動画タイトルを用いたサムネイル画像の自動選択手法の提案

嘉田 紗世[1] 山野 陽祐[1] 新美 茜[1] 田森 秀明[1] 小海 則人[1] 岡崎 直観[2] 乾 健太郎[3,4]

[1] 株式会社朝日新聞社 [2] 東京工業大学 [3] 東北大学 [4] 理化学研究所
{kada-s, yamano-y, niimi-a, tamori-h, kokai-n}@asahi.com,
okazaki@c.titech.ac.jp, inui@tohoku.ac.jp

概要

本研究では，動画のフレームから動画タイトルに関連した最適なサムネイル画像を自動で選択する手法を提案する．さらに，隣接するフレーム間の相関が高いという動画の特徴（冗長性）に対応し，性能を向上させるための手法についても言及する．本手法は画像とテキストのマルチモーダルな事前学習済みモデルを利用し，学習データに動画自体を必要としない．先行研究と性能を比較し，有効性を確認した．また，朝日新聞社の動画を使った実験で，構築したパイプラインの有効性を示した．

1 はじめに

サムネイルは画像や動画の情報量を削減したもので，ユーザの第一印象を左右する．ゆえに，その品質は重要である．一方，サムネイルの素材となるフレームを選ぶ作業（以下，サムネイル選択と呼ぶ）は労力を要するため，動画を大量に制作する現場では，効率化の観点からサムネイルの自動生成に対する期待が高い．

サムネイル選択の自動化では，動画とサムネイルのパラレルデータを用いた教師あり学習でモデルを構築するアプローチが考えられる．しかし現状，動画制作者によって手動で作成された動画・サムネイルデータは少なく，特に日本語のデータセットは筆者らの知る限り，存在しない．そこで本研究では，それらの代替として，画像とテキストのパラレルデータを用いたモデルを利用する．このようなデータセットは日本語でも存在し[1)]，特に朝日新聞社では，掲載された過去の記事から大規模かつ効率的にデータセットを構築することが可能である．

タイトルは動画制作者が動画を端的に説明するために付与したものであるため，動画とそのタイトルの間には，動画（画像の集合）からテキストへのマルチモーダルな「要約」の関係が成り立つと考えられる．そこで，タイトルとの関連性が高いフレームは動画の主題を表すと考え，画像とテキストの類似度を測るモデルを用いてサムネイルを選択する．さらに，動画の冗長性の排除を狙いとしたピーク検出[2)]と，複数のサムネイル候補の相対的な順位を考慮するランキング学習についても検討した．なお，本研究は，サムネイル画像として適切なフレームを動画から選択する作業の支援を目的としている．

事前に画像とテキストのパラレルデータで学習したモデルで各フレームと動画タイトルの類似度を計算し，最も類似度が高いフレームをサムネイルとする手法の有効性を，先行研究との性能を比較することにより確認した．また，朝日新聞社の報道動画を使った実験では，ピーク検出により動画の冗長性を排除し，効果的にサムネイル候補（キーフレーム）を抽出できることや，ランキングモデルを組み合わせることで性能が向上したことを確認した．

2 関連研究

サムネイル選択手法として，美的品質と代表性などの視覚情報に基づく手法が提案されている．代表的な手法として，Song らの研究 [2] がある．Song らは，暗い・ボケといった低品質のフレームの破棄，視覚的に類似したフレームのクラスタに基づく重複フレームの破棄，抽出したキーフレームの美的品質及び代表性（キーフレームを含むクラスタサイズから算出）による評価，という手順でサムネイル選択を行った．本研究では報道動画での実応用を見据え，視覚情報に加えて動画の主題との関連性も重視する．そこで，Song らとは異なり，動画のタイトル情報を活用する手法を検討する．

1） 代表的な日本語画像キャプションデータセットとして STAIR Captions [1] がある．

2） 評価値の極大値及びその位置を検出する処理．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

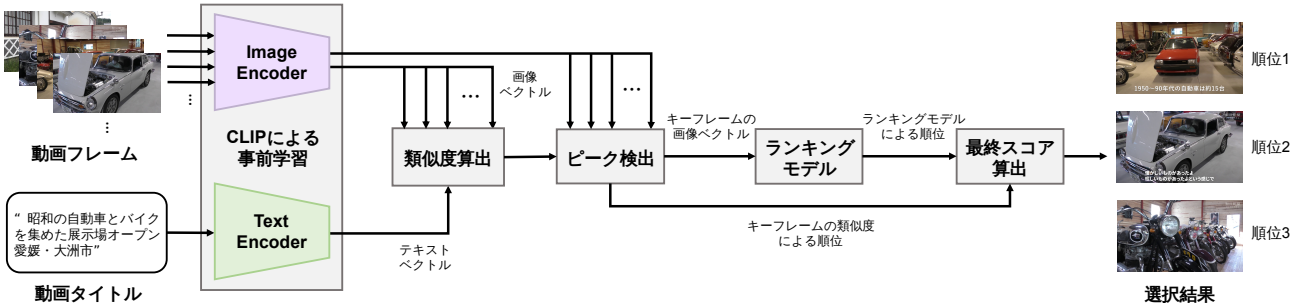

図 1　提案手法の概要．CLIP によりフレームとタイトルの類似度を計算する．ピーク検出によって冗長性を排除し，キーフレームを抽出する．CLIP で計算した類似度，ランキングモデルの結果を組み合わせ，キーフレームを順位付ける．

タイトル情報を活用する先行研究に Yu らの研究 [3] がある．Yu らは，視覚情報に加えて動画のタイトルや説明文，および動画中の音声を用いた埋め込み表現を学習したモデルを構築し，このモデルにより生成された動画を表すベクトルと最も類似度が高いフレームをサムネイルとして選択した．本研究では，動画を学習データに用いる手法とは異なり，画像とテキストのマルチモーダル事前学習済みモデルを活用したサムネイル選択手法を提案する．

3　提案手法

図 1 に本手法の概要を示す．前述のとおり，動画の主題をよく表すフレームを選択するために，画像とテキストの類似度を測るモデルを用いる．本研究では Contrastive Language-Image Pre-Training（CLIP）[4] を採用する（3.1 節）．しかし，動画中には同じようなフレームが多数存在することから，CLIP のみでサムネイルとなるフレームを選び出すのは困難である．そこで，テキストとの類似度に加えてピーク検出を導入する（3.2 節）．さらに，編集者の画像評価を再現するランキングモデルを構築し，より最適なサムネイルを候補から選択する（3.3 節）．

なお，本研究では動画から一つ以上のサムネイルを抽出する．一つ以上とするのは，本手法で選択した複数のフレームからサムネイルを人手で選択するという，実際の業務フローを想定するためである．

3.1　CLIP

CLIP は，2021 年に OpenAI が提案したニューラルネットワークである．大量の画像・テキストペアの対照学習を通してマルチモーダルな埋め込み表現を学習し，多くのタスクにおいてゼロショット学習で優れた精度を達成している．CLIP を使用することで画像とテキストをそれぞれ特徴量としてベクトル化し，その類似度を計算できる．

本研究で CLIP を用いる理由は，画像とテキストの対応関係を直接学習することで，その相互理解に向けたより汎用的なモデルが構築できると期待できるからである．タイトルに基づくサムネイル選択を実現するには，物体認識モデルの利用も考えられる．しかし，本研究の主な対象である報道動画においては，主題を表しているが即物的でないタイトル（例：「晩秋の尾瀬，金色に輝く 今だけ見られる景色」）も含まれているため，従来の物体認識モデルだけでは不十分である．そこで，テキストの文脈を考慮できるモデル構築が可能な CLIP を採用し，CLIP で計算される各フレームとタイトルの類似度（以下，CLIP スコアと呼ぶ）をサムネイル選択に利用する．与えられた動画の各フレームの CLIP スコアを算出し，CLIP スコアがより高いものを動画の主題をよく表すフレームとする．

3.2　ピーク検出によるキーフレーム抽出

冗長性を排除しつつ CLIP スコアの高いフレームを取得するため，以下の手順でピーク検出を行い，キーフレームを抽出する．

1. CLIP スコアを時系列データとする
2. CLIP スコアの移動平均を算出
3. CLIP スコアの移動平均からピークを検出
4. ピークの近傍からキーフレームを抽出

概要を図 2 に示す．手順 1 でプロットした CLIP スコアに対して，手順 2 で移動平均を算出して外れ値を除去する．手順 3 では，移動平均が極大となるフレームを見つける．手順 4 では，手順 3 で抽出したフレームの前後 30 枚から，同一シーン内で CLIP スコアが最大のフレームを抽出する．

この段階では，キーフレーム間の順位は CLIP スコアから決定する．また，再生時間は動画毎に異なるため，抽出されるキーフレーム数は不定である．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

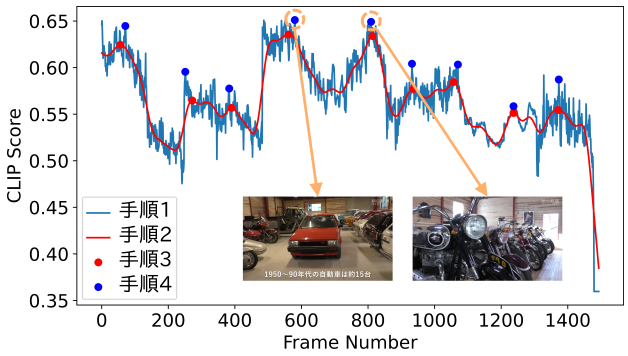

図 2　キーフレーム抽出の概要．凡例と本文中の手順 1〜4 は対応している．

3.3　ランキングモデル

サムネイルは動画の顔であるため，動画の主題を表すだけでなく，ユーザの興味を引くフレームを動画制作者は選択したい．そこで，3.2 節で抽出された複数のキーフレームに対し，より良い画像を順位付けするランキングモデルを構築する．このランキングモデルの学習データには，朝日新聞デジタルの記事に含まれる画像と，その表示順を利用する．記事に複数の画像が付与されているとき，それらの画像は編集者の考える優先度順に並べられる．この順序を学習させ，編集者の画像評価を再現するモデルを構築する．

キーフレームの最終スコア x は CLIP スコアでの順位 m，ランキングモデルでの順位 n を考慮し，以下の式から算出する．

$$x = 2^{-m} + 2^{-n} \tag{1}$$

ここでは，どちらかの順位が高ければ x が高くなるスコア付けをする．なお，x が等しい場合，ランキングモデルの順位を優先してフレームを採用する．

4　実験

本研究では，2 つの実験を通じて提案手法を評価する．4.1 節では，先行研究との性能比較により，各フレームと動画タイトルとの関連性によるサムネイル選択の有効性を検証する．4.2 節では，本研究の主な応用先である朝日新聞社の報道動画を用いて，提案手法の有効性を検証する．実験は以下の設定で行う．

- **CLIP**: 全フレームから CLIP スコアの上位 k 個を選択
- **CLIP+peak**: 3.2 節の手法でキーフレームを抽出・順位付けし，上位 k 個を選択
- **CLIP+peak+rank**: 3.2 節の手法で抽出したキーフレームを式（1）から算出した最終スコアで順位付けし，上位 k 個を選択

なお，4.1 節では単純に CLIP スコアによるサムネイル選択の性能を評価するため，**CLIP** の結果のみを示す．

4.1　先行研究との性能比較

この実験では，正解フレームとの一致に基づいた評価を行うため，正解フレームを収録した Yahoo データセット[3]を用い，Song ら，Yu らの手法と性能を比較する．評価指標として，Yu らの実験で採用されたものを用いる．

データセット　Yahoo データセットには，動画制作者が生成したサムネイルに対応するフレーム番号が含まれており，そのフレームを正解とする．Yu らの実験で使用された 71 件の評価データを特定できなかったため，Yahoo データセットの 429 件の動画（取得できた動画のうち，タイトルが英語のもの）からランダムに 71 件の動画を評価データとして抽出し，性能評価に用いる．評価データセットは 5 つ作成し，5 回の実験の平均値を結果として示す．

評価指標　Yu らの指標に従い，最上位（$k = 1$）のサムネイル候補 c が正解フレーム f^* と一致する場合に予測が正しいと見なし，適合率によって性能評価を行う．2 枚の一致は，VGG16 モデル [5] で抽出した特徴量の近さから判定する．VGG16 モデルの実装は Keras のものを用いた[4]．動画内には類似したフレームが存在することを考慮し，閾値 θ 以下で一致とみなす．与えられた θ に対して，サムネイル候補 c と正解フレーム f^* の間に，

$$\|v(f^*) - v(c)\|_2^2 \leq \theta \tag{2}$$

という関係が成り立つとき，サムネイル候補 c の抽出は正しい（真陽性）と判定する．ここで，$v(.)$ は VGG16 で計算されたフレームの特徴ベクトルを表す．θ の値を変えながら適合率を計算する．

CLIP モデル　Yahoo データセットのタイトルは英語のため，OpenAI 提供の CLIP を使用する[5]．

実験結果　表 1 に Song ら，Yu らのモデルと，**CLIP** でのサムネイル選択の実験結果を示す．**CLIP** による選択は，$\theta = \{750, 1000\}$ では先行研究よりも性能が劣るが，より厳密な評価である $\theta = 500$ では

3）https://github.com/yalesong/thumbnail
4）https://keras.io/ja/applications/
5）https://github.com/openai/CLIP

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1 Yahoo データセットにおける性能．Song ら，Yu らの手法を CLIP スコアのみを用いた手法と比較した．

モデル	適合率@θ ($k=1$)		
	$\theta=500$	$\theta=750$	$\theta=1000$
Song ら	0.113	0.267	0.601
Yu ら	0.197	**0.408**	**0.648**
CLIP	**0.256**	0.366	0.530

先行研究よりも性能が高い．この結果から，タイトル及び CLIP スコアがサムネイル選択に有効であることが分かった．

4.2 報道動画での実験

データセット 朝日新聞社 YouTube アカウントで 2021 年 9〜11 月に公開された動画を使用する．長時間のノーカット会見や，制作フローが違うと考えられるマンガ動画は対象から除外した．さらに，サムネイルに対応するフレームが動画内に存在しない動画を破棄し，204 本の動画を評価対象とした（平均再生時間 81.4 秒）．サムネイルに対応するフレームの有無と正解フレーム番号は目視で確認した．

評価指標 3 節で述べた実際の業務フローを想定し，本手法で予測した上位 k 個 ($k \in \{1,3,5\}$) の候補のうち，少なくとも一つが正解フレームと一致する場合に予測が正しいとみなす．4.1 節同様，真陽性は式（2）によって判定する．簡単のため $\theta = 500$ の結果のみを示す．

CLIP モデル 日本語の動画に適用するため，日本語の CLIP モデルを構築した．データセットとして，朝日新聞デジタルに掲載された記事に含まれる画像とテキストから作成したパラレルデータ，計 548,117 件を用いた．テキストには，画像キャプションと見出しを使用しており，1 枚の画像に対して 2 つのパラレルデータを作成した．実装は Moein のものを用いた[6]．Text Encoder には Sentence-BERT [6]，Image Encoder には Vision Transformer [7] を使用した．

キーフレーム抽出 3.2 節の手順による．手順 2 における移動平均の窓幅は，全フレーム数 N を用いて，$\lfloor \frac{N}{10} + 0.5 \rfloor$ とした．手順 3 では，移動平均線の極大値を検出する．ここで，極大値同士の距離 *distance* および極大値前後の高低差 *prominence* はそれぞれ，$distance \geq 60$，$prominence \geq 0.008$ とし，これらのパラメータは実験的に決定した．手順 4 では，検出したピークと異なるシーンのフレームの抽出を防ぐため，ニューラルネットワークを用いたシーン境界検出 [8] を行い，その結果を利用した．

6） https://github.com/moein-shariatnia/OpenAI-CLIP

表 2 構築したパイプラインの性能検証結果．朝日新聞社の動画を評価に用いた．

モデル	適合率@k ($\theta=500$)		
	$k=1$	$k=3$	$k=5$
CLIP	0.240	0.294	0.343
CLIP+peak	0.270	0.441	0.549
CLIP+peak+rank	**0.275**	**0.466**	**0.569**

ランキングモデル モデルは CatBoost [9]，損失関数は StochasticRank [10] を使用した．学習データには朝日新聞デジタルの記事画像と表示順序情報 70,737 件 (1 記事の画像が 5〜10 枚の記事 11,379 件分) を使用した．画像の特徴量は Vision Transformer で抽出した．評価尺度として nDCG@5 を用い，テストデータで nDCG@5 = 0.855 と確認した．

実験結果 3 つの実験設定での性能評価の結果を表 2 に示す．**CLIP** のみでは，動画の冗長性から，似通ったフレームが選択されてしまい，k を 3 や 5 に増やしても性能があまり上がらない．**CLIP+peak** の結果から，ピーク検出によるキーフレーム抽出は多様で適切なサムネイルを選択するのに効果的であることが示された．なお，ピーク検出で得られたキーフレーム数の平均値は 8.6 であった．さらに，**CLIP+peak+rank** において，$k = \{3,5\}$ では約 2%の性能向上が確認できた．以上の結果から，提案手法が実際の業務フローに役立つ可能性が示唆される．

5 おわりに

本研究では，画像とテキストの類似度，ピーク検出，ランキング学習を使って最適なサムネイル画像を選択する手法を提案した．実験から，画像とテキストのマルチモーダル事前学習済みモデルを用いたサムネイル選択の有効性を示した．また，ピーク検出，ランキングモデルを組み合わせることで，性能が向上することが示唆された．

本研究の今後の課題として，動画制作現場における実証実験が挙げられる．特に，本手法で不正解となったフレームに対して，動画制作者の主観的な評価を行い，実運用における性能検証を進めたい．また，現状では計算に GPU を用いているが，実運用では限られた計算資源で，短時間で大量の処理が必要となる．処理パイプラインの軽量化をしつつ，サービスとしての完成度の向上に取り組みたい．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] 吉川友也, 重藤優太郎, 竹内彰一. Stair captions: 大規模日本語画像キャプションデータセット. 言語処理学会 第 23 回年次大会, pp. 537–540, 2017.

[2] Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. To click or not to click: Automatic selection of beautiful thumbnails from videos. In **Proceedings of the 25th ACM International on Conference on Information and Knowledge Management**, pp. 659–668, 2016.

[3] Zhifeng Yu and Nanchun Shi. A multi-modal deep learning model for video thumbnail selection. **arXiv preprint arXiv:2101.00073**, 2020.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. **arXiv preprint arXiv:2103.00020**, 2021.

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

[6] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 11 2019.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In **International Conference on Learning Representations**, 2020.

[8] Tomáš Souček and Jakub Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection. **arXiv preprint arXiv:2008.04838**, 2020.

[9] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. **arXiv preprint arXiv:1810.11363**, 2018.

[10] Aleksei Ustimenko and Liudmila Prokhorenkova. Stochasticrank: Global optimization of scale-free discrete functions. In **International Conference on Machine Learning**, pp. 9669–9679. PMLR, 2020.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).