令和３年度　卒業研究

メタデータを用いない
コンテクスチュアルターゲティングシステムの開発

東京理科大学　理工学部　経営工学科
大和田研究室　7418105　薮谷瑞生

指導教員：大和田 勇人

目次

図目次

表目次

第１章　序論

序論では本研究の背景である動画広告の現状と課題，そして本研究の目的と構成について述べる

1.1　背景

インターネット広告では90年代頃の黎明期においてはバナー広告などやメール広告などのテキストを用いたものが主体であったが，近年では SNS の普及もあり動画を用いたものが多く見られる．また近年インターネット広告費が急激な勢いで増大しつつあり,図１に示す通りテレビ広告費を2019年に追い抜いた．そのためテレビ広告のみならず，インターネット広告における動画広告の重要度が広告全体において非常に大きなものとなっている．しかし現状の動画広告の課題として以下の２点が存在している．ひとつ目はテレビCMの場合大人数に広く・浅く届けることしかできない点である．テレビCMでも番組の内容などを通して主なターゲットとなる視聴者の年齢や性別などを設定し，視聴率などのデータから番組作りにフィードバックするという仕組みが存在する．しかし録画視聴が増え視聴率等のデータが視聴者の実態に即さなくなっていることや，そもそもインターネット広告でのユーザー情報を用いて個々人の好みや購買行動に合わせた広告を表示させる手法と比べてターゲティングの精度は低いということもあり,新たな広告の手法が求められている.ふたつ目はインターネット広告では近年個人情報保護の観点から，Cookie 等のユーザーデータを用いたターゲティングが難しくなりつつあることである．EU 等の各国ではデータ保護規制により，Cookie を利用する場合にはユーザーの許可を取らなければならなくなり，今までのようなCookieの利用は難しくなっている．またサードパーティの Cookie に関しては Google や Apple などがそれぞれのブラウザ上で規制に乗り出しており，今までの手法で広告主がCookieを用いて個々人のユーザーに合わせた広告の最適化を行うことは難しくなっていると言わざるを得ない．このようにテレビ広告，インターネット広告双方で現状の動画広告には課題が存在

し，ユーザー情報を用いずにコンテンツの動画の内容のみでターゲティングを可能にしたいという需要が存在している．

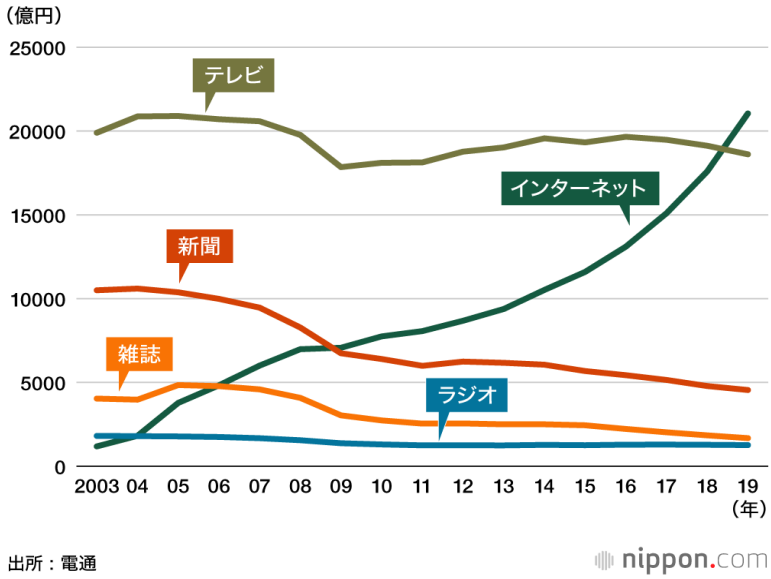

図 1.1: 媒体別広告費の推移

1.2 目的

背景で述べたように，現状ではユーザー情報を利用できることが前提に広告の最適化が行われている．しかし利用できなくなることを鑑みれば，新たな最適化手法が必要となる．そこで本研究では日本の広告代理店と協力し，テレビドラマの動画のシーンのコンテクストを特定することによりターゲティングを可能とするコンテクスチュアルターゲティングという手法を開発する．また本手法においてはコストのかかる台詞等の追加のメタデータを用いず，コンテンツの動画のみからターゲティングを可能とすることを目指す．

1.3 コンテクスチュアルターゲティングとは

コンテクスチュアルターゲティングの概要について説明する．図 2 に示すように，まず動画広告のコンテクストを広告主がタグ付けする．例えばカレーやシチューの CM であれば家族団欒のコンテクストがあるという具合である．次に本手法によりテレビドラマの本編内で CM に最適なシーンを発見し，CM を挿入する．そのためコンテクスチュアルターゲティングを行うためにはテレビドラマの各シーンが CM のコンテクストに合致するか分析する必要があり，本手法の中心となるのはテレビドラマ本編からコンテクストを抽出する部分となる．

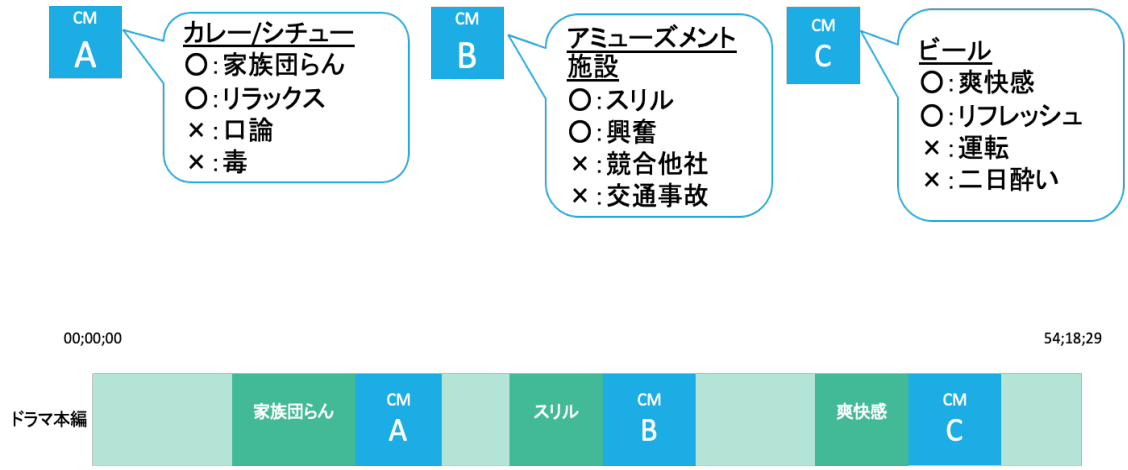

図　1.2: コンテクスチュアルターゲティングの概要

1.4　使用するデータセット

本研究には株式会社博報堂より提供されたあなたの番です第 1 話，マジで航海してます第 1 話，サバイバルウエディング第 1 話，グッドドクター第 1 話という計 4 本のテレビドラマのデータを用いた．各ドラマのデータにはシーン毎に発話内容，タイムスタンプ，BGM の有無とコンテクストに及ぼす効果，CM の直前か否か，登場する人物の感情，動作，写っている物，台詞が無いとコンテクストが判別できないかどうか，直前や直後のシーンにコンテクストが依存するか否か，そしてシーンのコンテクストが記載されている．また本手法によるシステムへの入力として各テレビドラマの動画も用いた．

1.5　提案手法

まずドラマの動画をシーン毎に分割するセグメンテーションを行う．この分割されたシーン毎にコンテクストが存在するか分析することになる．次にドラマ本編のコンテクストを特定するためには，コンテクストを 4 つの要素に分けて捉えることを考える．具体的には感情，動作，映っている物，そして台詞である．ありがとうという台詞が有れば，感謝というコンテクストというように，台詞のみで特定できるコンテクストについては台詞が検出されれば，そのコンテクストを割り振る．台詞のみで特定できないコンテクストについては，提供されたドラマのデータからコンテクストと感情，動作，映っている物の 3 要素の組みあわせを作成し，ドラマから 3 要素が検出された場合にコンテクストを割り振る．感情，動作，映っている物，台詞といった各要素を動画から検出するためには以下の技術を用いる．

・　感情の検出　Face Emotion Recognition
・　動作の検出　Slaw Fast
・　映っている物の検出　Yolo v5
・　台詞の検出　Google Speech to Text

そして最後にシーン毎のまとまりであるロジカルストーリーユニット(LSU)を導入し，コンテクストを検出できなかったシーンに対してコンテクストの補完を行う.

1.6　本論文の流れ

本論文の流れを以下に示す.

第１章　序論
　本研究の概要と，本論文の流れについて述べる

第２章　関連研究
　本研究と関連のある研究について述べる

第３章　使用するデータセット
　本研究で使用したデータセットについて述べる

第４章　提案手法
　本研究で提案する手法について述べる

第５章　実験
　本研究で行った実験とその流れについて述べる

第６章　結果と考察
　本研究の実験で得られた結果とその結果に対する考察を述べる

第７章　結論と今後の展望
　本研究の結論と今後の展望を述べる

1.7　本章のまとめ

本研究の背景，目的，データセット，提案手法，本論文の流れについて述べた．

1.8　本論文の構成

本論文の構成は，第１章にて広告のターゲティングが抱える問題について述べ，本研究の目的を述べた．第２章では関連研究を述べ，第３章では使用するデータセットについて説明する．第４章では提案手法について説明し，第５章では提案手法に基づいた実験について述べる．第６章では実験の結果と結果に対する考察を述べ，第７章では本論文の結論及び今後の展望について述べる．

第２章　関連研究

動画のコンテクストを利用する研究としては以前より映像要約の研究が行われている．本章では自動で要約映像を生成する手法として機械学習を用いないものと用いるものの２つの研究を挙げた上で，本研究の新規性について述べる．

2.1　機械学習を用いない要約映像作成

Tsoneva ら[1] は映画を対象として映像要約の研究を行った.字幕や映画の台本に含まれるテキスト情報を利用しており,キーワード,主要登場人物の名前,存在感などの特徴を抽出している.セグメンテーションは，時間情報付きの字幕と台本に従って手動でサブシーンの検出を行っている.重要度スコアの推定には，キーワード，登場人物，登場人物の存在感などのランキングを線形結合したもので定義している.ランキングには,PageRank アルゴリズムを用いており，各シーンからサブシーンを選出することで，最終的な要約映像の多様性を確保している.

2.2　機械学習を用いた要約映像作成

Yao ら[2]は，画像分類用の CNN（Convolutional Neural Network：畳み込みニューラルネットワーク）および動作分類用の３次元 CNN から抽出した特徴を入力として，各カットの重要度スコアを算出するニューラルネットワークを考案した。算出されたスコアが高いカットを選択することで，要約映像を自動生成する。ネットワーク構造はペアワイズネットワークと呼ばれるもので，人手で作成された要約映像に使われた正例カットの画像を入力するネットワークと，使われなかった負例カットの画像を入力するネットワークで構成される。２つのネットワークの内部パラメーターは共通とし，正例用ネットワークが出力する重要度が高く，かつ負例用ネットワークが出力する重要度が低くなるように学習する。

2.3　本研究の新規性

先行研究に共通する課題としては以下の 2 点が存在する. 1 つ目はコストや労力がかかるメタデータやアノテーションを必要とするデータを使用している点である. Tsoneva らの研究では映画の字幕や台本の台詞等のテキストの情報を利用しており, Yao らの研究では放送時に用いられたのか否かのアノテーションがなされた画像をモデルへの入力として使用している. これらの台詞等のメタデータやアノテーションがなされたデータを人手で作成するためにはコストや労力がかかるため用いないことが望ましいと言える. 2 つ目はどちらの研究もシーンの重要度を算出するにとどまり, どういったシーンなのかを特定することはできないという点である. どちらの研究も手法の違いはあれどシーンの重要度を算出することを中心に据えており, どういったシーンなのかまでは考慮していない. 本研究では以上の 2 つの課題を解決するため, ドラマの動画以外の人手で作成されたメタデータを利用することなくどういったシーンなのかコンテクストまで特定する. このことによりユーザーの情報を利用せずとも広告のターゲティングを可能にすることができると考えられる.

2.4　本章のまとめ

本章では動画のコンテクストを利用する研究として要約映像を自動で生成する研究を２つ挙げ, それらの問題点と参考点を示した. また, 関連研究との相違点を示し本研究の位置づけを明確にした.

第３章　データセット

本章では本研究で使用したデータセットについて述べる．

3.1　データセットの概略

本研究で使用したデータセットは株式会社博報堂より提供されたデータセットである．データはあなたの番です第１話，マジで航海してます第１話，サバイバルウエディング第１話，グッドドクター第 1 話という計４本のテレビドラマに対し，図３のように CSV ファイルとして作成され，それぞれのドラマに同じ内容のデータが収集されている．またそれぞれのドラマの動画も本システムの入力として利用している．

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	ドラマ「あなたの番です」第1話　2019年4月14日 22時30分-23時25分															
2	シーン概要	話者	発話内容	シーン解説文	発話開始TS	発話終了TS	BGM/効果	BGM/効果	CM直前/直	コンテクスト	感情	動作	写っているモノ（物体検知）	台詞	台詞がないと成立しない	直前直後のシーンに依存
3	役者・竹中直人が視聴者に殺人事件について語りかける	竹中直人	あなたには、殺したい人はいますか？		0:00:04.00	0:00:10.24	有	影響する	どちらでも	緊張	-	-	人	殺し	成立する	
4		竹中直人	誰だって周りに、嫌いな人、殴ってしまいたい人、目の前から消えてほしいなんて思う人は何人かいるでしょう		0:00:10.24	0:00:19.25	有	影響する	どちらでも	緊張	-	-	人	-	成立する	

図　3.1: データセットの一部

3.2　データセットの詳細

本研究に使用したデータセットの詳細について述べる．表 1 はデータセットに収録されたデータの表記と詳細についてまとめたものである．これらのデータはシーン毎に作成されており，CSV ファイルの 1 行が 1 つの発話に対応している．各ドラマの動画のデータ数はあなたの番です第 1 話が 757，マジで航海してます第 1 話が 371，サバイバルウエディング第 1 話が 984，グッドドクター第 1 話が 617 となっている．

表 3.1: データセットの詳細

表記	詳細
シーン概要	どういったシーンなのかの説明
話者	誰が話しているかの説明
発話内容	発言内容
シーン解説文	シーンの捕捉
発話開始 TS	発話を開始した時間
発話終了 TS	発話を終了した時間
BGM/効果音	BGM や効果音の有無
BGM/効果音がコンテクストに影響	BGM や効果音がコンテクストに影響しているか否か
CM 直前/直後	CM の直前または直後かどうか
コンテクスト	コンテクストの種類
感情	登場人物の表情から推定される感情
動作	登場人物の動作
映っているモノ	画面に映っているもの
台詞	発話内容の中でコンテクスト特定につながる単語
台詞がないと成立しない	台詞がコンテクスト特定に不可欠か否か
直前・直後のシーンに依存	コンテクストが直前・直後のシーンに依存するか否か

3.3　本章のまとめ

本章では，本研究で使用するデータセットについて概略を述べた後に，データの表記，詳細な説明そしてデータ数について述べた．

第４章　提案手法

本章では本研究で提案する手法について述べる．

4.1　提案手法の概要

本研究で提案する手法を大きく分けて２段階に分けることができる．1 つ目はコンテクストの分解であり，2 つ目は要素の検出である．コンテクストの分解とは図４に示すように，コンテクストをキーワード，感情，動作，物体の 4 つの要素に分けて考えることである．ただし例外としてコンテクストにはキーワードのみで特定することのできるものが存在する．例えば感謝というコンテクストの場合は，ありがとうというキーワードが台詞から検出されれば即座に特定することができる．詳しくは 4.3 節で述べるものの，こういったコンテクストとキーワードの組は 97 組存在する．この組以外のコンテクストの場合はキーワード以外の３要素を使って特定していくことになる．例えば家族団欒というコンテクストの場合，登場人物から楽しいという感情，食べるという動作，そして動画から大人数が検出された場合に特定できると考える．次に 2 つ目の要素の検出についてはコンテクストの分解された要素を検出する段階である．この要素の検出において使用する技術は以下の 4 つである．

- 感情の検出　Face Emotion Recognition
- 動作の検出　Slaw Fast
- 映っている物の検出　Yolo v5
- 台詞の検出　Google Speech to Text

これらの技術については 4.4 節にて詳細を述べる．他にも 4.2 節で説明する前処理としてのセグメンテーションや精度向上のための 4.5 節で述べる LSU の導入が本手法には存在する．次節からは本システムで行う処理の順番通りにそれぞれの処理の詳細について説明する．

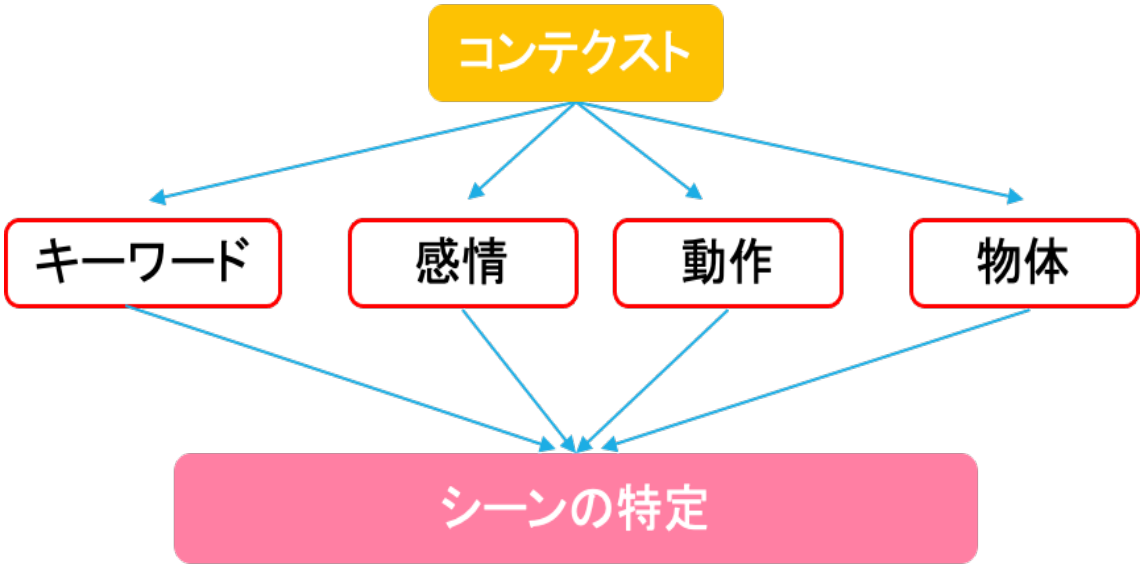

図 4.1: コンテクスト特定のイメージ

4.2 動画のセグメンテーション

本節では前処理にあたる動画のセグメンテーションについて説明する．セグメンテーションとは動画をシーンという小さな単位に分割することである．このセグメンテーションにはpythonライブラリであるPyscene detectを用いる．このライブラリではフレーム画像の輝度が大きく変化した場合にシーンを区切る．具体的には各フレーム画像の全ピクセルでチャンネル方向に輝度の平均を計算し，フレーム画像間でその値が30を超えて変化したときにシーンを分割する．このセグメンテーション後の各シーンに対して4要素の検出を行い，コンテクストの特定を行っていくこととなる．

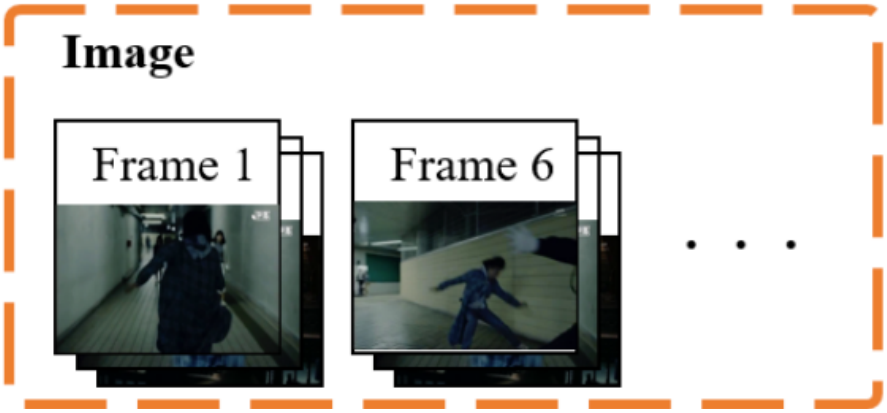

図 4.2: セグメンテーションのイメージ

4.3　コンテクスト特定のための辞書作成

4.1 節で述べたように，本手法ではコンテクストを 4 つの要素に分けて考える．本節ではいかにコンテクストを 4 つの要素に分けて考え，その結果をまとめた辞書を作成するかについて説明する．作成する辞書は 2 つである．1 つはキーワードのみで特定できるコンテクストについての辞書であり，もう一方は感情・動作・映っているものの 3 要素を使い特定するコンテクストについての辞書である．2 つの辞書のどちらも 3 章で説明したデータセットより作成される．1 つ目のキーワードのみで特定できるコンテクストについての辞書は，データセットの中の台詞がないと成立しないの値が「しない」の場合にコンテクストと台詞の値から作成した．その結果が表 2 である．この表に記載されたキーワードが検出された場合，即座に対応するコンテクストであると特定される．2 つ目の感情・動作・映っているものの 3 要素を使い特定するコンテクストについては，まずデータセットのコンテクストの欄が空欄でなく，加えて台詞がないと成立しないの値が「する」場合に感情・動作・映っているものの値を収集して辞書を作成した．次に動作検出と映っているものの検出に用いるモデルが検出できない動作や物体は辞書から削除した．この処理は，動作検出では kinetics400 のラベル[3]，映っているものの検出では COCO データセットのラベル[4]にそれぞれ含まれていない動作や物体を辞書から削除するものである．その結果が表 3 であり，表 2 と表 3 に記載されているコンテクストが本手法で検出できるコンテクストとなる．

表 4.1: キーワードのみで特定できるコンテクストの辞書

コンテクスト	キーワード	コンテクスト	キーワード
幸せ,感謝	ありがとう	遅い	まだ
空腹	お腹すいた	不健康	大丈夫
汚い	シミ	不健康	痛い
感謝	ありがとう	早い	早く
遅い	遅れ	爽快	ナイスショット
早い	まだ	不健康	ハァ
清潔	清める	不健康	熱
清潔	清掃	難しい	リスク
難しい	難し	遅い	遅い
緊張	殺し	安心	さすが，素晴らしい

表 4.1: キーワードのみで特定できるコンテクストの辞書

贅沢	万	安心	見事
贅沢	高い	安心,めでたい	無事，成功
すっきり,感謝	片付いて,ありがと	安心,めでたい	よかった
空腹	何食べる	感謝,めでたい,安心	ありがとう
健康	内腹斜筋	難しい	単純じゃない
難しい	分かんない	難しい	分かりません
美味しい	おいしかった	めでたい,幸せ	誕生日プレゼント
感謝	ありがと	めでたい,幸せ	誕生日
健康	糖質	めでたい,幸せ	ハッピーバースデー
健康	体形維持	めでたい,幸せ	おめでとう
不健康	太っちゃう	めでたい,幸せ	お祝い
一生懸命	お願い	めでたい	グッジョブ
一生懸命	頑張ります	空腹	食べてなかった
一生懸命	-	空腹	空腹
一生懸命	何とか	美味しい	おいしい
かんたん	簡単	不健康	バランスが偏る
感謝	感謝	めでたい	夢を実現
健康	ヘルシー	早い,簡単	早く
お得	安い	ダサい	ナルシスト
かんたん	余裕	緊張	緊張しろ
健康	顔色	不健康	糖質
おいしい	おいしい	めでたい	昇格
一生懸命	やる気	早い	スピード
緊張,難しい	たったの，特別な	贅沢	高価
難しい	大変	贅沢	つい買って
安心	これなら	汚い	やめて
安心,めでたい	終わり	汚い	着色
緊張	危険	感謝	サンキュ
遅い	いつ	感謝,幸せ	ありがとう

表 4.1: キーワードのみで特定できるコンテクストの辞書

早い	急げ	不健康	甘い物控える
安心,めでたい	無事，大丈夫	安心,健康	元気，安心
早い	長い	健康	元気，応援
難しい,リラックス	間違え	簡単	簡単
難しい,リラックス	難しい	感謝	センキュー
難しい	無理	美味しい	おいしそう
不健康	口内炎	難しい	食いづれぇ
遅い	遅く	難しい	食べづらい
空腹	おなかすいた	感謝,リラックス	センキュー

表 4.2: 感情・動作・映っているものを利用して特定するコンテクストの辞書

コンテクスト	感情	動作	映っているもの
緊張	fear	-	person
緊張	angry	-	person
幸せ,楽しい	happy	-	person
幸せ	happy	opening	person
早い		running	person
健康	-	streching	person
緊張	sad	-	person
乾杯	happy	-	person,wineglass
乾杯,美味しい	happy	drinking	person
飲む	-	drinking	person
美味しい	happy	eating	person
食べる	-	eating	person
リラックス	happy	-	bed
リラックス	-	-	bed
幸せ	happy	hugging	person
飲む	-	drinking	cup

4.4　４要素の検出

本節ではキーワード，動作，感情，映っているものをどうやって検出するのかについて述べる．

4.4.1　キーワードの検出

キーワードの検出では Google speech recognition[5]を用いる．api として整備されており，複数の言語に対応している．本手法ではセグメンテーションを終えた動画を wav ファイルに変換した後に api に読み込ませて使用する．検出した発話内容の中にキーワードが含まれていたら対応するコンテクストを割り振る．

4.4.2　動作の検出

動作の検出では SlawFast[6]という機械学習モデルを用いる．機械学習における動作認識とは動画のクラス分類のタスクであり，入力された動画に対して動画内で行われている動作を予測する．本モデルは kinetics400 データセットで学習されており，このデータセットに含まれる 400 個の動作を検出することができる．本モデルの特徴は slow pathway と呼ばれる低フレームレートで空間的情報を捉えるネットワークと fast pathway と呼ばれる高フレームレートで時間的情報を捉えるネットワークが組み合わさって構成されていることで高い精度と速い処理速度を兼ね備えた実用的なモデルとなった．本手法ではセグメンテーション済みの動画を入力とし，各シーンでの動作を検出する．

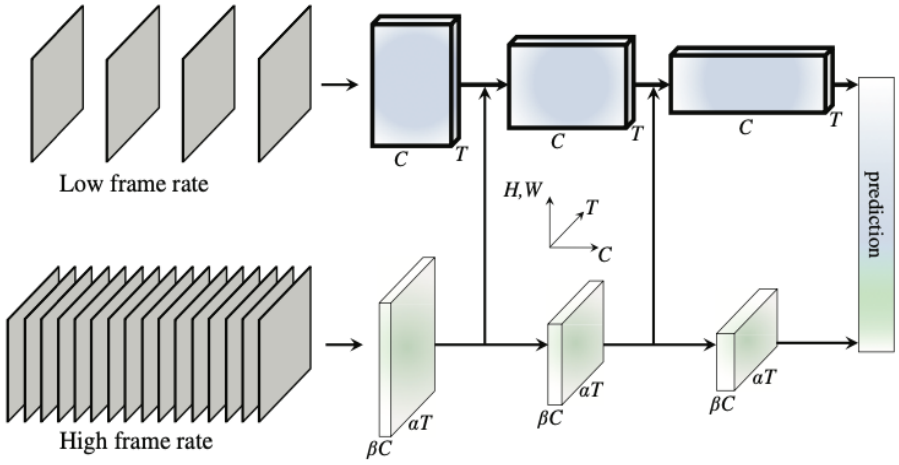

図　4 .3: SlawFast のネットワーク構成

4.4.3　感情の検出

感情の検出は顔認識の前処理を行なった後に行う．顔認識ではRetina Face[7]という機械学習モデルを利用する．この処理を行うことで感情認識が不能であることが分かっている動画に対しての処理時間を無くすことができる．感情検出ではFace Emotion Recognizer[8]という機械学習モデルを用いる．この機械学習モデルはMTCNN[9]というモデルをベースに構築されており，P-Net，R-Net，O-Net という 3 つの畳み込みニューラルネットワークを重ねて用いることで顔の位置の検出，顔の特徴点の検出，そして感情認識という一連のタスクを一つのモデルで行うことに成功している．本手法ではセグメンテーション済みの動画に対して顔が映っているかを検出し，検出された場合には感情の検出を行う．予測できる感情は以下の 7 つである．

・怒り
・嫌悪
・恐怖
・幸せ
・悲しみ
・驚き

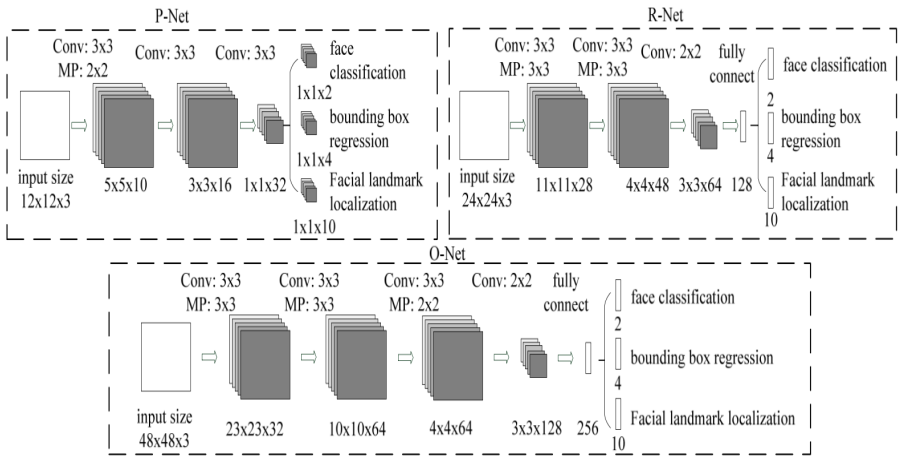

図　4 .4: Face Emotion Recognizer のネットワーク構成

4.4.4　映っているものの検出

映っているものの認識ではYolo [10]を用いる．Yolo とは検出窓と呼ばれるフィルターをスライドさせる仕組みを用いることなく，図８に示すような深い畳み込みニューラルネットワークを用いることで従来のモデルと比べて８倍程度

高速に処理を行うことを可能としたモデルである．本モデルは COCO データセットで学習されており，このデータセットに含まれている 91 個の物体を認識することができる．本手法ではセグメンテーション済みの動画を入力として，各シーンに映っているものを検出する．

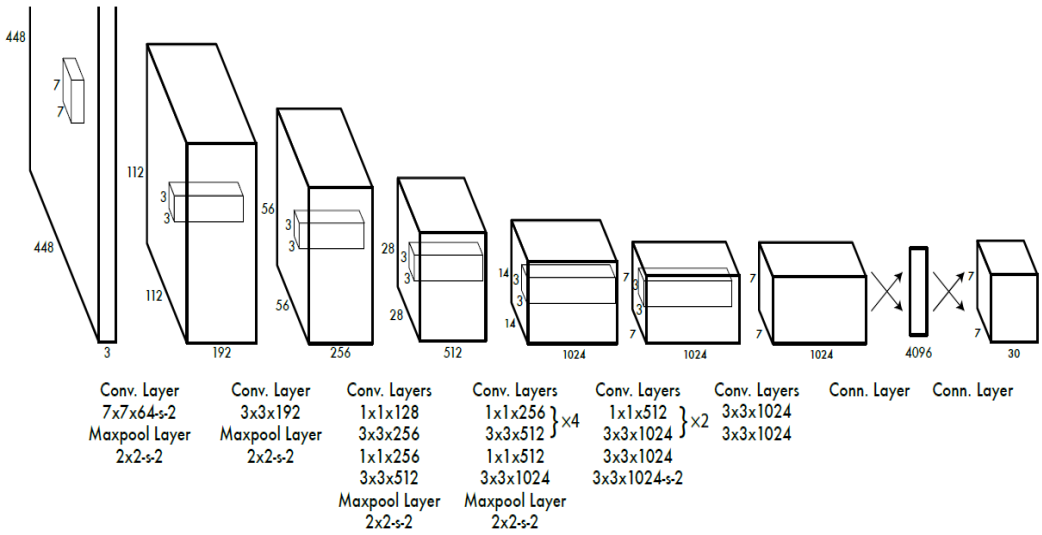

図　4 .5: Yolo v5 のネットワーク構成

4.5　LSU によるコンテクストの補完

本節では LSU と LSU によるコンテクストの補完について述べる．本手法により各シーンの動画に対してコンテクストが予測されるが，全ての動画で予測が可能という訳ではない．4.3 節で説明した 2 つの辞書に記載されていない組み合わせで要素が検出されるとコンテクストを割り振ることができないため，割り振られなかったシーンの動画に対してできるだけコンテクストを割り振ることは精度の向上につなげるために必要なことである．LSU は対話シーンを中心とした動画の単位である．LSU 単位でセグメンテーションを行うと 4.2 節で説明したフレーム画像でのセグメンテーションに比べて，1 つ 1 つのセグメントの時間が長くなる代わりに人にとって理解しやすい自然なセグメントとなる．コンテクストの補完では図 9 に示すように LSU 単位でセグメンテーションを行った後に LSU 内の全シーン動画で多数決を行い，一番多かったコンテクストを他のシーン動画に割り振る． LSU 単位のセグメンテーション手法は大塚[11]を参考にした．LSU 単位のセグメンテーションでは，各シーン動画の代表フレームを選出し，近傍の代表フレームとの類似度から LSU 境界の検出を行う．この

ため，始めにフレームのシーケンスに対してショット境界の検出を行い，得られたショット集合の部分集合からその集合を代表するようなフレームを選出する. 代表フレームの選出では各シーン動画内で最も画質のよいフレームを代表フレームとして定義した．各シーン動画から選出された代表フレームのシーケンスは対称行列で表現でき，i 番目と j 番目の代表フレームが類似している場合は i 行 j 列の要素$s_{i,j}$を 1, 類似していない場合は 0 とする．k 番目の代表フレームに着目した際，近傍の代表フレームにおいて対話シーンに現れる視覚パターンがあるかどうかは，対称行列の二重和である $s^{(k)} = \sum_{i>k, j<k} s_{i,j}$ $(1 < j < k < i < n)$が 1 以上かどうかに対応する．ここで n はショットの総数を表す．LSU 検出は$s^{(k)}$と$s^{(k-1)}$, (1<k<n, $s^{(1)} = 0$) を用いて，以下のルールに従って行われる．

・$s^{(k-1)}$ = 0 かつ$s^{(k)}$>1 を満たす場合，k-1 番目のシーン動画が LSU の開始を示す境界

・$s^{(k-1)}$ > 1 かつ$s^{(k)}$ = 1 を満たす場合，k 番目のシーン動画が LSU の終了を示す境界

上述したような手法の時間・空間計算量はともに $O(n^2)$ であるが，ドラマ 1 本の映像長は 1 時間未満の場合が多く，2fps でフレームのシーケンスを抽出しているため十分高速に計算可能である.

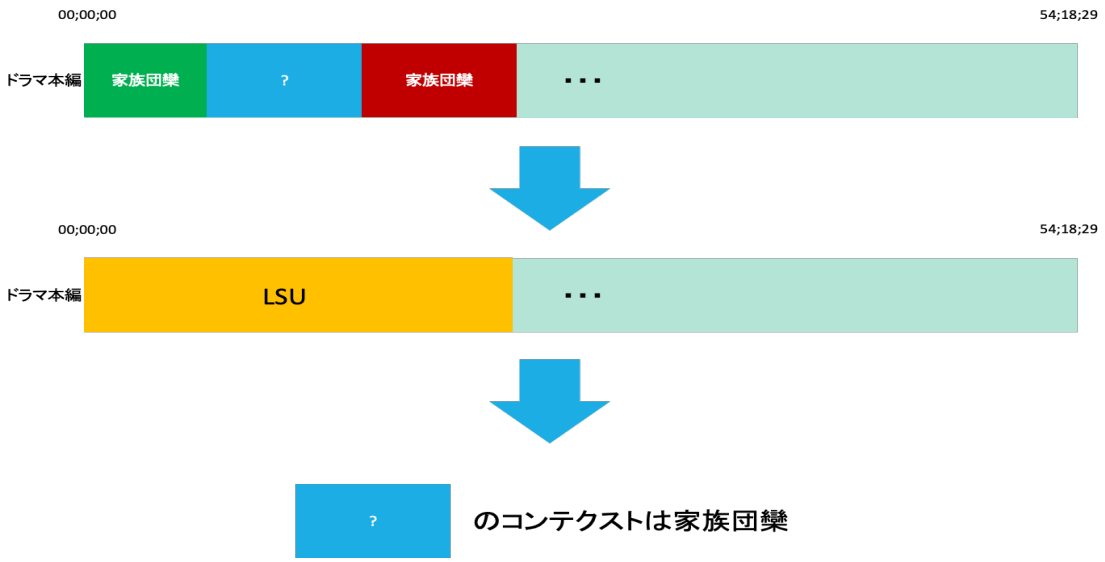

図 4 .6: LSU によるコンテクストの補完のイメージ

4.6　本章のまとめ

本章では提案手法について述べた．具体的には，まず手法の概要について説明した後，動画のセグメンテーション，コンテクスト特定のための辞書の作成，４要素の検出，LSU によるコンテクストに補完について述べた．

第５章　実験

本章では実験を行った環境，性能評価のための評価指標について述べる．

5.1　実験環境

本研究では，表４に示したマシン，表５.１と表 5.2 に示した開発環境で実験を行った．なおプログラミング環境は Google colab を用いて構築した．

表　５.1: 実験環境

OS	macOS Big Sur
CPU	2.7 GHz デュアルコア Intel Core i5
RAM	8.0GB

表　５.2: 使用したライブラリ

開発環境	バージョン
Python	3.7.12
NumPy	1.19.5
Pandas	1.1.5
face-detection	1.0.5
Mtcnn	0.1.1
OpenCV	4.1.2
yolov5	
Pytorch	1.7.0
Pytorchvideo	0.1.3

5.2　使用したデータセット

実験には３章で述べたデータセットを使用する．

5.3　実験手順

まず，LSU 単位でセグメンテーションされた各動画に対して正解のコンテクストを３章で説明したデータセットから割り振る．次に本システムの予測したコンテクストを各 LSU に割り振る．この処理は各シーン動画に対して本システムが割り振ったコンテクストを対応する LSU に割り振るものである．そして４．３の辞書の章で述べたキーワードのみで特定できるコンテクストと感情，動作，映っているもので特定するコンテクストに含まれない，つまり本手法では特定できないコンテクストを正解のコンテクストから削除する．最後に正解のコンテクストとシステムの出力を比較し，LSU の総数中いくつのコンテクストを正解することができるかで精度を算出する．つまり精度は以下の数式で表される．

$$精度 = \frac{正解と一致したコンテクスト数}{LSUの総数}$$

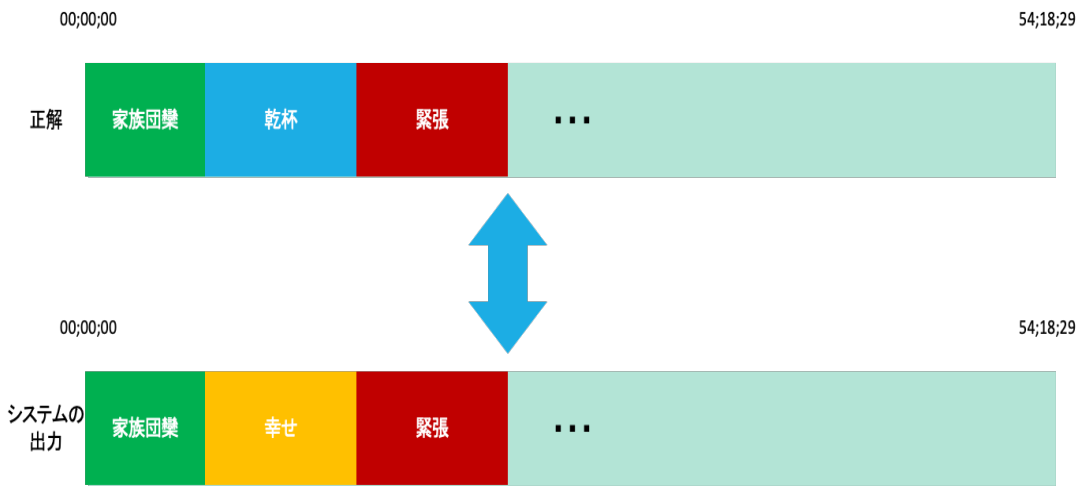

図　5.1: 性能評価のイメージ

5.4　本章のまとめ

本章では，本研究で用いた実験環境，データセット，実験内容と評価指標について説明した．

第６章　結果と考察

本章では実験とその結果を受けて考察を述べる．まずシステム適用結果の概要について説明した後に，各ドラマそれぞれへのシステム適用結果について考察する．

6.1　システム適用結果の概要

本節では４つのドラマへのシステム適用結果を示す．表６は 4 つのドラマへのシステム適用結果である．

表 6.1: システム適用結果

ドラマ名	精度
マジで航海してます	0.676
グッドドクター	0.803
あなたの番です	0.736
サバイバルウエディング	0.73
平均	0.736

コンテクスト特定精度は,マジで航海してますが 0.676,グッドドクターが 0.803,あなたの番ですが 0.736,サバイバルウエディングが 0.730 となった．
また 4 つのドラマのジャンルが，マジで航海してますはコメディ，グッドドクターは医療，あなたの番ですはサスペンス，サバイバルウエディングはラブコメディと幅広いことと，表６に示したコンテクスト特定精度の平均が 0.736 であることから本手法の汎用性が示せたと言える．

6.2　各ドラマの結果についての考察

以下では各ドラマの結果について考察を述べていく.

6.2.1　マジで航海してます

表 6.2 はマジで航海してますにおいてシステムが特定することができたコンテクストとその数をまとめたものである.

表 6.2: マジで航海してますにおいて特定に成功したコンテクストとその数

コンテクスト	個数
楽しい	3
幸せ	3
ダサい	1
緊張	12
不健康	2
贅沢	2
早い	2
汚い	3
健康	2
感謝	7
簡単	1
美味しい	1
飲む	1
空腹	1
リラックス	3

他のドラマに比べて精度がやや低くなった理由としては本ドラマがジャンルとしてはコメディであり，ドラマ内で多く見られる演出方法として台詞の言い回しによって笑わせるという方法をとっているという点が挙げられる．本手法におけるキーワードの検出では台詞の文脈を考慮することはできないため，言葉で表現されたコンテクストを特定しきれていない.

6.2.2　グッドドクター

表 6.3 はグッドドクターにおいてシステムが特定することができたコンテクストとその数をまとめたものである.

表 6.3: グッドドクターにおいて特定に成功したコンテクストとその数

コンテクスト	個数
緊張	13
安心	3
遅い	4
感謝	4
リラックス	1
不健康	5
早い	1
難しい	1

本ドラマは医療ドラマであるため，手術のシーンなど緊迫感あふれるシーンが多い．精度が平均を上回った理由としても数ある緊迫感あふれるシーンを特定することができたことが挙げられる.

6.2.3　あなたの番です

表 6.4 はあなたの番ですにおいてシステムが特定することができたコンテクストとその数をまとめたものである.

表 6.4: あなたの番ですにおいて特定に成功したコンテクストとその数

コンテクスト	個数
緊張	29
幸せ	11
感謝	7
遅い	2
清潔	2
難しい	1

表 6.4: あなたの番ですにおいて特定に成功したコンテクストとその数

健康	2
空腹	1

本ドラマのジャンルはサスペンスであり，グッドドクターと同じく緊迫感あふれるシーンが多い.本ドラマではBGMを使った演出が用いられたシーンが多く，特定しきれなかったコンテクストとしては和やかな BGM が流れた幸せを表現したシーンなどが挙げられる.

6.2.4 サバイバルウエディング

表 6.5 はサバイバルウエディングにおいてシステムが特定することができたコンテクストとその数をまとめたものである.

表 6.5: サバイバルウエディングにおいて特定に成功したコンテクストとその数

コンテクスト	個数
感謝	2
一生懸命	6
早い	1
簡単	2
健康	3
乾杯	3
お得	1
楽しい	1
美味しい	5

本ドラマのジャンルはラブコメディであり，大勢で楽しく飲食をするシーンがあるのが特徴である.本手法では乾杯として特定されており，酒等の飲み物のコマーシャル挿入に適していると考えられる.

6.3　コンテクストの特徴による本手法の適否

本節では前節で述べた各ドラマにおけるコンテクスト特定結果から考えられる本手法が特定しやすいコンテクストと特定しにくいコンテクストの特徴を述べる．本手法では，第４章で述べたようにキーワードのみで特定できるコンテクスト以外のコンテクストは感情，動作，映っているものの３要素でコンテクストの検出を行なっている．そのため当然ながら感情，動作，映っているものからシーンの特徴を検出できなければコンテクストの検出も不可能である．よってグッドドクターやあなたの番ですにおける表情から緊張が読み取れる場合，サバイバルドクターでの乾杯のシーン，マジで航海してますにおける食べ物が映っているシーンなどの映像内にコンテクストに直結する３要素の特徴が多く含まれる場合は精度が高くなる傾向がある．しかし顔が映っていないために感情が検知できなかったり，BGM によって緊迫感を表現していたり，台詞の言い回しでコンテクストを表現していたり等の映像の特徴がコンテクストに直結しないシーンが多く含まれる場合は精度が低くなりやすい．また緊迫感あふれるシーンの後に登場人物の家の外観等を見せることで日常の幸せを表現する場合などの前後のシーンと合わせてコンテクストの表現を行なっている場合も特定が難しい傾向にある．

6.4　本章のまとめ

本章では前章で述べた実験の結果を示し，結果に対する考察を述べた．また各ドラマにおけるコンテクストの特定結果から見られる本手法の特徴についても述べた．

第７章　結論と今後の展望

本章では，本研究のまとめ及び今後の展望について述べる．

7.1　まとめ

本研究ではドラマの動画に対して自動でどのようなシーンなのか，コンテクストを特定する研究を行った．提案手法ではキーワード，感情，動作，映っているものの４要素にコンテクストを分解して各要素を検出することにより，先行研究のように台本等のメタデータを利用することなくコンテクストを特定することを可能にした．様々なジャンルのドラマでコンテクストの特定精度の検証を行ったところ，平均の精度は 0.736 を達成しシステムの有用性を示すことができた．また各ドラマへのシステム適用結果からは，本システムの特徴として４要素の検出精度にコンテクストの特定精度が左右されることが分かった．本手法を用いることによりテレビ・インターネットの双方においてユーザーの情報を使うことなくターゲティングを行うことができる．

7.2　今後の展望

6.３節で述べたように本手法では BGM と台詞の文脈を考慮することができていない．BGM についてはコードの推定から BGM の曲調を特定することが考えられる．例えば緊迫感を醸し出す BGM では G の音を基音とするコード進行が用いられており，このコード進行の検出は BGM によるコンテクスト特定に直結する可能性がある．また台詞の文脈については Bert 等の言語モデルを使用することが考えられる．音声認識の今以上の精度向上が現状では不可欠と考えられるが，正確に音声を認識できるならば言語モデルによる文脈の考慮は，今まで特定できなかった前後のシーンを考慮したコンテクストの特定を可能にする可能性がある．

参考文献

[1] Tsvetomira Tsoneva, Mauro Barbieri, and Hans Weda. Automated summarization of narrative video on a semantic level. In *International conference on semantic computing (ICSC 2007)*, pages 169–176. IEEE, 2007.

[2]T. Yao, T. Mei and Y. Rui, "Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 982-990, doi: 10.1109/CVPR.2016.112.

[3] https://deepmind.com/research/open-source/kinetics

アクセス日 1 月 13 日

[4] https://cocodataset.org/#home

アクセス日 1 月 13 日

[5] https://github.com/Uberi/speech_recognition

アクセス日 1 月 13 日

[6] C. Feichtenhofer, H. Fan, J. Malik and K. He, "SlowFast Networks for Video Recognition," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6201-6210, doi: 10.1109/ICCV.2019.00630.

[7] https://github.com/serengil/retinaface

アクセス日 1 月 13 日

[8] https://github.com/gitliber/FaceEmotionRecognizer

アクセス日 1 月 13 日

[9] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," in IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[10] https://pytorch.org/hub/ultralytics_yolov5/

アクセス日 1 月 13 日

[11]大塚優斗．メタデータを必要としないドラマ要約映像の自動生成．
Master's thesis，東京理科大学 理工学研究科 経営工学専攻，2020.