論文では資料動画と呼び，例えば， Zoom 会議における画面共有機能を用いたリアルタイム形式のものや，講義動画等の YouTube や Udemy 等の Web サイトに公開されているオンデマンド形式のものが挙げられる．

一般に，資料動画の視聴においては，動画とは別に，その動画に対応する元データ（主に PDF 形式）が入手可能である場合が多い．この元データは資料動画の理解を助けるものであるが，一方で，資料動画・元データの 2 種類のデータ間の対応関係を把握することは困難である．これは，以下 2 つのタスクのためにユーザにかかる物理的・認知的負荷が大きいからであると考えられる．

- **動画に基づく資料の探索タスク（V2D タスク）**：視聴中の動画コンテンツに関連する情報を資料内で探索するタスク．例：Zoom ミーティングにおいて画面共有されている資料の 1 つ前のページを参照する，講義動画のあるタイミングで言及されたある専門用語が初めて定義されたスライドのページを振り返る，等．
- **資料に基づく動画の探索タスク（D2V タスク）**：閲覧中の資料（元データ）に関連する情報を資料動画で探索するタスク．例：資料で表示しているページについて動画内で言及されている箇所を探す，等．

これまで資料動画の理解や探索支援のための研究は多数なされてきた[1][2][3][4]ものの，上記の資料・動画間の相互の探索タスクの難しさの問題はあまり扱われてこなかった．

そこで本研究では，リアルタイムまたはオンデマンドの資料動画を対象とした，動画ビューアと資料ビューアを統合する新たなユーザインタフェース「SwapVid」を提案する．本インタフェースのコアアイデアは，（1）動画と資料の各コンテンツ間の対応関係をシステムが把握し，それにより，（2）動画と資料を 1 つのビューに統合してそれらをユーザの操作に応じて適応的に切り替え表示することである．より具体的には，本システムは動画の各フレームでの表示されている資料位置（スクロール位置とズームレベル）を，OCR 用いて推定する．これに基づき，動画の直接操作インタフェース[5][6][7]を参考に，あたかも動画として表示されている資料を直接スクロール・ズーム操作するかのような体験を可能とすることで，V2D タスクを支援する．加えて，資料閲覧時には，関連する動画部分をシークバーや資

†1 東北大学電気通信研究所
†2 School of Computing and Information Systems, Singapore Management University

料上に直接ハイライト表示することで，D2V タスクを支援する．本稿では，SwapVid プロトタイプの設計・実装，およびそのアプリケーション例について述べる．

2. SwapVid

SwapVid は，資料動画の閲覧を支援するための資料ビューア･動画ビューアの統合型ユーザインタフェースである．SwapVid のシステムは，動画と資料のコンテンツの対応関係を解析する「シーケンスアナライザ」，および動画ビューア・資料ビューアの「統合型ユーザインタフェース」の 2 つのコンポーネントからなる．以下では， PDF 形式のスライドまたは論文資料に対応し，Web アプリケーションとして幅広いデバイス (PC, タブレット端末, スマートフォン) で動作する SwapVid プロトタイプについて述べる．

2.1 シーケンスアナライザ

動画と資料とのコンテンツマッチング方法として，本システムでは，先行研究[8][9]と同様，OCR を採用した．動画と資料の元データの両方に対し，Tesseract OCR[a]による OCR を適用させた．

図 2 に，シーケンスアナライザの全体的なワークフローを示す．まず，システムは事前に，PDF 資料に含まれる文字列セグメント（文字列及びその位置）に関するインデックスデータを作成する．インデックスデータは，抽出されたテキスト文字列とその位置（ページ番号とページ内のバウンディングボックスの座標）がページ毎に格納されている JSON ファイルである．一方で，システムはキーフレームごとに動画のフレーム画像に対し OCR を実行し，文字列とその位置を抽出する．キーフレームは，監視中の動画でシーン変化が検出されたときに，フレーム間の水平および垂直投影プロファイルの差を計算することによって得られ，本実装ではサンプリングレートを 1Hz とした．以上の方法で，動画と資料の間で抽出された文字列セグメントをマッチングすることで，キーフレームに含まれるコンテンツが資料のどの部分であるかを推定する（ビューポート推定）．このプロセスは，(1) スクロール位置の推定，(2) ズームレベルの推定，の 2 段階で構成される．

2.1.1 スクロール位置の推定

本システム実装においては，動画のキーフレームから抽出された文字列セグメントについて，資料のインデックスデータ内に一致する文字列セグメントが存在するかを総当たり検索でチェックする．OCR の検出の結果が不正確である可能性を考慮し，一致検出は，(1) 文字列長の類似性，(2) N-gram（N＝2）の類似性，および (3) 文字列配列の類似性，の 3 つの類似性指標のそれぞれが，所定のしきい値を超えるかを判定することで行われる．文字列の類似度

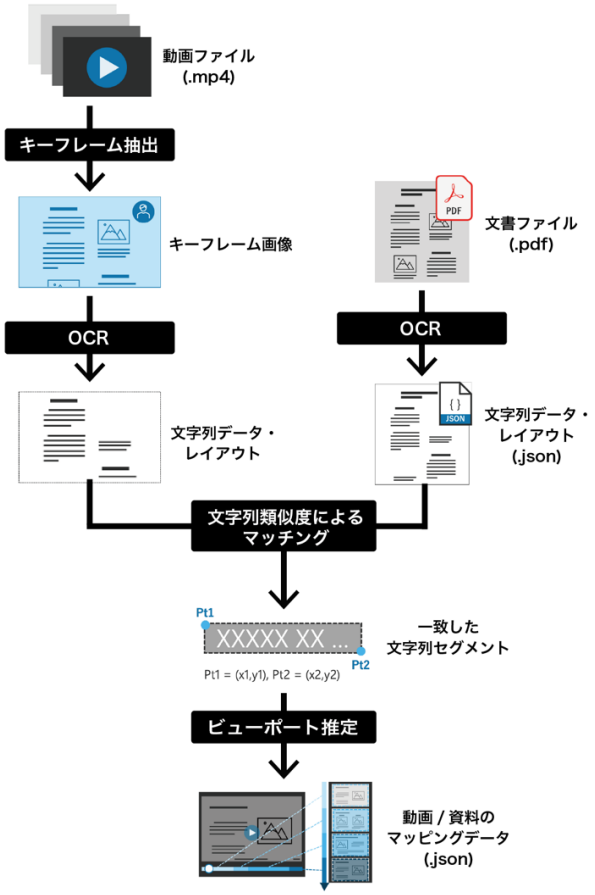

図 2 シーケンスアナライザの処理フロー

判定に基づく都合上，極端に短い文字列はマッチング精度の低下を招くため，事前のフィルタリングによって文字列長が 10 以上のもののみを検索対象とした．

通常，スライド形式の資料と，論文形式の資料ではスクロールの連続性に違いがある（すなわち，前者は動画のページ単位，後者は連続的にスクロールされることが多い）ため，スクロール位置の推定処理は資料の種類によって異なる．資料がスライド形式の場合，ある文字列が動画と資料の間で一致することを検出すると，その文字列を含む資料の元データ中のページ番号を推定候補として返す．キーフレームで画像の先頭からこのプロセスを繰り返した後，システムは最終的に最も一致可能性の高いページ番号を，スクロール位置の推定結果として出力する．資料が論文形式の場合，動画キーフレームと資料内のそれぞれから抽出された文字列セグメントの先頭から 3 行分を比較し，一致を検出した場合，先頭行の位置を推定スクロール位置の候補として返し，この処理を繰り返すことで最終的な推定結果を出力する．

2.1.2 ズームレベルの推定

資料が拡大または縮小されたシーンにおいてもビューポートを正しく推定するため，本プロトタイプでは動画と資

a https://github.com/tesseract-ocr/tesseract

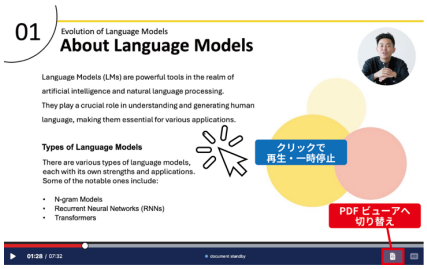

図 3 動画モード

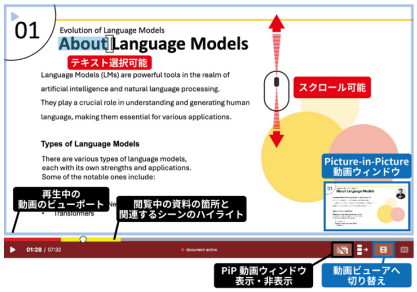

図 4 資料モード

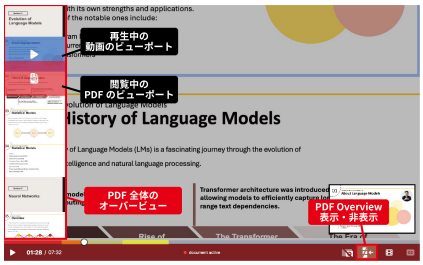

図 5 資料モード（サムネイル表示時）

料の間で一致する特定の文字列セグメントのサイズ差に注目する．具体的には，両データ間でマッチした文字列のバウンディングボックスの高さの比率を計算することで，ズームレベルを推定する．スクロール位置，ズームレベルの推定結果，並びに動画・資料のサイズ情報を用いて，PDF 内のどの領域がキーフレームに映し出されているかを計算するビューポート推定が行われ，動画と資料のマッピングデータが作成される．

2.1.3 ストリーミングモード

シーケンスアナライザは事前処理を前提としているが，Zoom 等のリアルタイム形式の資料動画にも対応する（ストリーミングモード）．ストリーミングモード時，シーケンスアナライザは独立したサーバとして動作する．ユーザが視聴中の動画フレームの画像データを取得し，随時シーケンスアナライザに送信する．この動画フレームの画像データに対する PDF 資料データとのマッチング処理（ビューポート推定）を実行後，その結果を JSON 文字列として返す．なお，高速化のため，ストリーミングモード時にはキーフレーム抽出処理は行わない．

2.2 統合型ユーザインタフェース

図 3-図 5 に，SwapVid ユーザインタフェースの概要を示す．本インタフェースは，動画モードのビュー（図 3）と資料モードのビュー（図 4）を 1 つのウィンドウ内で切り替えることにより動作する．モード間の切り替えは，各ビューアに関連するインタラクションによって自動的に行われる．画面下部のボタンにより手動でも切り替えることができる．

以下では，モードを切り替える際の具体的なインタラクションと動作について説明する．

2.2.1 動画から資料への遷移インタフェース

動画視聴時に，ユーザによる資料に対するインタラクション（スクロールや資料内テキストの選択）を検出した場合，自動的に資料ビューアに切り替える．これにより，ユーザは動画内の資料に対して直接インタラクションしているかのような体験を得ることができる．動画の再生中に資料モードへ遷移した際には，動画はバックグラウンドで再生され続けるため，ユーザは閲覧中の資料とは非同期的に，動画の音声を聞き続けることができる．これにより，ユーザはリアルタイム形式の資料動画（例：Zoom での画面共有）においても，本ユーザインタフェースが使用できる．

2.2.2 資料から動画への遷移インタフェース

資料モード（図 4）では，ユーザによる動画に対するインタラクション（シークバーの操作，および，動画の Picture-in-Picture（PiP）表示のクリック）を検出した場合，動画モードに切り替える．加えて，閲覧中の資料箇所に対応する時間帯を，シークバー上に黄色でハイライトする．これは，シーケンスアナライザによって生成されたマッピングに基づき，現在表示されている PDF 資料上のコンテンツが動画に存在するかを常時監視することにより実現する．ハイライト部分は，ビューポート内に映り込む資料の該当箇所の面積が増加するにつれて，より濃い色で表示されるため，ユーザは資料内における，動画で言及される箇所が視界に入ってきたことを直感的に認識することができる．さらに，PDF ビューの一部に，現在再生中の動画内で注目される領域が，資料ビューア上に青い四角形として表示される．

2.2.3 各ビューアのユーザインタフェースについて

動画ビューアの基本的なインタフェースとして，再生，一時停止，シークバーの制御（シークバーにカーソルを合わせるとプレビューが表示される），字幕表示を実装した．資料ビューアはスクロール，テキスト選択，テキスト検索などの，多くの既存の資料ビューアのような，基本的なインタラクションが可能である．また，図 5 に示すように，画面左側にサムネイルビューを表示する．各ビューアはレスポンシブデザインで設計しており，タッチ入力を伴うモバイルデバイス（i.e.,タブレット端末）上でも使用することができる．

3. アプリケーション SwapVid Desktop

任意の資料動画で SwapVid を使用できるようにするため，既存のアプリ（Zoom や Web ブラウザ等）にオーバーレイして SwapVid インタフェースを適用可能なアプリケーションである SwapVid Desktop を実装した．

SwapVid Desktop では，デスクトップ領域の一部またはアプリケーションウィンドウを画面録画した映像をソースとして読み込むことで，映像に対して SwapVid インタフェー

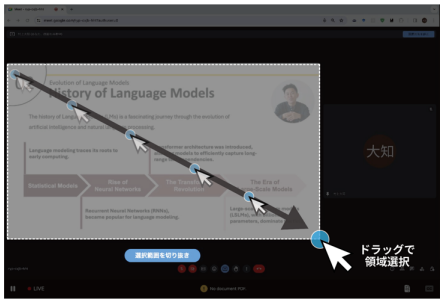

図 6　SwapVid Desktop 動画領域指定

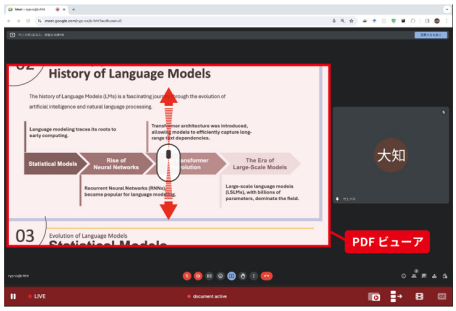

図 7　SwapVid Desktop
Google Meet 画面への資料ビューア表示例

スを適用する．ユーザはまず，資料ソースとなる PDF データ，および，動画ソースとなる任意ウィンドウを選択し，次に選択したウィンドウ内で資料動画が表示されている領域をドラッグ操作で指定する（図 6）．その後，SwapVid システムはストリーミングモードで動作し，資料動画のウィンドウ位置に重畳するように資料ビューアを適応的に表示する．図 7 は，Google Meet の画面共有のビューに，資料ビューアを重畳表示させた例を示している．

4. 結論

本研究では，動画・資料データ間の往来をスムーズにする資料動画の閲覧・探索インタフェースである SwapVid を設計・実装した．また，本インタフェースを既存アプリで利用可能とする SwapVid Desktop アプリケーションを実現した．今後の展望として，画像処理の併用によるより広範な資料形式への対応や，スライド内アニメーションや埋め込みビデオ等の動的コンテンツへの対応等が挙げられる．

参考文献

[1] Li N., Kidzi ´nski Ł. and Dillenbourg P.: Augmenting Collaborative MOOC Video Viewing with Synchronized Textbook, *INTERACT 2015*, pp. 81–88 (2015).
[2] Pavel A., Reed C., Hartmann B. and Agrawala M.: Video Digests: A Browsable, Skimmable Format for Informational Lecture Videos, *ACM UIST 2014*, pp.573–582 (2014).
[3] Wang F., Ngo C.-W. and Pong T.-C.: Synchronization of Lecture Videos and Electronic Slides by Video Text Analysis, *ACM MM 2003*, pp. 315–318 (2003).
[4] Zhao B., Xu S., Lin S., Wang R. and Luo X.: A New Visual Interface for Searching and Navigating Slide-Based Lecture Videos, *IEEE ICME 2019*, pp. 928–933 (2019).
[5] Dragicevic P., Ramos G., Bibliowitcz J., Nowrouzezahrai D., Balakrishnan R.and Singh K.: Video Browsing by Direct Manipulation, *ACM CHI 2008*, pp. 237–246 (2008).
[6] Karrer T., Weiss M., Lee E. and Borchers J.: DRAGON: A Direct Manipulation Interface for Frame-Accurate in-Scene Video Navigation, *ACM CHI 2008*, pp. 247–250 (2008).
[7] Denoue L., Carter S., Cooper M. and Adcock J.: Real-Time Direct Manipulation of Screen-Based Videos, *ACM IUI 2013*, pp. 43–44 (2013).
[8] Chengpei Xu, Ruomei Wang, Shujin Lin, Xiaonan Luo, Baoquan Zhao, Lijie Shao, and Mengqiu Hu. Lecture2note: automatic generation of lecture notes from slide-based educational videos. *IEEE ICME 2019*, pp. 898–903 (2019).
[9] Wang Feng, Ngo Chong-Wah, and Pong Ting-Chuen. Synchronization of lecture videos and electronic slides by video text analysis. In Proceedings of the Eleventh *ACM MM 2003*, pp. 315–318 (2003).