文字起こしに基づく動画の切抜き部分推定

小城 凱[1,a)]　伊東 栄典[2]

概要：近年，YouTube 等の動画サイトでは膨大な数の動画が毎日投稿されている。長い動画を短く見るための要約動画や切り抜き動画も増加している。我々は長時間動画から，ハイライト部分を機械的に切り抜く手法について検討している。機械学習に用いる学習用データとして，人手による切抜き動画と，その元動画の比較を想定している。今回，YouTube の人気動画と，その切り抜き動画を対象に，切り抜かれた部分が元動画のどの部分かを推定するツールを試作した。切抜き部分の推定には，動画の文字起こしテキストを用いた。本発表では，切抜き部分推定手法と，試作したシステム，および推定精度を発表する。

キーワード：切り抜き動画, 機械学習, 推定, 文字起こし

Clipped video parts estimation based on transcription

Kai Kojo[1,a)]　Eisuke Ito[2]

Abstract: In recent years, a huge number of videos have been posted every day on video sharing sites such as YouTube. Summarised videos and clipped videos become populare, and they allow longer videos to be viewed in shorter formats We are interested in some methods to realize video summarization and video clipping methods using machine learning. We assume that a manually cut video and its original video are better for machine learning as training data. For the future machine learning using massive amount of video, we develop a tool that estimates clipped video parts in its original video using transcription texts. In this paper, we describe the method of clipped video parts estimation based on transcription, a prototype tool, and estimation results.

Keywords: clipped video, machine learning, estimation, transcription

1. はじめに

近年，YouTube を始めとする動画サイトが人気である。これらのサイトには毎日膨大な数の動画が投稿されている。ライブ配信も人気で，視聴者の多い配信者は定期的にライブ動画を配信している。ライブ配信動画は，ライブ（同時刻）視聴だけでなく，YouTube の機能を活かしての後から視聴も多い。

動画数や視聴環境の変化より人々の視聴傾向も変化している。スマートホンが普及した 2010 年代から，短時間動画の需要が高まっている。稲田は著書 [1] の中で，時間効率向上のため映画を早送りで見る人が増えたと示している。時間短縮需要に応えて，TikTok やショート動画が流行し，2010 年代には長時間動画の一部を短く切抜き編集した動画（以降，切抜き動画）が出現した。その後，元動画と切抜き動画の両方が利益を得る仕組みができたことで，さらに切抜き動画が増加した。

切抜き動画は，視聴者が長時間動画全てを見ずに，面白いシーンだけを視聴できるため人気を得ている。そのためゲーム実況や VTuber 等のライブ配信の面白い場面を切抜く動画が増えている。比較的長時間のライブ配信動画を見るのが面倒で，盛り上がり部分や要約のみ見たいという要望の存在を示す。

従来より動画のハイライト部分抽出は，映画やスポーツの動画に対して研究されてきた。我々は機械学習による動画のハイライト部分および要約部分の抽出を目指してい

[1] 九州大学大学院システム情報科学府
[2] 九州大学情報基盤研究開発センター
[a)] kojo.kai.183@s.kyushu-u.ac.jp

る [2]。ハイライト抽出のための機械学習データとして切抜き動画を考えている。元動画のハイライト部分を人力で特定したものが切抜き動画と仮定することで，機械学習で元動画からの切抜き動画部分推定が可能になる。切り抜き部分の自動推定器が実現できれば，同種の動画に対するハイライト部分自動抽出が可能になる。

本論文では，Vtuber の実況動画を対象に，切り抜かれる前のライブ配信動画と，切り抜かれた切り抜き動画を比較し，切り抜かれた部分が元動画のどの部分に相当するのかを推定する手法を提案する。また試作した推定ツールを紹介し，いくつかの動画に適用した結果を報告する。

本論文の構成は以下の通りである。第 2 節で関連研究を述べる。第 3 節で文字起こしによる切り抜き部分の推定手法と試作ツールを説明する。第 4 節で実験と結果を述べる。最後に第 5 で，まとめと今後の課題を述べる。

2. 関連研究

動画のハイライト部分抽出について，動画内容を対象とするコンテンツベースの手法が提案されている。

Liang らは文献 [3] で，教師なし学習による動画要約手法 CAAN (Convolusional attensive adversarial network) を提案している。動画要約では従来，教師あり学習を用いた手法が成果を上げてきた。学習用の動画要約データセットでは，人間が人手で動画を視聴し，どの部分を要約箇所として選択するかを選び，選択箇所についてのアノテーションデータを作成していた。学習用の動画要約データを多数作成するのは困難かつ費用がかかる。そこで Liang らは，教師なし学習での動画要約を試みている。その手法として CAAN を提案している。CAAN では，深層学習に基づく敵対的生成ネットワーク (GAN) を活用した，教師なし学習で動画要約を生成する枠組みである。

この枠組みを，生成器と識別器の 2 つから構成している。生成器では動画の全フレームに対する重要度スコアを予測する。予測には，全体的な表現を構築する完全畳み込みシーケンスネットワークと，正規化された重要度スコアを予測するためのセルフアテンションネットワークを使用している。これにより，動画フレームのグローバルおよびローカルな時間関係を捉えている。識別器では，生成器で重み付けされたフレームの特徴と，元のフレーム特徴を区別している. 動画要約データセットである SumMe と TVSum に提案手法を適用し，動画要約結果を他の手法と比較している。その結果，他の教師なし学習による動画要約手法と比べて，最も高い精度で要約部分の抽出を実現している。

Zhao らは文献 [4] で，動画要約および要約部分のトピック推定を行う手法を提案している。そのために Multimodal transformer model を用いている。このモデルでは，動画から抽出された複数のモーダル (視覚的・聴覚的・言語的な特徴) を適応的に融合し，トピックに関する要約動画を生成する。生成は以下のような手順で行われる。

- 特徴抽出：ビデオから視覚的，聴覚的，言語的特徴を抽出する。視覚特徴は CLIP モデル，テキスト特徴は BART モデル，音声特徴は PANNs モデルを用いて抽出する。
- 特徴学習：抽出された複数のモーダル特徴を統合し，時間的情報をモデル化する。Multimodal Transformer Encoder を用いて特徴を融合し，時間 Modeling Encoder で時間情報を捉える。
- トピック分類：更新された視覚特徴を用いて，ビデオフレームをトピッククラスに分類する。複数トピックであることを考慮し，マルチラベル分類タスクとして取組む。
- フレーム選択：更新された融合特徴を用いて，フレームレベルの重要度スコアを予測し，要約を生成する.

3. 切抜き部分の推定手法

機械学習で，長時間動画からハイライト部分を切り抜くためには，切抜き動画で切り抜かれた部分が，元動画のどの部分であるか知る必要がある。元動画の切り抜き部分を推定するために，動画の文字起こしテキストを利用する。

切り抜き動画および元動画は，YouTube に投稿された動画を対象とする。YouTube には膨大な動画が蓄積されており，配信等の長時間動画も多い。また切り抜き動画も数多く投稿されている。動画チャンネル 1 つに対し，複数の切り抜き動画作成者がいる場合もある。切り抜きかれた動画群のうち，どの動画が最も人気があるのかを知るには，動画再生数を見れば良い。切り抜き動画の品質は，動画再生数で推定できる。

3.1 文字起こしテキストの構造

YouTube Data API v3 [6] を用いると，YouTube 動画のメタデータに加え，視聴者コメントや字幕（文字起こし）テキストが入手できる。投稿チャンネルの動画リストも取得できる。字幕テキストには，YouTube の自動文字起こしで生成されるものと，動画投稿者が埋め込むものがある。本論文の対象は自動文字起こしされたものである。

字幕データには，表示テキストと，表示時刻間隔が含まれている。切り抜き動画の文字起こしテキスト（字幕）と，その元動画の文字起こしテキストを入手し，2 つを比べる。元動画の文字起こしデータを A，切り抜き動画の文字起こしデータを B とする。A と B は以下の構造になる。

$$A = (a_1, a_2, \cdots, a_n),$$

$$B = (b_1, b_2, \cdots, b_m).$$

ここで $a_1, a_2, ..., a_n$ が動画 A の文字起こしテキストであ

る。実際のデータには，テキストと共に時刻情報が記述されている。動画再生時に字幕としてテキスト a_i の表示される期間として，再生開始からの経過時間が 2 つ記述されている。

3.2 文字起こしテキストによる切り抜き箇所推定

一般に元動画より切り抜き動画が短い。そこで，切り抜き動画 B のテキスト $b \in b_1, b_2, ..., b_m$ を，元動画 A の文字おこしテキスト $a_1, a_2, ..., a_n$ と比較する。

YouTube の自動文字起こしの精度は高くない。日本語の場合，かな漢字変換の誤りも多い。そのため切り抜き動画 B が元動画 A の一部を切り抜いたものだとしても，B の文字列 b が，A の要素に含まれるとは限らない。そこで文字列の類似度を計算することにした。類似度は Python 言語の `diff.SequenceMathcer` を用いて計算した。この類似度は以下の計算式で算出され，0〜1 の値を取る。1 のとき完全一致である。

$$sim(a,b) = \frac{2 \times length(LCS(a,b))}{lengtha + length(b)}$$

上記の式にある $LCS(a,b)$ は，文字列 a と b の最長共通部分文字列 (longest common subsequence) である。分子で LCS を 2 倍しているのは，元の文字列に対する共通部分列の寄与を強調するためである。これに共通部分列の重要度が考慮され，文字列 a, b の長さに依存せずに類似度を計算できる。

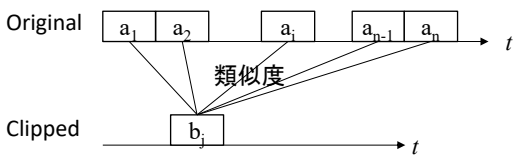

図 1 切り抜き部分の推定方法

類似度計算により，B の b_j が，A の a_i と類似度が高いことが判明したとする。このとき，切り抜き動画 B における b_j の表示期間が，元動画 A の a_i の表示期間を切り抜いたと推定する。

図 1 に，切り抜き部分推定手法を示す。これらの処理を行うプログラムを Python 言語で作成した。切り抜き部分の候補は，CSV ファイルとして出力する。

3.3 推定部分チェック HTML

前節の推定手法で抽出した切り抜き部分が，本当に正しいかを調査するには，人間が 2 つの動画を見るしかない。そこで，2 つの YouTube 動画を表示させるための HTML, JavaScript プログラムを作成した。

CSV ファイルに出力された切り抜き部分の候補情報を用いて，HTML を生成する。HTML の上部には，元動画 A と切り抜き動画 B を差し込む。その下に，切り抜き候補をリストとして並べる。HTML,JavaScript の記述で，YouTube 動画の再生開始時間を指定する。また推定の合否を「None, ◯, ×, △」で評価し，それを保存する仕組みも作成した。これにより機械的に推定した切り抜き部分チェックが省力化出来た。図 2 にチェック画面を示す。

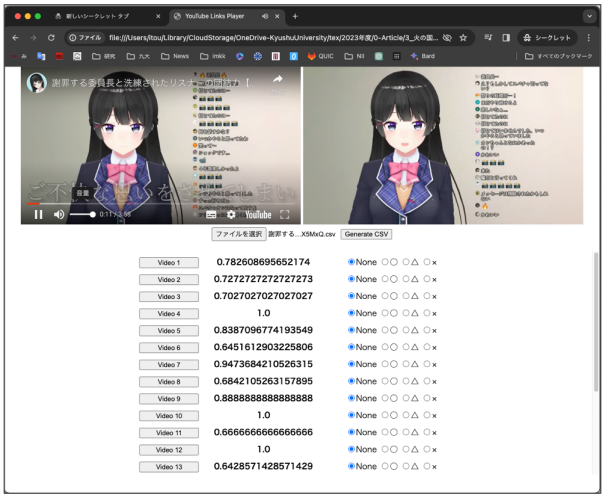

図 2 推定部分チェックツール

4. 実験と評価

実験として，表 1 に記載した Vtuber の動画を対象とした。これらの動画は，Vtuber の再生回数の多い動画と，それを切り抜いた動画のうち再生数最大の動画を選んだ。これらの動画について，YouTube API で字幕テキスト（文字起こし）を入手し，第 3 節の切り抜き部分推定手法を適用した。ただし字幕テキスト（文字起こし）が存在し無い動画や，字幕の言語が日本語に設定されていない動画は対象から外した。

表 1 に分析した動画の ID と，切り抜き動画 1 つにおける，切り抜き部分の平均推定精度を示す。なお，`https://www.youtube.com/watch?v=` の後に動画 ID をつけた URL にアクセスすると YouTube の動画を閲覧できる。全動画の推定精度の平均は 57.0%であった。この結果から，本研究で提案および実装した，文字起こしテキストを用いた切り抜き部分推定手法の精度は高いと言えない。

精度が高くない理由はいくつかある。まず，YouTube の自動文字起こし精度の低さがある。元動画と切り抜き動画の音声が同じであっても，異なる文字列を抽出している場合がある。複数人が同時に話す会話は，文字起こしを間違いやすい。ゲーム実況動画で，背景にゲームの音楽や効果音が流れている場合も，会話の文字起こしを間違いやすい。

次に Vtuber によるキャラクタ設定に合わせた言い回しも低精度の理由になる。ここで言い回しとは，「〜なのだ」「〜のじゃ」のように語尾を特徴的する場合などがある。こ

のような喋り方をすると，文字起こしの文字列に類似文字列が出現しやすくなる。そのため推定精度が高くならない。

5. おわりに

本論文では YouTube の人気動画と，その切り抜き動画を対象に，切り抜かれた部分が元動画のどの部分かを推定する手法を提案し，ツールを試作した。切抜き部分の推定には，動画の文字起こしテキストからの，文字列類似度算出を用いた。

数名の Vtuber の動画を対象に，元動画と切り抜き動画に対して提案手法を適用してみた。その結果，全動画における推定精度の平均は 57.0%であった。YouTube が自動的に生成する文字起こしテキストだけでは十分な精度を得られない。推定精度向上には，動画の画像や音声を用いたほうが良いかもしれない。今後はそれらの情報用いて推定精度上げ，最終的な目標である機械学習による動画ハイライト部分および要約部分の抽出を実現したい。

参考文献

[1] 稲田豊史, 映画を早送りで観る人たち, 光文社 (2022).
[2] 小城凱, 伊東栄典: 動画のハイライト部分自動抽出に向けた検討, FIT2023, F-035 (2023).
[3] Guoqiang Liang, et.al.: Video summarization with a convolutional attentive adversarial network, Pattern Recognition, vol.131 (2022).
[4] Yubo Zhao, , et.al.: Topic-aware video summarization using multimodal transformer, Pattern Recognition, vol.140 (2023).
[5] 赤木信也: 盛り上がり検出のための音声解析の一考察, FIT2023, F-014 (2023).
[6] Google: YouTube Data API v3, `https://developers.google.com/youtube/v3?hl=ja` (最終アクセス 2023/12/16).

表 1 対象動画と精度

元動画 ID	切り抜き動画 ID	平均精度 (%)
壱百満天原サロメ		
AniQdstq4KM	ww1DfXDM4Dk	47.0
sleq-_yFHdw	NREMaz_xzSI	27.0
-mHkg7tA83Q	EDHkI3KkMAw	15.0
jagxPn2LiFg	-CADozkx28A	39.0
HBmYTvq4EPw	F2QoYBU339c	42.0
葛葉		
fbaoedS3Vcs	6Jw8HgA8Vik	56.0
X7CO4IUkBTQ	9S4qtw07KMo	59.0
cCQza5SvQ4E	zzN-_fS_tjw	20.0
叶		
MzskmTemxNc	uXi9E0psumE	53.0
NBrowMosk9c	uVZH-XGAnFk	63.0
ZaHXxFkLwu0	Wd7Hk3pkIeY	62.5
AZUeOoTvxAU	Le7MqrkLa6E	47.0
月ノ美兎		
93_cEPLCHfo	vLo_fqNCFHg	64.0
jpJXyaX5MxQ	PBkpTcb7xko	85.0
gB2E54dzk30	7MvKqWuoTvQ	55.6
GpBhQCVyfp0	E7uGBzYxwmc	73.0
TPa8Gqj75kI	R3ZYG1z1y3U	64.9
星川サラ		
xcTuEvgB4r8	a0R0pDy0-pM	54.0
7IVtIixUHBI	ZbsA1bvMPbU	78.2
HuFhYJDzsfQ	3KRkmiVY84g	92.0
CGq1gAPETvg	kFNyRvA4_wA	58.1
HCtBjFTvcBI	dEu43I5Iec0	50.0
剣持刀也		
RW-usgHwOEA	hdw8NO6U5b4	42.1
SBEJ-mR1FrA	V-3KX5JNlEg	35.0
P8vUJX6Afsg	wREt3BDtsDc	62.6
Blmi20mq2cA	Y-GN_3Krcw8	87.0
DtpuybgZ2ik	XvukiLwjmOg	61.5
AGHqitxXJoI	PFXCa6eZRlc	83.3
	全体の平均精度	57.0