FastPerson: ユーザ中心の学習体験のための視覚・音声情報に基づく講義動画要約

河村 和紀[1,2,a)]　暦本 純一[1,2,b)]

概要：時間が限られている学習者や多様なトピックに関心を持つ学習者にとって，長時間の講義動画を短時間で理解することは学習効率の向上に不可欠である．動画の重要なシーンのみを閲覧することを支援する動画要約は広く研究されているが，これらの研究は主に動画の視覚情報または音声情報のいずれか一方に焦点を当て，動画中の重要なセグメントを抽出する．そのため，講義動画のように教師の発言と黒板やスライドの視覚情報が双方とも重要である場合，重要な情報を見落とすリスクがある．そこで，本研究では講義動画の視覚情報と音声情報の両方を考慮して要約動画を生成する「FastPerson」を提案する．この手法は，動画の音声の書き起こしと動画中の画像やテキストを用いて要約動画を生成し，個々の学習者にとって重要な情報を見逃さないようにする．さらに，動画の各チャプターごとに要約動画と元の動画を切り替える機能を提供し，学習者が興味や理解度に応じてチャプターごとに適切な動画を選択し，学習のペースを柔軟に調整することが可能である．この手法の有効性を検証するために 40 名の実験参加者による評価を行った結果，提案手法を用いることで元の動画を視聴するのと同等の理解度を維持しつつ，視聴時間を 53%削減することが可能であることが確認された．

1. 序論

動画共有プラットフォームの普及により，映画，音楽，ブログ，教育といった多岐にわたる分野での動画の利用が増加しており，教育や学習手法にも影響を及ぼしている．教育分野での動画活用の一例として，反転授業では学生が授業の内容に関連する動画を事前に視聴しておくことで，授業時間はディスカッションや実践的な活動に専念することができる [5]．また，MOOC（Massive Open Online Courses）では，多数の講義や教材が動画形式で提供され，世界各地の学生がオンライン上で自由に学習することが可能となっている [14], [18]．さらに，多くの先行研究において，これらのような動画を活用した学習方法の有効性が確認されている．一例として，Zhang らはインタラクティブな動画の利用が学習者の理解度向上に寄与することを明示している [52]．また，Kay も動画の使用が学生の理解，適応能力，および達成度の向上に寄与することを報告している [22].

これらの研究成果から，動画を中心とした学習方法の利点が確認されている一方で，大量の動画の中から適切な動画を選択するのが困難であることや，個別の理解や興味に合わせた学習ペースの調整が容易ではないという課題がある．実際，ユーザが動画視聴時間の 80%をブラウジングや部分的な視聴に使用していることが報告されている [7]．これは，大量の動画の中から求める内容を効率的に探し出すことが，現在の動画共有プラットフォームにおける主要な課題であることを示唆している．さらに，YouTube の教育チャンネルや Coursera，EdX などのオンライン学習プラットフォームにおいては，30 分から 2 時間の長尺の動画が多く含まれる．しかし，動画は文章コンテンツとは異なり，スキミングのように全体の概要を掴むことや，特定のチャプターごとに閲覧する情報の量を調整することが難しい．この制約は，動画を学習ツールとして利用する際の学習者の障害となり得る．もし動画にこのような柔軟性があれば，学習者は長い動画の重要なポイントだけを素早く確認することで，最初から最後まで動画を見る価値があるのか素早く判断したり，短時間で多くの動画の大要を把握したりすることができる．また，十分理解できている不要な部分は大筋だけをつかみ，重要な部分や理解できない部分は繰り返し，あるいはゆっくり閲覧することができるため，学習効率の向上も期待できる．

これらの課題に対処するため，動画の内容を短時間で把握することを支援する技術として，動画要約が提案されて

[1] 東京大学大学院情報学環・学際情報学府
[2] ソニーコンピュータサイエンス研究所京都研究室
[a)] kwmr@acm.org
[b)] rekimoto@acm.org

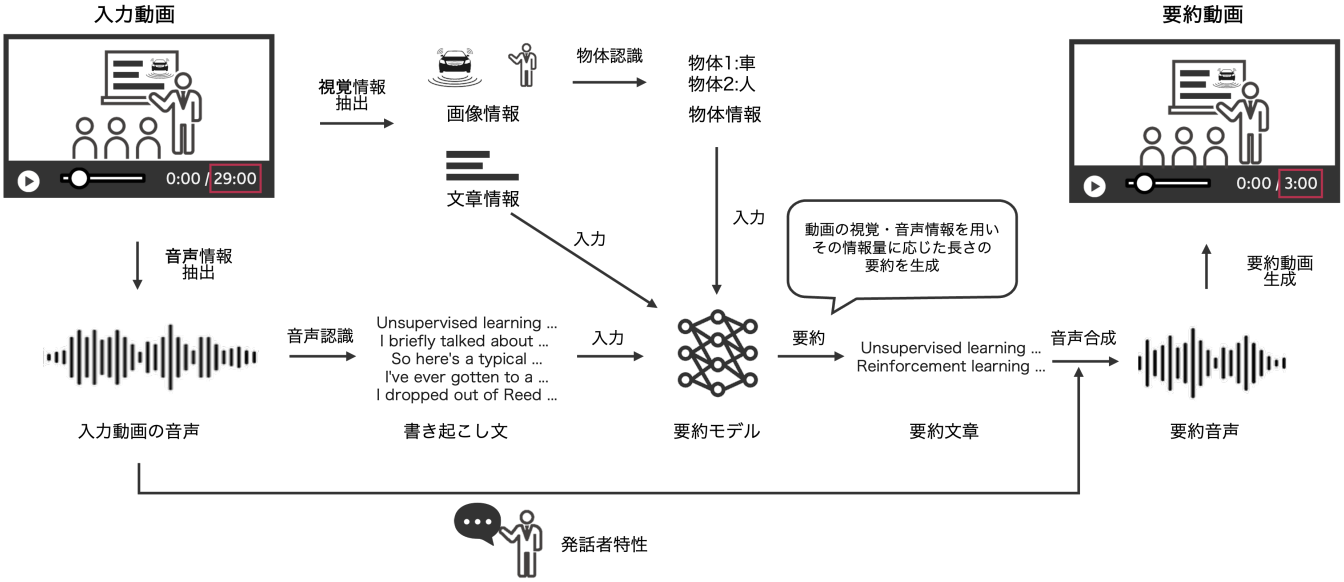

図 1: FastPerson: 視覚情報（画面上の物体や文章）と音声情報（発話者の音声）を用いて要約文章を生成し，その文章を元の発話者の発話特性を用いて音声合成する動画要約

おり，動画から主要な箇所を抽出する方法が様々検討されている [27], [45]．特に，教育コンテンツのように情報が画面への表示と話者の発話の両方に散在する動画を要約する際の課題は，重要な情報が必ずしも映像と音声の同じ箇所に存在するわけではないという点である．すなわち，ユーザにとって重要な要素が映像に含まれる時間帯と音声に含まれる時間帯とが必ずしも被っておらず，どちらかの情報だけに着目して重要な箇所を抽出するとユーザが必要とする要素を取り逃がす可能性があるということである．しかし，既存の動画要約手法は，映像または音声の情報のいずれかに主に焦点を合わせているため，両方の情報を均等に扱う手法はまだ限定的である．そこで，本研究では，視覚情報と音声情報の両方を総合的に考慮する新しい動画要約技術，FastPerson を提案する．通常の動画要約では，重要な箇所を抽出するのに対し，我々の手法では，視覚情報と音声情報を加味した動画の要約文章を音声合成したものと対応する元の映像箇所を組み合わせることで要約動画を生成する．この際，元の動画の発話者の声質を反映させた要約動画を生成することで要約動画と元の動画のシームレスな切り替えを実現している．さらに，要約動画を生成する際の要約率に関しても，視覚情報と音声情報それぞれの情報の量を反映させることで要約率を調整し，学習者にとって重要な内容を維持しながら，情報の過不足を防ぐ．これにより，例えば，話者が長い時間話していたり，映像内で多くの文章情報を含むスライドが表示されていたりするような内容が豊富な部分に関して，より詳細な要約が生成されることとなる．これは，学習者が重要な視覚的または音声情報を逃すリスクを抑えるのに寄与する．

本稿ではさらに，ユーザ中心の学習体験を提供するためのインタラクション方法の設計にも焦点を当てる．提案システムは，動画の各チャプターで，要約動画と元の動画の両方を提供し，ボタンによりそれらを適宜切り替えることを可能にする．これにより，学習者は内容，興味，または理解度に応じて，どちらの動画を視聴するか選択することができる．例えば，要約動画を通じて学習者は迅速に各チャプターの概要を把握でき，更に詳細な情報が必要な場合はそのチャプターの元のフルバージョンの動画を視聴することができる．もちろん，要約動画のみを閲覧することで，動画の概要を把握するという使い方も可能である．また，各動画チャプターのタイトル，要約文，サムネイルが一覧表示され，学習者は関心のあるチャプターだけを直接選択して視聴することも可能である．さらに，提案手法の学習効果とユーザ体験を評価するため，40 名の参加者を対象に評価実験を実施した．二つの実際の講義動画を閲覧後，動画に関する確認テストを実施したところ，FastPerson を使用することにより，視聴時間が 53.24%，53.11%削減され，元の動画を閲覧した場合と同程度の理解が得られることを確認した．また，78%のユーザが要約動画と元の動画の切り替え機能を有用と評価した一方，動画チャプターの遷移や要約動画の音声的な質などいくつかの改善が可能であることが示唆された．

本論文の貢献を以下に示す．

- 講義動画の視覚情報と音声情報を総合的に考慮することで両方の重要な情報を反映させた要約動画を生成する手法「FastPerson」を提案する．
- 元動画と要約動画を自由に切り替える機能により動画の各チャプターで簡潔に情報を取得するか詳細な情報を取得するかの選択を可能にするユーザ中心の学習体

験の設計を提案する．
- 提案手法を動画を用いた学習に適用することにより，学習者が長い動画コンテンツから効率的に情報を習得できることを示す．

2. 関連研究

動画要約は，元の長い動画から短い形式のコンテンツを作成し，ユーザが興味のある主要なコンテンツを素早く把握できるようにする作業であり，主に視覚的に重要な要素に基づく手法，聴覚的に重要な要素に基づく手法，学習者にとって重要な要素に基づく手法の 3 つのカテゴリに分類される．

視覚的要素に基づく要約手法は，視覚情報を利用して重要なフレームやセグメントの識別を目指し，コンピュータビジョンの分野で継続的に研究されている．初期の研究では，顔，物体，美的特徴などの動画の特徴から動画の各セグ面との重要度を教師あり学習手法で予測する [15]．近年の手法では，convolutional neural network (CNN) [12], [24] や recurrent neural network (RNN) [20] といった深層ニューラルネットワークから抽出される高レベルの特徴を組み込み，動画要約の性能向上を図っている [15], [51], [53], [55]．注意機構 [56] や graph convolutional network (GCN) [29]，Transformer [25] も動画要約の性能向上に寄与している．他にも，タイトルに関連するオブジェクトを持つ動画のセグメントを検出する手法もある [43]．これらの手法は，動画監視やスポーツ動画のハイライト生成において有望な結果を示しているが，本研究で対象とするオンライン講義や HowTo 動画などの画面上に文字情報が多く含まれるスライドや板書が表示されることの多い動画への適用は充分に検討されていない．また，料理やスポーツの動画を中心に，視覚特徴に加えて音声特徴を用いて動画の重要なセグメントを予測する手法もいくつかあるが [40], [54]，これらの手法を使用するにはラベル付けされた教師ありデータを多数用意する必要がある．しかし，幅広い分野に広がる講義動画に対しそのようなデータを用意するのは非常に困難であるため，我々は大規模言語モデル（LLM）を活用することで，教師データがなくても講義動画に対して要約された動画を生成できる技術の確立を目指す．

聴覚的要素に基づく要約手法は動画内の音声を書き起こし文へと変換し，この文章に対して文書要約技術 [9], [38] を適用して動画の要約文章を生成するものである．Seq2seq モデルを利用した要約技術 [28] はその一例であり，原文の内容を改変しながら要約を生成する特徴がある．近年の動向としては，ChatGPT [26] のような LLM を利用した動画要約文書を生成する事例が増加している[*1]．これらの手法は言語情報の重要性を十分に考慮しており，講義やプレゼンテーションといった音声が中心となる動画の内容理解に有効である．しかしながら，これらの文書中心の要約手法が提供する情報は，動画としての視覚的要素の重要性を軽視する恐れがある．特に，講義動画においては，音声だけではなくスライドや図表などの視覚情報が重要な役割を果たしている場面が多く，学習者が発話者の要約文だけを読んで動画の内容を理解するのは困難である．我々は，動画内の発話者の発言内容の重要性を認識するとともに，動画の視覚な表現も重要であると考え，音声情報と言語情報の両方を加味し，出力形式が文章ではなく動画として得られる動画要約技術の確立を目指す．他にも，発話内容ではなく，音声の音高に着目し，音高が強い部分を講義動画の重要なセグメントとして検出する手法も存在するが [17]，発話内容と比較すると情報が少なく，一般的なノイズに対してロバスト性が低いため，我々の手法では発話内容を直接要約に用いる．

学習者の嗜好に基づく手法としては，インタラクション履歴やユーザの嗜好情報を考慮した動画要約手法提案が多く見られる [1], [10]．ユーザのインタラクションデータを基盤とした要約生成には，他の学習者が多く閲覧するフレームを取得する Lecturescape [23] や，ソーシャルメディアの視聴者のインタラクション情報を活用してスポーツ動画の主要な瞬間を識別する EpicPlay [44] といった方法が存在する．さらに，ユーザの視聴行動を分析することで，ユーザの嗜好を推測する研究もある [3]．一方で，個々のユーザの関心度に基づく動画要約も開発されている．例えば，Varini らはユーザが直接自分の好みのジャンルを提供することでパーソナライズされた旅行の要約動画を生成する手法を提案しており [47]，EgoScanning ではユーザが関心を示すイベントに関するキーワードを入力することで，該当しない部分の再生速度を高速化することが可能である [19]．さらに，ElasticPlay は要約動画の全体的な長さをユーザがインタラクティブに調整することができる特徴を持つ [21]．我々の手法では，これらの手法のようにユーザの嗜好を直接入力したり，推定したりすることはおこなわない一方で，ユーザ各自で自分の嗜好に合わせて動画の視聴ペースを自由に調整することのできる機能を提供する．

3. FastPerson の動画要約手法

本章では，図 1 に示される FastPerson のアーキテクチャを中心に，その要約手法について述べる．

3.1 動画のチャプター分け

FastPerson では，動画をまとまりのあるチャプターごとに要約し動画要約を生成するため，まず，動画内のシーンの遷移や無音の期間を検出し，各セグメントとして区切る．シーンの遷移検出に関しては，連続するフレーム間の色分布（ヒストグラム）の変動を分析し，変動の大きさに基づきセグメントの境界を決定する [34]．Truong らの研究に

[*1] https://glasp.co/youtube-summary

よれば，連続するフレーム間での明瞭な変動はシーンの変更を示すことが知られており [45]，ヒストグラムの変動は動画のムードや背景の変化を示す可能性が高い．また，視覚的な区切りと音声的な区切りが必ずしも一致しないことを考慮し，音の波形が特定の閾値以下の振幅を持つ期間を無音として検出するアルゴリズムを使用している．振幅の低い期間や特定の周波数の欠如を検出し，無音や静寂のセグメントを同定する [39]．これらの視覚的なシーンの遷移と音声における無音期間の両方を考慮して切のいい場所で動画を区切り，セグメント化する．この機能を用いることで，まとまりのあるセグメントごとにタグをつけずとも自動でチャプター分け（セグメント分割）が可能である．

3.2 視覚・音声情報の抽出

提案手法では，動画の視覚・音声両方の情報を加味した要約動画を生成するため，こららの情報を一度文字情報に変換する．視覚的な情報を取得するためには，光学的文字認識（OCR）および物体検出技術を利用する．OCR [42] を用いることで,，画像や映像中の文字を検出し，これを文書に変換する．例えば講義動画において，スライドの内容やホワイトボードの手書きノートなど，映像内の文書情報を取得するのに有効である．この情報を用いることで，映像内の文書が示すテーマや内容を正確に要約に反映させることができる．一方，物体検出は映像内の各フレームから特定の物体やエンティティを検出・識別する技術である [37]．映像内での物体は，そのセグメントの話の中心やテーマに密接に関連していることが多い．したがって，物体検出を利用することで，映像のキーシーンや重要なエンティティを特定し，その情報をもとにした効果的な要約を生成することが可能となる．我々のシステムでは，OCR に Tesseract-OCR Engine を用い，物体物体モデルとして ResNet-50-FPN をバックボーンとして持つ Faster R-CNN [37] を用いる．

動画における発話内容，特に講義やセミナーなどの学術的な内容は，音声情報の正確な取得も必要である．本手法では，動画の音声を音声認識モデルに入力として供給し，文書としての書き起こしを得る．この段階での精度が要約の品質に影響を及ぼすため，高精度に音声信号を文書データに変換することが望ましい．近年の音声認識の進歩は，RNN [13], [16] や Transformer [48], [50] を基盤とするモデルに起因している．本手法では，その一つである Whisper を音声認識のモデルとして使用する．Whisper は，Web から収集した約 68 万時間の多言語音声で学習されており，その認識精度は人間に匹敵することが確認されている [31]．この手法の性能は高いものの誤認識率はゼロではない，そこで，我々は OCR による映像内の文字の認識により音声認識の誤認識をカバーし，動画内の重要な単語を要約から取りこぼさないよう工夫する．

3.3 要約音声の生成

通常の動画要約では，動画の重要なポイントにラベル付けされたデータを用い，教師あり学習手法で動画内の重要な場所を検出することがおこなわれる．しかし，我々が対象とする教育動画は，多様な領域に広がっており，そのような重要な箇所を抽出する学習データを用いることは困難である．近年，自然言語処理の領域において，BERT [11] や，T5 [35]，GPT [6], [32], [33] のような Transformer アーキテクチャをもとにした LLM が多数提案されている．これらのモデルは様々なドメインの文書や対話を元に学習されているため広範な領域の知識を持ち合わせており，要約生成を含め多様な自然言語処理をおこなうことが可能である [4]．本システムではこの LLM を活用することで，教師データを必要とせずに，講義動画の要約を可能にする．発話の書き起こし文に加えて，OCR や物体検出を介して取得された視覚的なメタデータの両方をもとに動画の要約を生成することで，動画の視覚情報を要約に反映させる．次のように視覚・音声両方の情報を加味した要約を生成するような指示文を LLM [36] に入力し，各動画セグメントの文書要約を生成する：

指示文

Using the provided transcription of spoken content, OCR-derived textual data, and object detection information, synthesize a comprehensive summary. This summary should highlight the key themes and actions depicted in the video. Focus on distilling the essence of the video by combining insights from the transcription (which captures the spoken words), the OCR data (which provides text found within the video), and object detection (which identifies significant objects and actions).

この指示文とともに，各動画のセグメントの書き起こし文と OCR，物体検出された結果の視覚情報が文書として LLM に入力される．この方法を採用することで，要約は発話内容の中には含まれていないニュアンスや，話者の動き，視覚的なオブジェクト，スライドや視覚的なプレゼンテーション内の文書を強調した要約を生成することが可能となる．

3.4 要約長の計算

動画の要約においては，元の動画の情報量を適切に反映させて文書要約の長さを決定することが求められる．つまり，動画が長い場合や内容が豊富な場合には，要約としても十分な情報を提供する必要があり，元の動画で短く済まされている箇所は要約でもその分短くすることで冗長性をなくすことが必要である．そこで，我々は，書き起こされた文書の量に応じて，要約後の出力文字数を調整する．それだけなく，今回の我々手法では，動画の音声だけでなく，視覚情報も元に動画要約を生成するため，要約後の出力文

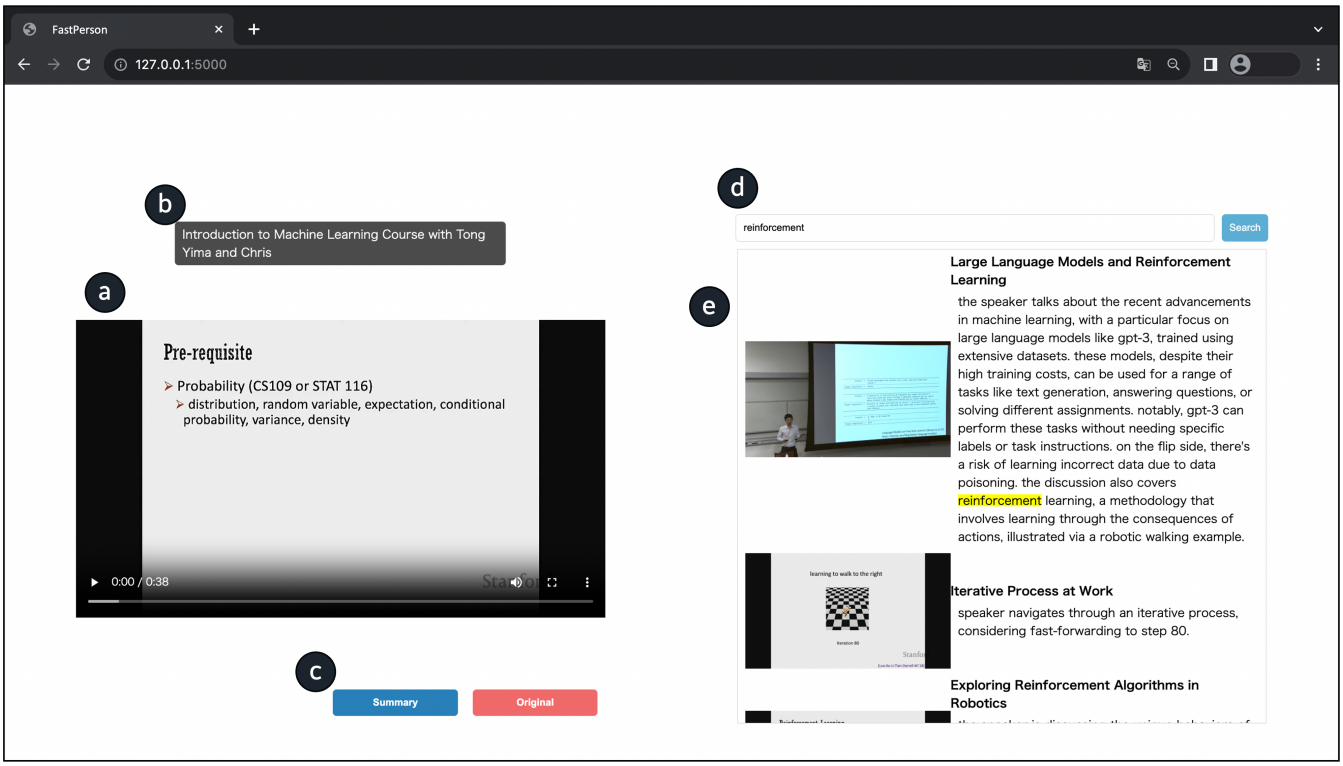

図 2: FastPerson のユーザインタフェース．任意の場所で要約動画と元の動画を切り替えることができるように設計されている．

字数には視覚情報の量も鑑みて調整する．例えば，スライドの文字量が多い場合，話者の発言内容が短いとしても，そのスライドの情報内容を正確に反映して要約することが求められる．この観点から，提案するシステムでは要約の出力文字数 N を以下の式を用いて計算する．

$$N = \max(50, w_s * L_t + w_i * (L_o + L_c))$$

ここで，N は要約の最終的な出力文字数，L_t は書き起こし文書の長さ，L_o は画面上で識別されたオブジェクトの数，L_c は OCR された視覚要素の単語数であり，w_s，w_i は音声情報と視覚情報のどちらを重要視するかを調整するための重み係数である．重み係数 w_s と w_i は，実験データやユーザの好みに基づいて選定される．二つの係数を変更し，視覚的または音声的情報の比重を調整することができる．ユーザの関心や理解度に応じてこの比重を変えることで，個別化された要約動画を提供することもできるが，本研究では固定値としてそれぞれ 0.3 と 2.5 を設定した．

3.5 要約動画の生成

従来のオーディオブックなどの音声要約手法 [49] では，各音声の書き起こしに対して一文単位で重要性を評価し，重要な音声セグメントを直接抽出し結合する手法が採用されていた．しかし，この方法は抽出されたセグメント結合時に音声や動画の連続性に課題をもたらす可能性がある．また，この手法では動画の視覚的な表現を反映させた動画要約を作ることができない．そこで，本手法では，音声合成 [41], [46] を用いた手法を採用している．視覚・音声情報を用いた文書要約を入力として，音声合成技術を用いてその読み上げ音声を生成する．合成音声の生成後，その音声の長さを基準にして，元の動画から最適な映像セグメントを選択する．この映像セグメントは，要約された動画の長さに合わせて，元の動画の最初，中間，または最後から切り出される．簡易的なユーザ実験の結果を鑑み，元の動画の中間に位置するの映像セグメントをデフォルトの結合方法として採用している．この選択された映像セグメントと合成された音声を組み合わせることで，連続性のある簡潔な動画要約を実現する．また，本手法をユーザ使用する際に，頻繁に元の動画と要約された動画を切り替えることが想定されるため，シームレスな体験を提供する必要がある．そこで，音声の合成に際しては，話者適応技術 [2] を利用して，元の音声と要約動画の音声の変化を最小化する方法を採用する．この手法により，元の動画と要約された動画を切り替えた際の音声特徴の不一致を軽減することができ，シームレスな切り替えを実現できる．

4. FastPerson のアプリケーションとインタラクション

4.1 アプリケーション

FastPerson はユーザが自身の興味や理解度に応じて，各動画でセグメントごとに元の長い動画と要約された短い動画を切り替えることができるよう設計されている．さらに，動画の横側に配置された静的コンポーネントでは，各動画セグメントのサムネイルと要約文を表示することでユーザが対応するビデオのチャプターに素早くアクセスできるようにする．

ユーザインタフェース全体は図 2 のようであり，ⓐ **動画ウィンドウ**では，ユーザは，要約された形式もしくは元の形式の動画をこのウィンドウで確認することができる．また，ⓑ **動画タイトル**は動画ウィンドウの上部に配置され，現在再生中のセグメントのタイトルが表示される．ⓒ **Summary ボタンと Original ボタン**により，要約版と元の完全版の動画を迅速に切り替えが可能である．動画の横には，動画のチャプターを表現するためのテキスト要約とサムネイルのペアを表示するウィンドウがある．このようなウィンドウはユーザが興味のある動画のセグメントに素早くアクセスのに有効であることがられている [8], [30]．ⓓ **検索ウィンドウ**では，動画セグメント内の特定のキーワードを検索するための文書入力ボックスおよび検索ボタンが提供される．ⓔ **セグメントウィンドウ**では，各動画セグメントのサムネイルおよびタイトルと要約文が表示される．ユーザは，対応するサムネイルをクリックすることで，該当する動画セグメントへ迅速にアクセスできる．

4.2 ユーザとのインタラクション

FastPerson はユーザが動画をどの程度時間をかけて見るか，また，どこに時間をかけどこに時間をかけないかを調整することを可能にすることで，ユーザ個々の興味や理解度に応じて個人化された学習体験を提供するものである．以下にユーザがこのシステムを用いてどのように動画を閲覧することができるのかの例をいくつか示す．

(1) **要約全体の確認**：ユーザは **Summary ボタン** ⓒ を活用することで，動画の要約版を視聴することができる．この状態で動画を通しで閲覧することで動画の概要を素早く把握することができる．この動作は，一冊の本の大まかな内容を掴む行為と相似性を持っている．

(2) **特定部分の詳細な確認**：**セグメントウィンドウ** ⓔ を使用することで，関連するサムネイルや文書によりセグメントの内容を事前に知ることができる．興味を持ったセグメントを選択すれば，その部分のみを詳しく視聴することができる．深い理解を求める場合，**Original ボタン**（c）を選択し，該当セグメントの完全版を閲覧できる．この機能は，本で特定の節や章をピックアップして読む行為との共通点を持つ．

(3) **特定部分の繰り返し視聴**：ユーザは，理解を深めるために特定のセグメントを何度も再生することができる．これは，読書時に重要と感じた節を再三読み返す状況と似ている．

(4) **キーワードによる動画内の情報検索**：**検索ウィンドウ** ⓓ により，特定の用語やキーワードに関連するセグメントを直ちに検索し，アクセスすることができる．これは，本の索引を参照して情報を見つけ出す行為と同様の体験を提供する．

表 1: 評価実験に用いた動画に関する情報

	動画 1	**動画 2**
タイトル	The Origin of Life	The Market Revolution
サムネイル		
長さ	12:58	21:55
質問 1	What molecule, also called the biological blueprint for life, do all living organisms possess?	The activity shown in the image contributed most directly to which of the following?
質問 2	All living things on Earth share important processes. What are those four processes?	Which of the following developments most directly relates to the overall trend from 1800 to 1840 depicted on the graph?
質問 3	In 1828, a German chemist, Friedrich Wohler, proved that organic life is composed of the same components as inorganic matter. How did he do this?	According to the passage, which of the following best explains the most important effect that technological developments had on American society?
質問 4	What ingredient(s) are seen as necessary for the creation of life?	A significant long-term result of the major pattern depicted on the map was which of the following?
質問 5	What quality of life allows for its slow diversification?	なし

(5) **簡略表示と詳細表示の選択：Summary ボタンや Original ボタン** ⓒ を利用して，ユーザは動画の要約版と完全版を自由に切り替えることができる．この操作は，読書時に速読と詳読を組み合わせることとの類似性がある．速読と詳読を組み合わせることにより，読者は全体の概要を迅速に把握しつつ，必要な部分では深い理解を得ることができる．FastPerson のこの機能は，この読書戦略を動画視聴に応用し，ユーザが関心のあるセグメントに集中しつつ，全体のコンテキストを失わないようにすることを可能にしている．

5. 評価実験

FastPerson システムの学習効果とユーザ体験を評価するために，一連の定量的・定性的評価実験を実施した．

5.1 実験手順

本実験では，参加者を，FastPerson を用いる介入群と一般的な動画プレイヤーを用いる対照群にランダムに割り振り，動画を用いた学習の効果の測定とアンケートを実施する．参加者が視聴する動画はオンライン学習サービスの一つである Khan Academy から二つの動画を選択し，付属する設問を使用して動画の学習効果を測定した．参加者の前提知識によって動画の理解度が大幅に変化しないよう，前提知識を必要とせずにその動画単体で理解ができる難易度の講義動画を選択した．選択した動画は，表 1 に示すように，生物分野と歴史分野から選択され，それぞれ，12 分 58 秒，21 分 55 秒の動画である．動画には付属する設問があるが，一つ目の動画に関しては，動画以外の教材を閲覧しないと正解できない設問が含まれていたため，それらを除

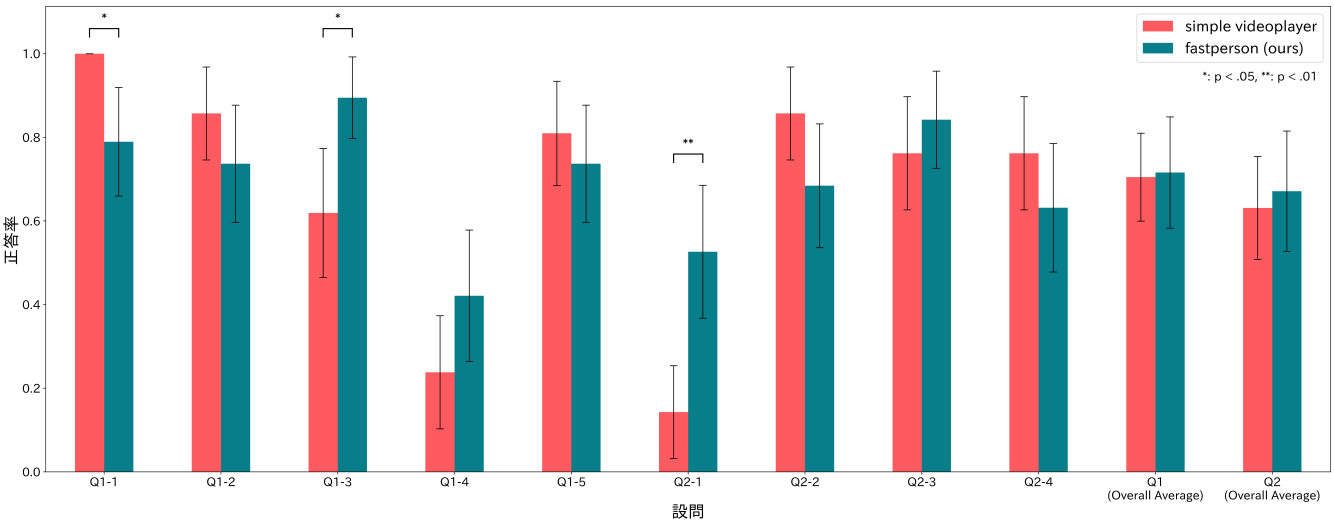

図 3: 動画の設問ごとの正答率

いた 5 問を動画の理解度を確認するために使用した．一つ目の動画は，発話者やビジュアル的な要素がメインの講義動画になっており，二つ目の動画は板書中心の講義動画である．FastPerson を用いる介入郡に関しては，追加でシステムに関するアンケートを収集した．実験はクラウドソーシングプラットフォームの Prolific を通じておこなわれ，作業時間の中央値は 46 分 26 秒で，各参加者には 7£ の報酬が支払われた．

5.2 参加者

本研究では，合計で 40 名の参加者が研究に参加し，各参加者ごとにランダムに FastPerson を使用する介入群か一般的な動画を使用する対照群のいずれかに割り当てられた．結果的に，介入群の参加者が 19 名，対照群の参加者が 21 名となった．参加者の選択基準としては，動画の言語が英語であることから，英語を流暢に話せること，動画の内容を理解できる必要があることから，少なくとも中等教育（secondary education）を終了していることとした．参加者の人口統計学的データ（性別，年齢，教育背景等）は，実験の影響を評価するための追加情報として収集された．参加者の性別は，男性 23 名，女性 17 名で，年齢は平均 26.05 歳（$\sigma = 7.05$）であり，最終学歴は，高校卒業または A レベルが 35%，大学学士号（BA/BSc/その他）が 35%，専門学校・コミュニティカレッジが 15%，大学院学位（MA/MSc/MPhil/その他）が 15%であった．また，得点が上位だった実験参加者にはボーナス報酬が払われることを明記しており，参加者は高い得点を取ることを動機付けられていた．

5.3 学習効果の評価

学習効果の評価のために，一般的な動画プレイヤーを使用する対照群と FastPerson を用いる介入群それぞれに対しての二つの同じ動画を閲覧後の動画に関する設問の正答率と動画の閲覧にかかった時間を分析した．まず，図 3

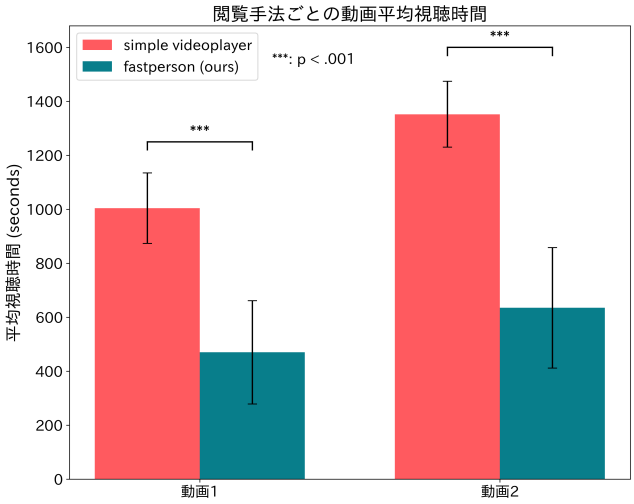

図 4: 閲覧手法ごとの動画平均視聴時間

右端に，動画 1，動画 2 の設問におけるユーザの平均的な正答率を比較したグラフ（それぞれ，「Q1」，「Q2」）を示す．動画 1 は単純なプレイヤーの正答率（赤棒）が，0.70（$\sigma = 0.33$），FastPerson（青棒）が，0.71（$\sigma = 0.42$）であり，動画 2 はそれぞれ，0.63（$\sigma = 0.39$）と 0.67（$\sigma = 0.46$）であることから，単純な動画プレイヤーを使用した場合と，FastPerson を用いた場合で正答率に明らかな差は見られない．実際，二群の間で t 検定を行った結果，一つ目の動画で，t 統計量は 0.170，p 値は 0.864，二つ目の動画で，t 統計量は 0.528，p 値は 0.598 となり，有意水準を 0.05 とするといずれも統計的に有意な差はない．より詳細に各設問ごとの正答率を見ると（図 3 左側），動画 1 の設問 1（Q1-1）は一般的な動画プレイヤーでの視聴が，動画 1 の設問 3（Q1-3）と動画 2 の設問 1（Q2-1）は FastPerson での視聴の得点が高い．Q1-1 は設問が “What molecule, also called the biological blueprint for life, do all living organisms possess?” で，回答が “DNA” であり，FastPerson での要約にもキーワードとしてはピックアップされており，“do all living organisms possess” であるこ

とは示されているが，“biological blueprint for life” と呼ばれているということの説明が不足しているため正答率が下がっているものと考えられる．一方で，Q1-3 と Q2-1 は単純な動画再生でも，FastPerson でも十分に取り上げられているものの，FastPerson では設問の他の間違った選択肢の話題が相対的に少なくなっており，FastPerson での正答率が高くなっていると予測される．動画の視聴にかかった時間は図 4 に結果がまとめられており，通常の動画プレイヤーでは，動画 1，動画 2 それぞれで，ユーザ平均 1003（$\sigma = 312$）秒，1352（$\sigma = 270$）秒かかっているのに対して，FastPerson では，動画 1，動画 2 それぞれで，ユーザ平均 469（$\sigma = 424$）秒，634（$\sigma = 496$）秒で動画の閲覧が終わっている．FastPerson を使用することにより動画 1 の視聴時間は約 53.24%，動画 2 の視聴時間は約 53.11%削減されたことになる．また，二群の間で t 検定を行った結果，一つ目の動画で，t 統計量は 4.409，p 値は $1.063 \cdot 10^{-03}$，二つ目の動画で，t 統計量は 5.281，p 値は $9.955 \cdot 10^{-06}$ となり，いずれも $p < 0.001$ で有意差が認められる．したがって，FastPerson を使用して学習することで，元の動画を閲覧するよりも短い時間で，元の動画と同定度の理解度を得ることができるといえる．

5.4 ユーザ体験の評価

システムのユーザビリティ，学習効果およびユーザ体験を評価するために，FastPerson を用いた介入群の 19 人が動画での学習を終わった後，提案手法に関するアンケート調査をおこなった．今回のアンケート調査では以下の観点で質問をおこなった．

(1) インタフェースの使いやすさ
- Q1: インタフェースの直感性（5 段階スケール）
- Q2: 情報の検索容易性（5 段階スケール）
- Q3: Summary/Original ボタンの使用感（自由記述）
- Q4: UI デザインの改善点（自由記述）

(2) 学習体験の満足度
- Q5: 提案手法を用いた学習の楽しさ（5 段階スケール）
- Q6: 通常の動画プレイヤーとの比較（5 段階スケール）
- Q7: 他の手法と比較した場合の学習体験の評価（自由記述）

(3) 要約動画の有用性
- Q8: 要約動画の有用性（5 段階スケール）
- Q9: 動画の切り替え機能の有用性（5 段階スケール）
- Q10: 要約動画の品質（自由記述）

(4) 特定機能の評価
- Q11: サムネイルや要約文の便利さ（5 段階スケール）
- Q12: 検索ウィンドウの使いやすさ（5 段階スケール）

質問には，数値評価のものと自由記述のものが含まれるが，数値評価の質問に対する結果を図 5 に示す．各質問に対して，5 段階のそれぞれに回答したユーザの割合が棒グラフで表されている．

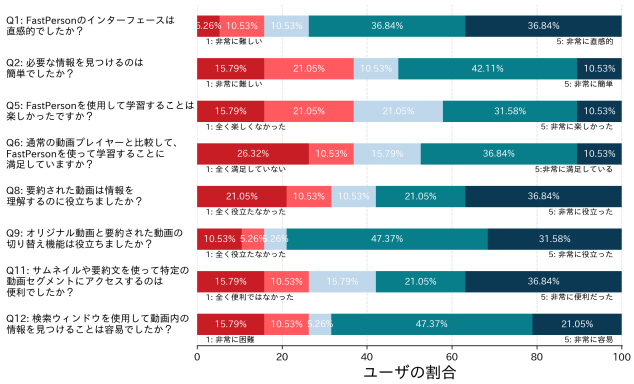

図 5: ユーザ体験に関するアンケート結果

5.4.1 インタフェースの使いやすさの評価

インタフェースの使いやすさ，特に直感性に関する質問（Q1）に対しては，平均評価が 3.89（$\sigma = 1.20$）であった．これは，ユーザが平均的にインタフェースを直感的と感じたことを示している．システム内での情報の見つけやすさに関する質問（Q2）では，平均評価が 3.11（$\sigma = 1.33$）であった．この評価は，ユーザが情報を見つけることにおいて適度な容易さを感じていることを反映している．Summary ボタンと Original ボタンの使用感に関して，多くの参加者は「非常に使いやすく理解しやすい」と評価し，異なる動画モード間の切り替えの有用性と快適さを評価していた（Q3）．しかし，あるユーザは「インタフェースがあちこちにあり，理解するのに多くの時間を費やす必要がある.」と意見し，ボタンや各コンポーネントの配置に改善の余地があることを示唆している．ユーザインタフェースデザインの改善点に関する提案（Q4）では，回答が多岐にわたった．要約された説明文や読み上げ音声の人間味の向上が必要と感じる参加者がいる一方で，サイトの全体的な視覚的側面の向上を求める声もあった．他にも，「動画が断片的になっているため，ユーザが別のパートに移るには，それぞれのパートが終わるのを待たなければならない」という意見があった．実際には，別のパートに移るには動画セグメントのサムネイルをクリックすれば自由に移動することが可能になっていたが，ユーザに伝わっておらず，インタフェースの改善の余地があることが分かる．

5.4.2 学習体験の満足度の評価

まず，FastPerson を使用した学習の楽しさに関する量的測定が行われ，平均評価は 3.00 となった（Q5）．この評価は, ユーザの間で適度な楽しさのレベルがあることを示しているが，1.29 の標準偏差は参加者の体験に幅広い差異が存在することを示唆している．次に，通常の動画プレイヤーとの比較における学習満足度を測定した結果，平均満足度はやや低く 2.95（$\sigma = 1.43$）となった（Q6）．一部のユーザには従来のビデオプレイヤーに対するわずかな嗜好や快適さがあることを強調している．より高い標準偏差は，

ユーザ満足度におけるより広範な差異を示しており，全てのユーザの学習嗜好や期待と完全に一致していない可能性があることを示唆している．オープンクエスチョンを通じて受け取った質的フィードバック（Q7）において，他の動画プレイヤーと比較した場合に肯定的な体験であると評価したユーザは，提案手法により必要な情報を素早く取得することができる特徴やインタラクティブに学習が可能になる性質が楽しさや満足度に寄与する重要な要因であると指摘していた．一方で，満足度が低いと評価したユーザは，インタフェースの使い慣れなさや従来の動画プレイヤーのより単純なアプローチに対する嗜好を理由に挙げており，構成要素をより単純に示すことがユーザの満足度向上に寄与する可能性があることを示唆している．

5.4.3 要約動画の有用性の評価

情報理解における要約動画の有用性に対する平均評価は 3.42（$\sigma = 1.61$）であった（Q8）．この評価は，要約動画が情報を伝える効果に関して，中程度の肯定的な受容を示している．同様に，要約動画と元の動画の形式を切り替える機能に対する平均評価は 3.84（$\sigma = 1.26$）であった（Q9）．ユーザは，詳細と簡潔な動画形式の間を切り替えることができることを価値あるものとしており，学習プロセスにおける自身のコントロールを強化している．要約動画の品質に関するフィードバック（Q10）では，多くのユーザは要約動画のコンセプトを称賛し，特にトピックの復習や時間が限られている場合に，主要なコンセプトを迅速に理解するのに役立つと指摘している．しかし，要約動画の内容の品質に関する懸念も提起された．主な問題点は，内容の深さと要約動画の発音に関連していた．要約は有用であると感じられているが，複雑なトピックを完全に理解するためには，十分な詳細が欠けている場合があるとの意見があった．また，要約動画に含まれる単語の発音が異なっていると指摘するユーザもいくつか存在しており，より高い性能の音声合成を用いることで要約動画の質を向上させることができることが確認された．

5.4.4 特定機能の評価

サムネイルと要約テキストを使用したビデオセグメントへのアクセスの利便性については，平均評価が 3.53（$\sigma = 1.50$）であった（Q11）．同様に，検索ウィンドウを使用したビデオ内の情報の検索容易性は，平均で 3.47（$\sigma = 1.39$）と評価された（Q12）．これらの機能は提案システムに特有なものではないが，要約動画や元の動画との切り替え機能と両立し得ることが分かる．また，サムネイルと要約テキストの評価およびオリジナル動画と要約された動画の切り替えの評価のいずれも過半数のユーザが有用であると評価していることを考えると，動画の要約文だけを提示するのではなく，要約動画および要約文の両方をユーザに提示することが利便性を向上させることが推測される．

6. 結論

本研究では，講義動画の視覚情報と音声情報を総合的に考慮し，学習体験を効率化する新しい動画要約技術「FastPerson」を提案し，その効果を検証した．FastPerson は，動画の視覚的および聴覚的要素を文章化し，それらを統合した要約文を生成することにより，要約動画を作成する．評価実験では，FastPerson を用いた場合と従来の動画再生方法を用いた場合とで，動画に関する理解度に有意な差は見られなかったが，FastPerson は視聴時間を 53%削減することに成功した．この結果は，FastPerson が効率的な学習体験を提供できることを示唆している．さらに，ユーザビリティとインタラクションに関するフィードバックは，インタフェースの直感性や，要約動画とオリジナル動画の切り替え機能の有用性を示している．一方で，要約の内容の深さや音声の質に関する改善の余地も明らかになった．

謝辞 本研究は JST ムーンショット型研究開発事業グラント番号 JPMJMS2012, JST CREST グラント番号 JPMJCR17A3, 国立研究開発法人情報通信研究機構の委託研究 02901 の支援を受けたものである．

参考文献

[1] Agnihotri, L., Kender, J., Dimitrova, N. and Zimmerman, J.: Framework for personalized multimedia summarization, *Proc. SIGMM international workshop on Multimedia information retrieval*, pp. 81–88 (2005).
[2] Arik, S. O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S. and Shoeybi, M.: Deep Voice: Real-Time Neural Text-to-Speech, *Proc. ICML*, p. 195–204 (2017).
[3] Babaguchi, N., Ohara, K. and Ogura, T.: Learning personal preference from viewer's operations for browsing and its application to baseball video retrieval and summarization, *IEEE transactions on multimedia*, Vol. 9, No. 5, pp. 1016–1025 (2007).
[4] Basyal, L. and Sanghvi, M.: Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models, *arXiv preprint* (2023).
[5] Betihavas, V., Bridgman, H., Kornhaber, R. and Cross, M.: The evidence for 'flipping out': A systematic review of the flipped classroom in nursing education, *Nurse Education Today*, Vol. 38, pp. 15–21 (2016).
[6] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al.: Language models are few-shot learners, *Proc. NeurIPS*, Vol. 33, pp. 1877–1901 (2020).
[7] Chen, L., Zhou, Y. and Chiu, D. M.: Video Browsing - A Study of User Behavior in Online VoD Services, *Proc. ICCCN*, pp. 1–7 (2013).
[8] Chi, P.-Y., Ahn, S., Ren, A., Dontcheva, M., Li, W. and Hartmann, B.: MixT: automatic generation of step-by-step mixed media tutorials, *Proc. UIST*, pp. 93–102 (2012).
[9] Chopra, S., Auli, M. and Rush, A. M.: Abstractive sentence summarization with attentive recurrent neural networks, *Proc. NAACL*, pp. 93–98 (2016).
[10] del Molino, A. G., Boix, X., Lim, J.-H. and Tan, A.-H.: Active video summarization: Customized summaries via on-line interaction with the user, *Proc. AAAI*, Vol. 31, No. 1 (2017).
[11] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint* (2018).

[12] Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, *Biological cybernetics*, Vol. 36, No. 4, pp. 193–202 (1980).
[13] Graves, A., Mohamed, A.-r. and Hinton, G.: Speech recognition with deep recurrent neural networks, *Proc. ICASSP*, pp. 6645–6649 (2013).
[14] Guo, P. J., Kim, J. and Rubin, R.: How Video Production Affects Student Engagement: An Empirical Study of MOOC Videos, *Proc. L@s*, p. 41–50 (2014).
[15] Gygli, M., Grabner, H., Riemenschneider, H. and Van Gool, L.: Creating summaries from user videos, *Proc. ECCV*, pp. 505–520 (2014).
[16] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A. et al.: Deep speech: Scaling up end-to-end speech recognition, *arXiv preprint* (2014).
[17] He, L., Sanocki, E., Gupta, A. and Grudin, J.: Auto-summarization of audio-video presentations, *Proc. MM*, pp. 489–498 (1999).
[18] Hew, K. F. and Cheung, W. S.: Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges, *Educational Research Review*, Vol. 12, pp. 45–58 (2014).
[19] Higuchi, K., Yonetani, R. and Sato, Y.: Egoscanning: Quickly scanning first-person videos with egocentric elastic timelines, *Proc. CHI*, pp. 6536–6546 (2017).
[20] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, *Neural computation*, Vol. 9, No. 8, pp. 1735–1780 (1997).
[21] Jin, H., Song, Y. and Yatani, K.: Elasticplay: Interactive video summarization with dynamic time budgets, *Proc. MM*, pp. 1164–1172 (2017).
[22] Kay, R. H.: Review: Exploring the Use of Video Podcasts in Education: A Comprehensive Review of the Literature, *CHB*, Vol. 28, No. 3, pp. 820–831 (2012).
[23] Kim, J., Guo, P. J., Cai, C. J., Li, S.-W., Gajos, K. Z. and Miller, R. C.: Data-driven interaction techniques for improving navigation of educational videos, *Proc. UIST*, pp. 563–572 (2014).
[24] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. D.: Backpropagation applied to handwritten zip code recognition, *Neural computation*, Vol. 1, No. 4, pp. 541–551 (1989).
[25] Liu, Y.-T., Li, Y.-J. and Wang, Y.-C. F.: Transforming multi-concept attention into video summarization, *Proc. ACCV* (2020).
[26] Lund, B. D. and Wang, T.: Chatting about ChatGPT: how may AI and GPT impact academia and libraries?, *Library Hi Tech News*, Vol. 40, No. 3, pp. 26–29 (2023).
[27] Money, A. G. and Agius, H.: Video summarisation: A conceptual framework and survey of the state of the art, *J. Visual Communication and Image Representation*, Vol. 19, No. 2, pp. 121–143 (2008).
[28] Palaskar, S., Libovickỳ, J., Gella, S. and Metze, F.: Multimodal abstractive summarization for how2 videos, *arXiv preprint* (2019).
[29] Park, J., Lee, J., Kim, I.-J. and Sohn, K.: Sumgraph: Video summarization via recursive graph modeling, *Proc. ECCV*, pp. 647–663 (2020).
[30] Pavel, A., Reed, C., Hartmann, B. and Agrawala, M.: Video digests: a browsable, skimmable format for informational lecture videos., *Proc. UIST*, Vol. 10, pp. 2642918–2647400 (2014).
[31] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I.: Robust speech recognition via large-scale weak supervision, *Proc. ICML*, pp. 28492–28518 (2023).
[32] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. et al.: Improving language understanding by generative pre-training (2018).
[33] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al.: Language models are unsupervised multitask learners, *OpenAI blog*, Vol. 1, No. 8, p. 9 (2019).
[34] Radwan, N. I., Salem, N. M. and El Adawy, M. I.: Histogram Correlation for Video Scene Change Detection, *Proc. ICCSEA*, pp. 765–773 (2012).
[35] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P. J.: Exploring the limits of transfer learning with a unified text-to-text transformer, *J. Machine Learning Research*, Vol. 21, No. 1, pp. 5485–5551 (2020).
[36] Ray, P. P.: ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, *Internet of Things and Cyber-Physical Systems* (2023).
[37] Ren, S., He, K., Girshick, R. and Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *Proc. NeurIPS*, Vol. 28 (2015).
[38] Rush, A. M., Chopra, S. and Weston, J.: A neural attention model for abstractive sentence summarization, *arXiv preprint* (2015).
[39] Scheirer, E. and Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator, *Proc. ICASSP*, Vol. 2, pp. 1331–1334 (1997).
[40] Shang, X., Yuan, Z., Wang, A. and Wang, C.: Multimodal video summarization via time-aware transformers, *Proc. MM*, pp. 1756–1765 (2021).
[41] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y. and Wu, Y.: Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions, *Proc. ICASSP*, pp. 4779–4783 (2018).
[42] Smith, R.: An Overview of the Tesseract OCR Engine, *Proc. ICDAR*, Vol. 2, pp. 629–633 (2007).
[43] Song, Y., Vallmitjana, J., Stent, A. and Jaimes, A.: Tvsum: Summarizing web videos using titles, *Proc. CVPR*, pp. 5179–5187 (2015).
[44] Tang, A. and Boring, S.: EpicPlay: Crowd-Sourcing Sports Video Highlights, *Proc. CHI*, p. 1569–1572 (2012).
[45] Truong, B. T. and Venkatesh, S.: Video Abstraction: A Systematic Review and Classification, *Trans. Multimedia Comput. Commun. Appl.*, Vol. 3, No. 1, p. 3–es (2007).
[46] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio, *Proc. SSW*, p. 125 (2016).
[47] Varini, P., Serra, G. and Cucchiara, R.: Egocentric video summarization of cultural tour based on user preferences, *Proc. MM*, pp. 931–934 (2015).
[48] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *Proc. NeurIPS*, Vol. 30 (2017).
[49] Wang, B., Jin, Z. and Mysore, G.: Record Once, Post Everywhere: Automatic Shortening of Audio Stories for Social Media, *Proc. UIST*, pp. 1–11 (2022).
[50] Wang, C., Tang, Y., Ma, X., Wu, A., Okhonko, D. and Pino, J.: Fairseq S2T: Fast Speech-to-Text Modeling with Fairseq, *Proc. AACL/IJCNLP*, pp. 33–39 (2020).
[51] Yao, T., Mei, T. and Rui, Y.: Highlight detection with pairwise deep ranking for first-person video summarization, *Proc. CVPR*, pp. 982–990 (2016).
[52] Zhang, D., Zhou, L., Briggs, R. O. and Nunamaker, J. F.: Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness, *Information & Management*, Vol. 43, No. 1, pp. 15–27 (2006).
[53] Zhang, K., Chao, W.-L., Sha, F. and Grauman, K.: Video summarization with long short-term memory, *Proc. ECCV*, pp. 766–782 (2016).
[54] Zhao, B., Gong, M. and Li, X.: Audiovisual video summarization, *IEEE Transactions on Neural Networks and Learning Systems* (2021).
[55] Zhao, B., Li, X. and Lu, X.: Hierarchical recurrent neural network for video summarization, *Proc. MM*, pp. 863–871 (2017).
[56] Zhou, K., Qiao, Y. and Xiang, T.: Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward, *Proc. AAAI*, Vol. 32, No. 1 (2018).