ユーザによって定義される意味空間を用いた能動的動画視聴インタフェース

岡井 尚也[1,a)]　宮下 芳明[1]

概要：本稿では，動画をセグメントに分割し疑似的に意味空間上に展開することで，時間ではなく意味に基づいた能動的な動画視聴を行えるインタフェースを提案する．動画は等間隔で分割され，マルチモーダルモデルによって 2 つの自然言語軸を持つ意味空間にプロットされる．軸は柔軟かつ容易に指定可能である．例えば長時間のライブ配信動画であっても，意味空間に展開すれば各ユーザがその嗜好や移り行く興味に合わせて視聴できる．

1. はじめに

アニメやスポーツ，ドラマなどの映像コンテンツにおいて，印象的な名シーンを繰り返し見返すという視聴行動は，昔から多くの視聴者にとって自然な楽しみ方の一つであった．感動的なセリフや演出，象徴的な構図などは，時間が経ってもなお人々の記憶に残り，視聴者はそれらの場面に個人的な意味を見出しながら，再び再生したり，そこからコンテンツ自体に興味を持ったりしてきた．

一方，近年 YouTube などの動画投稿サイトにおいては，ゲーム実況やライブ配信といった長尺の動画コンテンツが著しく増加している．これらの動画は多様な展開や情報を内包しているため，視聴者が自身の関心に合ったシーンを見つけ出すことは容易ではない．また，スポーツ中継のような長時間の試合動画においても，注目する選手やプレー，試合の転換点などは個々の視聴者によって異なり，画一的な視聴体験では満足させることが難しくなっている．

このような背景から，第三者や運営が長尺動画から特定の場面を抜き出して再編集した「切り抜き動画」という文化が広く普及した．例えば，人気 VTuber の数時間に及ぶライブ配信から面白いトーク部分だけをまとめた動画が挙げられる．アニメの名シーンやスポーツのハイライトも切り抜き動画の一種であるといえる．切り抜き動画は，視聴時間の短縮や要点の把握に有効である一方で，その編集は制作者の恣意的な判断に委ねられるという課題を抱えている．これにより，視聴者が本当に見たい場面が必ずしも含まれているとは限らず，個々の多様な関心に応えきれていない．また，元動画の文脈から切り離されることで，意図していなかった面白い場面との偶然の出会い（セレンディピティ）の機会も失われがちである．加えて，このような編集プロセスは，文脈を無視した悪質な編集による誤情報の拡散や，元となる動画の制作者へ収益が適切に還元されないといった問題を引き起こすこともある．

本稿では，ユーザに主導権のない視聴体験を「受動的視聴」と位置づける．切り抜き動画を選ぶ行為は一見能動的に見えるが，一度再生が始まれば，それは編集者が構築した一本の筋道をたどる受動的な体験に過ぎない．もちろん，映画やプレゼンテーションのように，作り手の意図に沿って物語を追う受動的な体験が有効に働く場面も多い．しかし，近年のようにコンテンツが多様化・長尺化する中，従来の受動的モデルだけでは応えきれない視聴者の多様な興味に対し，より主体的で能動的な関わり方を可能にするインタフェースが有効であると考えられる．

そこで本研究では，こうした課題を解決するため，元動画の価値を損なうことなく，視聴者一人ひとりの興味関心に基づいて内容を探索できる「能動的視聴」を実現するインタフェースを提案する．本システムは，動画を時間でセグメントに分割し，それらをマルチモーダルモデルを用いてユーザーが自然言語で指定した二つの軸（例：「登場人物」「感情」）に従って二次元の疑似的な意味空間上にプロットする．視聴者は，この意味空間を俯瞰し，興味を引かれた点をクリックすることで，動画の該当箇所へ瞬時にジャンプすることができる．これにより，視聴者は時間軸による制約から解放され，より主体的な動画体験を得ることが可能となる．本稿では，システムの設計と処理構成を示した上で，システム適用例を通じて，本手法の有効性を考察する．

[1] 明治大学
[a)] ev230514@meiji.ac.jp

図 1　システムのメイン画面 (動画は SPOTV NOW より [1])

2. 関連研究

2.1 シークバーの拡張

YouTube などで用いられる一般的なシークバーを拡張した研究が存在する．本稿第二著者らは，シークバーをロープに見立て，曲げる・切るといったインタラクションを行うシステムを提案した [2], [3]．ロープの配置や結合によって編集や可視化を行い，自由で直感的なコンテンツ操作を実現した．高嶋らは，動画内容に応じて再生速度を変化させるシステムを提案した [4]．サッカーの試合といった概観が分かればいい箇所と細かに見たい箇所がある動画に対し有効性を示している．

2.2 意味に基づく動画インタフェース

動画を意味的特徴に基づいて可視化・操作する試みは，初期には低レベル特長 (色・音・動きなど) に基づくクラスタリングを用いた代表フレーム抽出から始まった．Campanella らは，動画に含まれる構成要素を視覚的に分類し，ユーザによる意味的なアノテーションの補助を目的としたインタフェースを提案した [5]．

CLIP に代表されるマルチモーダルモデルの登場により，動画フレームを意味空間に埋め込み，類似性に基づいて整理，探索するインタフェースが提案されている．Ablaoui らは，動画フレームを意味空間にマッピングし，視覚的に探索できる UI を提案している [6]．Lin らは，動画の各フレームを複数の特徴の潜在空間 (色，形状，意味など) にプロットし，異なるレンズを切り替えながら動画全体を探索するシステムを提案した [7]．

本研究は，これらの研究とは異なり，ユーザが自然言語を用いて任意の視聴観点（意味軸）を定義できる点に新規性がある．これにより，元動画の価値を損なうことなく，個々の興味関心に応じた探索が可能となる．本手法は，視聴インタフェースの変更のみで，従来の時間軸に沿った受動的な視聴を，ユーザ主体の能動的な視聴体験へと変えるものである．

3. 提案システム

3.1 システム概要

本稿では，マルチモーダルモデルを用いて動画を解析し，従来のシークバー型視聴ではなく，動画セグメントを疑似的な意味空間上に展開した二次元動画視聴インタフェースを提案する．意味空間はユーザが自然言語で入力したカテゴリや尺度によって構築され，各セグメントがどこかに割り当てられる．本システムにより，ユーザは時間軸に沿った受動的視聴ではなく，意味的な関心に基づいた能動的視聴を行うことが可能となる．システムのメイン画面を図 1，全体処理の概要を図 2 に示す．

3.2 意味空間へのプロット

意味空間へのプロット処理は，ユーザがカテゴリまたは数値で指定可能な二つの意味軸とその内容を自然言語で入力し，1 セグメントの秒数を決定後 (図 1④)，「分析開始」

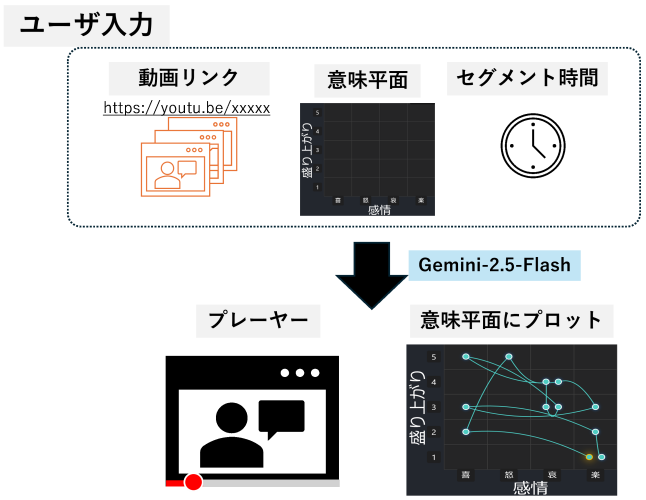

図 2　システムの処理概要

ボタン (図 1⑤) を押すことで行われる．入力された意味軸とセグメント秒数に基づき，各動画セグメントがモデルによって解析される．動画は複数同時に選択でき，それぞれ並列で処理され，同一の意味空間にプロットされる仕様となっている．

解析では，モデルにより各セグメントにユーザが入力した意味軸の要素が割り振られる．この結果をもとに，二次元平面上にセグメントがプロットされる (図 1③)．リンクを入力した動画は画面に表示される (図 1①)．プロットされたセグメントをクリックすることで，動画の該当セグメントにジャンプし，即座に再生を開始できる．同一のカテゴリに複数の点がプロットされる場合は，円を描くように配置される．時系列が連続するセグメントは，プロット点に被らないように曲線で結ばれる．

3.3　セグメントの解説

動画下に再生中のセグメントの解説が表示される (図 1②)．解説は，モデルによって出力された評価結果と簡単な概要である．モデルがハルシネーションを起こしていないか，評価は適切か等を確認する役割を持つ．

3.4　実装環境

フロントエンドは JavaScript で実装されており，インタラクティブな操作やプロット描画を担っている．YouTube の再生は Iframe Player API を用いて行っている．バックエンドは Node.js 上に構築し，動画のアップロード処理，セグメント生成，およびモデルへのリクエスト処理を行う．セグメントの評価には Gemini 2.5 Flash[8] を使用した．同モデルは，YouTube リンク及びタイムスタンプに基づく動画処理が可能であり，動画ファイルを直接扱うことなくシステムを実現している．

図 3　アニメ作品へのシステム適用例
(動画は名探偵コナン公式より [9], [10])

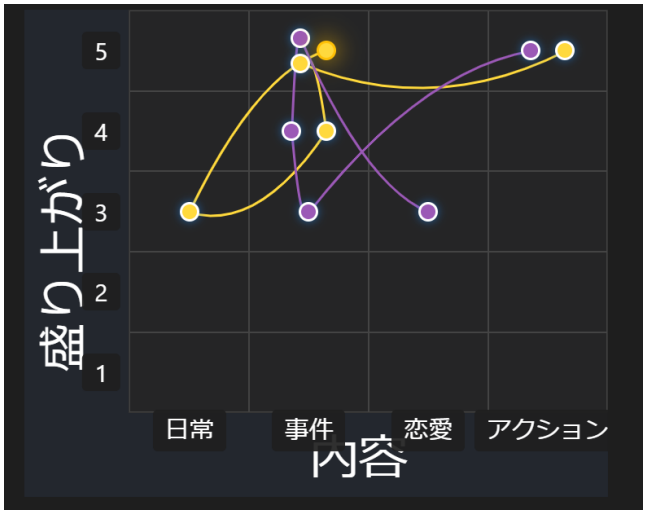

図 4　図 3 右上部の拡大

4.　システム適用例

本システムの使用例として，アニメ作品，スポーツ (野球) 動画，VTuber の配信を紹介する．注意点として，本システムは長時間の動画に適応することを目標としているが，モデルの制約から編集が既に入っている短い動画を使用している．

4.1　アニメ作品

アニメ作品において，視聴者が名シーンを繰り返し視聴する行為は一般的であり，制作会社からは公式に「名場面集」などが発信されることもある．しかしながら，これらの編集済み映像は多様な視聴者の嗜好を必ずしも網羅するものではない．たとえば，アクションシーンを重視する視聴者と，ストーリーの展開や登場人物間の関係性に関心を持つ視聴者では，名場面と捉える基準が異なる．

本システムは，こうした多様な嗜好に即してシーンを選択できる点で有効である．図 3，4 は，YouTube にて公開されているアニメ『名探偵コナン』第 1 話，第 3 話に本システムを適用した例である．同作品は，日常・推理・恋愛・

図 5　スポーツ動画へのシステムの適用例
(システム画面は図 1 と同一)

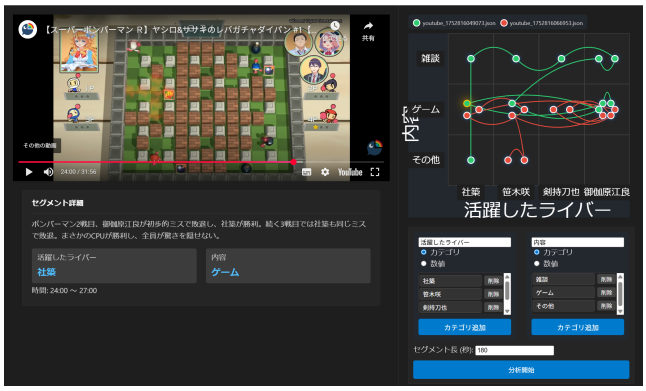

図 7　VTuber の動画へのシステムの適用例
(動画はにじさんじより [11], [12])

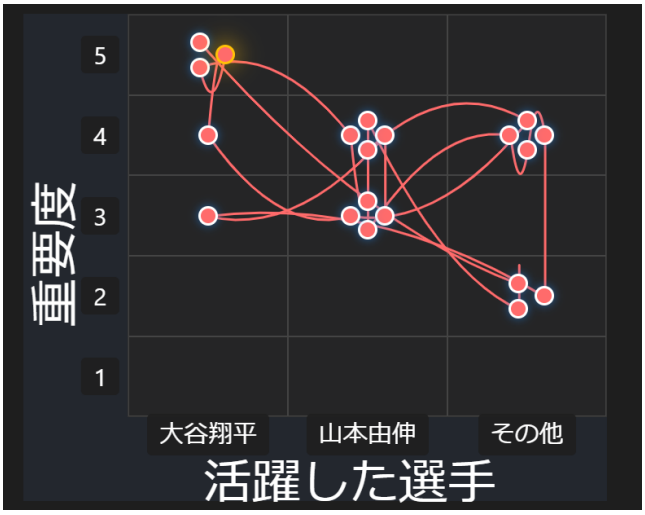

図 6　図 5 右上部の拡大

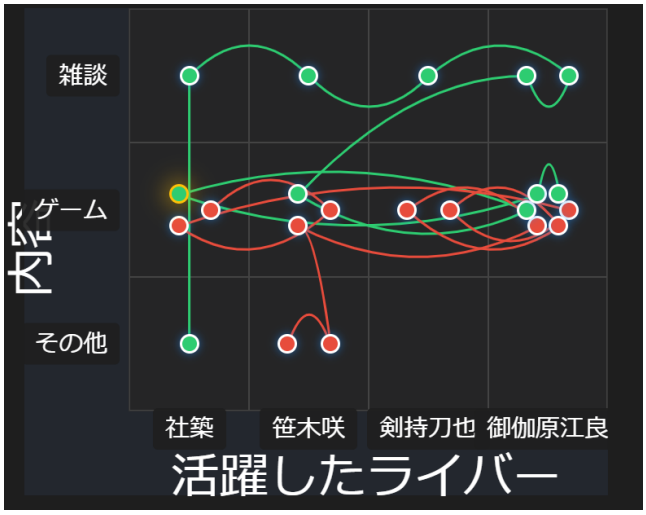

図 8　図 7 右上部の拡大

アクションといった複数の要素を内包しており，幅広い層のファンを持つ．すべてのファンに向けた名シーンの網羅的提示は困難であるが，本システムを活用することで，「事件に関連する場面のみを再視聴したい」「日常的なやり取りに注目して振り返りたい」といった個別の関心に基づいた視聴が可能となる．

4.2　スポーツ動画

スポーツにおける切り抜き動画は，劇的な逆転や名場面を収めたコンテンツとして，数年にわたり繰り返し視聴されることも多い．一方で，視聴者が「名場面」と感じる基準は一様ではなく，多様である．あるユーザは特定の選手のプレーに注目したいと考える一方で，別のユーザは試合展開の転換点や解説者によるコメントなどに関心を持つ場合もある．このような多様な関心に対応した柔軟な切り抜きは，従来の一括編集型のハイライト動画では実現が難しい．提案システムは，任意の選手名やプレー内容といった自然言語による意味軸に基づいてセグメントを意味空間上にプロットできるため，ユーザの関心に即したパーソナライズドな動画視聴体験を提供する．

図 5，図 6 は，YouTube の SPOTV NOW チャンネルに掲載されている「2025 年 7 月 2 日 ロサンゼルス・ドジャース vs シカゴ・ホワイトソックス」のハイライト動画 [1] に本システムを適用したものである．このハイライト動画に本システムを適用することにより，選手ごとに活躍を別で見る，試合の山場を把握する，試合を主役となった選手を確認するといったことができる．本システムを試合のフル動画に適応できれば，よりニッチな場面に注目して動画を閲覧できる可能性がある．

4.3　VTuber の配信

YouTube などの動画 SNS においては，VTuber によるゲーム実況を中心としたライブ配信コンテンツが高い人気を集めており，それに付随して視聴者や第三者による「切り抜き動画」も多数制作・共有されている．これらの切り抜き動画は，配信の特定場面を短時間で把握できる利便性がある一方で，編集者の恣意的な判断に基づくため，視聴者の多様な関心に応じた柔軟な視聴体験を提供することは難しい．例えばゲーム実況など，プレイ内容と語りが同時に進行する配信では，視聴者が注目する場面が人物ごと・

視点ごとに異なる傾向があり，従来のハイライト構成では対応しきれない．

図 7，図 8 は，YouTube のにじさんじ公式チャンネルに掲載されている「ヤシロ＆ササキのレバガチャダイパン」の第 1 回，第 2 回に本システムを適用したものである．動画は，ライブ配信ではなく約 30 分ほどのテレビ番組風に構成されており，MC の「社築」「笹木咲」がゲストと共に雑談やゲームを行っている．本システムにより，「推し」の活躍に絞って動画を視聴したり，複数の動画の中からゲームパートに絞って視聴できる．本システムを，より時間の長くなる傾向のあるライブ配信に適応できれば，長いアーカイブの中から自分の興味関心のある部分の抽出や新たな「推し」の発見につながる可能性がある．

5. 考察

5.1 意味軸指定の柔軟性と課題

本システムの特長は，ユーザが任意の自然言語によって意味軸を指定できる点にある．この柔軟性により，従来手法と比して個別の関心に即した探索が可能となる．実験では，「感情（喜怒哀楽）」や「登場人物名」などの軸に基づいたセグメント分類が概ね意味に整合して行われることを確認した．これは，マルチモーダルモデルを用いた意味空間の構築が，視聴者の多様な関心を反映した情報抽出に有効であることを示唆する．

一方で，軸の指定にはある程度の事前知識が必要であり，動画全体の構造を把握していないユーザにとっては入力のハードルが高い．また，指定された意味軸がすべてのセグメントに適用可能とは限らず，プロットされない区間が生じる可能性もある．この問題に対しては，今後，動画の特徴量に基づいて意味軸候補を自動抽出・提案する機能の導入が有効であると考えられる．これにより，ユーザの入力負担を軽減しつつ，探索支援の精度と利便性を向上させることが期待される．

5.2 セグメント分割の精度と視聴体験

本システムにおけるセグメントの分割は，原則として等間隔の時間単位で行っている．しかし，実際の視聴体験においては，意味的まとまりに基づく区切り（場面転換や感情の変化など）が重要であり，不適切な分割は意味的連続性を損ね，探索効率や理解度の低下を招く．たとえば，感情の転換点や台詞の起承転結の途中でセグメントが切断されると，プロット上の意味位置が曖昧になり，視聴者の意図にそぐわない導線が生じる恐れがある．

さらに，分割の単位時間そのものもプロット結果に影響を与える．本システムでは任意の自然数で分割長を指定可能としたが，極端に短い時間を設定した場合，一部のセグメントにおいて隣接セグメントの内容が誤って評価に影響を与える例が確認された．一方で，短時間分割による高密度な遷移の可視化は，従来の長尺セグメントとは異なる視聴価値を提供し得る．今後は，モデルの時間分解能と視覚的負荷のバランスを考慮しながら，最適な分割単位を設計していく必要がある．

5.3 モデル依存性とシステム制約

本システムのセグメント評価処理は，Gemini-2.5-Flash に依存している．本モデルは高精度かつ高汎用な性能を持つが，45 分を超える長尺動画には対応しておらず，実運用においては処理対象の制約が生じる．また，YouTube 再生には IFrame Player API を用いており，再生が許可されていない動画は埋め込み視聴が不可能である点も運用上の制限となる．

加えて，分類および要約処理においてはモデルの出力結果に依存しており，誤分類やハルシネーションといった問題も確認されている．特に動画時間が長くなるにつれて，出力の曖昧さが増加する傾向が見られた．現時点では，各セグメントに対して簡潔な要約と分類結果を提示することで，ユーザ自身が意味的妥当性を確認できる設計としているが，今後はユーザからのフィードバックによる精度改善などが求められる．

6. おわりに

本稿では，マルチモーダルモデルを用いて動画を疑似的な意味空間上に展開し，ユーザが自然言語で指定した意味軸に基づいて能動的に視聴できるインタフェースを提案した．これは，編集者の主観に依存し，元の文脈から切り離されがちであった従来の「切り抜き動画」に対する一つのアンチテーゼである．本システムにより，視聴者は時間軸の制約から解放され，自身の興味関心に基づいた直感的で主体的な探索が可能となった．

本システムは，多様なコンテンツの視聴体験を個人に最適化する可能性を秘めている．例えばアニメ作品では，アクションシーンを重視する視聴者と登場人物の関係性に注目する視聴者が，それぞれの基準で名場面を再発見できる．スポーツ動画では，特定の選手の活躍だけに注目したり，試合の転換点のみを追いかけたりといった，画一的なハイライトでは満たせなかった多様な関心に応える．さらに，数時間に及ぶこともある VTuber のライブ配信においても，自身の「推し」の活躍場面や，複数の配信から特定のゲームパートだけを横断して視聴することが可能となり，新たな魅力の発見にも繋がるだろう．

今後は，ユーザスタディによる評価や意味軸の自動提案機能，プロンプト設計の最適化などを通じ，本研究で提示した「第三者の編集に頼らず，視聴者自身の興味を道しるべとして時間軸から解放された探索」という新たな視聴体験を，さらに洗練させていく予定である．

参考文献

[1] SPOTVNOW:【日本人コンビが躍動！大谷 5 年連続 30 号山本 7 回 1 失点 8 奪三振で 8 勝目！】ホワイトソックス vs ドジャース 試合ハイライト MLB2025 シーズン 7.2, http://www.youtube.com/watch?v=y1pPGYU1omw (2025). Accessed: 2025-07-18.

[2] 佐藤剛 and 宮下芳明: Seek Rope: 曲げて切って結べるシークバー, インタラクション 2010 論文集, pp. 197–200 (2010).

[3] 青木惇季 and 宮下芳明: SeekRopes: 複数スライダとシークロープによる音楽制作, インタラクション 2011 論文集, pp. 429–432 (2011).

[4] 高嶋章雄: 時間を利用した視覚的表現における情報理解のためのインタラクション, *The Japanese Society for Artificial Intelligence*, Vol. 20, No. 1, pp. 114–221 (2005).

[5] Campanella, M., Leonardi, R. and Migliorati, P.: Interactive visualization of video content and associated description for semantic annotation, *Signal, image and video processing*, Vol. 3, pp. 183–196 (2009).

[6] Ablaoui, L., Marcilio-Jr, W. E., Ng, L. X., Jouffrais, C. and Hurter, C.: Interactive Content Retrieval in Egocentric Videos Based on Vague Semantic Queries, *Multimodal Technologies and Interaction*, Vol. 9, No. 7 (online), DOI: 10.3390/mti9070066 (2025).

[7] Lin, D. C.-E., Caba Heilbron, F., Lee, J.-Y., Wang, O. and Martelaro, N.: VideoMap: Supporting Video Exploration, Brainstorming, and Prototyping in the Latent Space, *Proceedings of the 16th Conference on Creativity & Cognition*, CC '24, New York, NY, USA, Association for Computing Machinery, p. 311–327 (online), DOI: 10.1145/3635636.3656192 (2024).

[8] Google: Gemini 2.5 Flash, https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash (2025). Accessed:2025-07-18.

[9] 【アニメ】名探偵コナン公式: 【公式】名探偵コナン「ジェットコースター殺人事件」| シーズン 1 第 1 話, https://youtu.be/RtJCQk-dSnQ (2020). Accessed: 2025-07-18.

[10] 【アニメ】名探偵コナン公式: 【公式】名探偵コナン「アイドル密室殺人事件」| シーズン 1 第 3 話, https://youtu.be/y4dRWQKJrYs (2020). Accessed: 2025-07-18.

[11] にじさんじ: 【スーパーボンバーマン R】ヤシロササキのレバガチャダイパン 1【にじさんじ】, https://youtu.be/QVOgoELCCqs (2020). Accessed: 2025-07-18.

[12] にじさんじ: 【スーパーボンバーマン R】ヤシロササキのレバガチャダイパン 2【にじさんじ】, https://youtu.be/dO7egHj1O3U (2020). Accessed: 2025-07-18.