SNS に適した調理要約動画の半自動作成システムに関する研究

TP18008 祖父江亮　　指導教授：山下隆義

1. はじめに

ソーシャルネットワーキングサービス (SNS) に投稿される調理レシピ動画 (要約動画) は，工程や出来上がりを簡単・短時間に把握できるため，投稿数が年々増加している．一方で，要約動画の作成は負担が大きいため，ユーザが手軽に作成できる半自動のレシピ動画生成システムが提案されている．このシステムでは，調理レシピサイトの手順画像と撮影した動画から，画像特徴量の類似度比較によって各調理工程に適したシーンを検索する．しかし，カメラの設置位置により，調理に関係ない物体が映りこみ，調理部分の画像特徴が捉えられず，適切なシーンを検索することが困難となる場合がある．

本研究では，レシピに付属する説明文や動画中の物体および動作情報を利用することにより，システムの利便性を向上させる．また，システムの利便性を客観評価できるように新たな評価指標を提案する．評価実験では，提案した評価指標によって求めたスコアを比較することで，調理工程に適したシーンの検索精度が向上することを示す．

2. 調理レシピ動画作成システム

調理手順に沿った調理レシピの要約動画を作成するには，調理レシピの各調理工程に類似したシーンを撮影した動画から検索して動画クリップとし，各シーンのクリップを結合する必要がある．本システムでは，手順画像に類似したシーンの候補をユーザに提示し，ユーザの好みのフレームを選択することで手軽に短時間でレシピ動画の要約を可能とする．

2.1 類似シーン候補の抽出

本システムは，レシピサイトの手順画像と類似する動画フレームを選択して要約動画のシーンとする．まず，手順画像と動画の各フレームに対して，InceptionV3[2] の Mixed5 層から，128 次元の特徴ベクトルを抽出する．その後，各手順画像の特徴ベクトルとの類似度をユークリッド距離で算出し，各上位 5 フレームをそれぞれ調理工程のシーン候補とする．

ここでは，料理画像に特化して画像間の類似度をより正確に算出するために，Triplet loss[1] を用いて InceptionV3 の学習を行っている．手順画像を anchor，手順画像と同一シーンの動画フレームを positive，手順画像と異なるシーンの動画フレームを negative としてネットワークに入力する．そして，手順画像と類似する動画フレームの特徴ベクトルが近くなるように学習させ，入力画像に映る調理物体やシーンに応じて異なる特徴ベクトルがになるように特徴抽出器を学習している．

2.2 要約動画作成手順

前節で述べた特徴抽出器を用いたシステムのインターフェースを図 1 に示す．各手順画像に対して，類似度の高いフレームを含む数秒間の動画クリップを候補として 5 つ提示し，ユーザに選択してもらう．また，ユーザによる操作で候補シーンのフレーム再生時刻の微調整や，各シーンへの字幕の挿入も可能である．

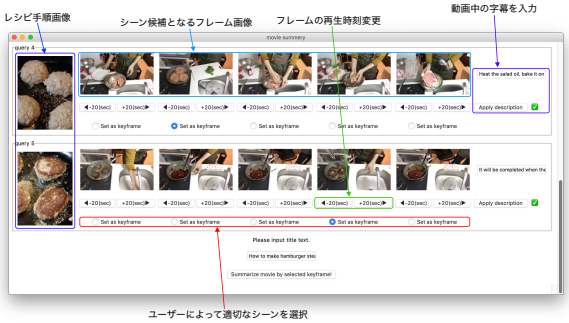


図 1： システムのインターフェース

3. 類似シーン検索精度の向上

本システムでは，レシピ画像と撮影動画の視点が異なる際の類似シーン抽出精度が低下する問題がある．その原因の 1 つとして，調理に関係ない物体が映ることが挙げられる．これにより，類似シーン候補の検索に用いる特徴ベクトルが適切に得られず，調理工程とは異なるシーンが抽出される．そこで本研究では，レシピ説明文の活用や調理領域の検出によって類似シーン検索の精度を向上させる．また，新たに定義する評価指標により，システムが提案するシーンを定量的に評価する．

3.1 レシピに付属する説明文の活用

レシピサイトにある調理手順画像には，各手順に対する説明文がついている．そのため，動画中にある物体ラベルと同様の単語が説明文中に含まれる場合，そのフレームは手順に適したシーンであると期待できる．そこで，入力動画フレームから調理に関連する物体を検出し，検出した物体とレシピ説明文中の単語を照合する．照合の流れを図 2 に示す．

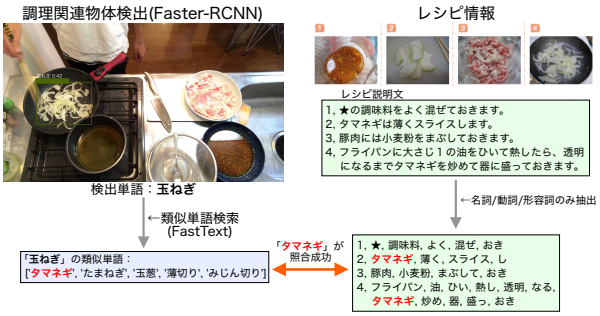


図 2： 物体検出結果とレシピ説明文との照合の流れ

1) 調理関連物体の検出

まず，調理画像に対する物体検出データセットとして，EPIC-KITCHEN dataset[3] を用い，Faster-RCNN により調理関連物体の検出を行う．検出する物体のクラス数は 295 クラス，学習に使用した画像は 173,672 枚，バウンディングボックスの個数は 274,701 個である．

2) 調理関連単語の類似単語検索

レシピサイトの説明文は，レシピ情報を投稿するユーザによって表現方法が異なる．本手法では，検出した物体ラベルに対して類似単語の検索を行い，レシピ説明文中の単語と照合する際の表記ゆれによる照合失敗を抑制する．

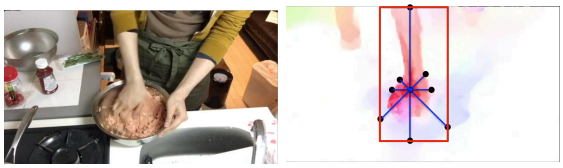
ここでは，類似単語検索を行うため，FastText によって単語の分散表現を獲得する．FastText は同様の手法である Word2Vec と異なり，品詞の活用形による表記ゆれを吸収することが可能である．本研究では，楽天レシピのレシピ説明文 3,048,617 文，40,067 種の単語を学習データとして用いている．またレシピ説明文には，「これら」といった代名詞や「中火」といった火加減の説明，「混ぜる」といった動詞など，複数の手順において何度も出現する単語がある．そのため，レシピ説明文の頻出 100 単語と名詞以外の単語を照合の対象から除外し，照合対象の絞り込みを行う．

1) と 2) を用いてレシピ説明文との照合に成功したフレーム画像が，システムによって提示される．

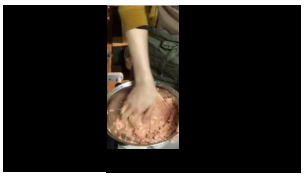
3.2 オプティカルフローによる調理領域の検出

ユーザの撮影した調理動画には調理に関係ない背景が多く映り込み，動画フレーム中の調理領域が著しく小さくなることがある．そのため，調理領域では動きが発生していると仮定し，オプティカルフローによる調理領域の検出を行う．そして，画像全体の代わりに調理領域を利用することで，提案精度を向上させる．

調理領域を検出する手順を図 3 に示す．まず，隣接動画フレーム間のオプティカルフローを算出する．ここでは，密なオプティカルフローを算出する手法である Gunnar Farnelback 法を用いる．その後，算出したフローの中で，最も動きが大きい画像内の点から 8 方向にフローの大きさが閾値より小さくなるまで点を延ばす．なお本手法では，フローの大きさを示すマグニチュードの閾値を 0.02 とする．最後に，8 点を囲む最小の矩形を作成し，これを調理領域とする．そして，動画フレームから調理領域以外を黒くマスキ



(a) フレーム画像　(b) 調理領域の矩形作成例



(c) 調理領域

図 3：調理領域の検出例

ングすることで図 3(c) の調理画像を作成する．

3.3 シーン検索精度の評価指標

従来の動画要約の評価方法は，要約された動画をユーザに見せアンケートに答えてもらう定量的な評価が主流となっている．そのため，年齢層や性別などのアンケートの回答者の属性が偏ることや，質問文などからユーザにバイアスが生じることにより，適切な評価ができないことが考えられる．

本研究では，システムが手順画像に対して適切なシーンを提示できているかを定量的に評価するための，評価関数を定義する．式中の $S_{candidate}$ は，システムが提示する候補ごとに算出する．なお，gt_{start}，gt_{end} は，各手順における正解範囲となる動画フレームの開始時刻と終了時刻である．システムから提示された動画フレームの時刻 t_p が $gt_{start} \leq t_p \leq gt_{end}$ となる時，スコアは 1 となる．一方で，提示されたフレーム時刻の正解範囲から離れていくにつれ，スコアは 0 に近づき，正解範囲から a 離れた際に，スコアは 0 となる．また，$S_{suggestion}$ は，$S_{candidate}$ の最大値を手順ごとに求め，全ての手順の平均を取った値である．

$$S_{candidate} = \begin{cases} 1.0 & if\ gt_{start} \leq t_p \leq gt_{end} \\ \max(1.0 - \frac{d}{a},\ 0) & otherwise \end{cases} \tag{1}$$

$$d = \begin{cases} gt_{start} - t_p & if\ t_p < gt_{start} \\ t_p - gt_{end} & if\ gt_{end} < t_p \end{cases} \tag{2}$$

$$S_{suggestion} = \frac{1}{N} \sum^{N} \max(S_{candidate}) \tag{3}$$

ここで，N は手順の数である．動画を 30 フレーム/秒として，提案したフレーム画像の時刻が正解範囲から $a/30$ 秒離れると $S_{candidate}$ は 0 となる．

4. 評価実験

類似シーンの提示精度を向上させる手法の有効性を確かめるため，$S_{suggestion}$ を比較することによって評価実験を行う．評価には 8 種類のレシピをもとに調理した 25 本の動画を用いる．なお，1 つの料理における手順数は最大 13 手順，最小 6 手順であり，$a = 3000$ とする．

ここでは，物体検出とレシピ説明文の照合について，改善手法のあり，なしで比較する．なお，改善手法ありの手法では，追加でレシピ説明文と照合させる対象の単語を増減させた場合や，絞り込みを行った場合のスコアも算出する．また，いずれの比較手法についても，入力画像に動画フレーム全体を用いた場合，調理領域のみを用いた場合のスコアをそれぞれ算出する．

4.1 評価結果

従来システム，および提案手法を導入したシステムによるスコアの比較結果を表 1 に示す．調理領域を入力画像とした場合，動画フレーム全体で入力した場合に比べて，提案スコアが向上した．また，レシピ説明文と物体検出による照合を行なった場合，照合させる単語の対象を絞り込むことで，より高いスコアを獲得した．従来システムはスコアが 0.5527 であるが，提案手法を導入するとシステムの $S_{suggestion}$ が最大で約 0.04 ポイント向上した．また，図 4 に示す導入前と導入後のシステムのインターフェースの例でも，手順画像により適した動画クリップが提示されたことを確認した．

表 1：提案スコアの比較

物体検出結果と説明文による照合	入力画像	
	全体	調理領域
なし (従来システム)	0.5527	0.5771
あり (類似 10 語, 対象絞り込み)	0.5653	**0.5935**
あり (類似 10 語)	0.5591	0.5737
あり (類似 2 語)	**0.5656**	0.5699
あり (類似 0 語)	0.5455	0.5608

動画フレーム全体および調理領域に対して物体検出を行った例を図 5 に示す．上の例では動画フレーム全体を入力した場合，画面の端にある蓋を検出してしまい，誤った手順との照合を起こす可能性がある．しかし，調理領域を物体検出の対象にすることで，調理領域外にある蓋は検出されないことから，誤照合を防止できる．また，下の例では動画フレーム全体を入力した場合に検出できなかった物体を，調理領域の画像を用いて検出，レシピ説明文と照合している．これにより，2 つの手法を組み合わせることで，シーン提示により良い影響を与えることを示した．

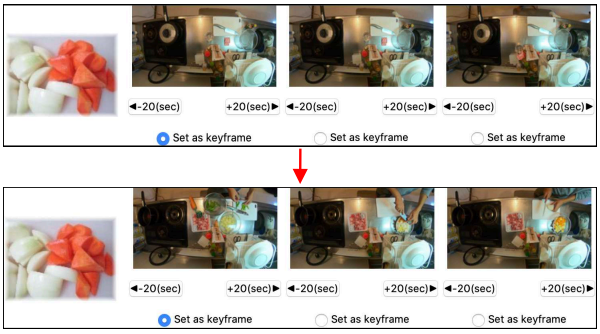


図 4：改善手法によるシーン提案の改善例

(a) フレーム画像　(b) 調理領域

図 5：フレーム画像と調理領域の物体検出結果例

5. おわりに

本研究では，システムの提示精度を向上させることを目的とし，レシピの説明文や動画中のフローを用いるシーン提示の改善を提案した．また，シーンに適した動画フレームが提示されているかを定量的に評価する評価関数を定義した．

評価実験では，システムが提示したフレームをもとに評価関数によりスコアを算出し，調理領域の検出や物体検出，レシピ説明文の活用によってシステムの提示精度が向上することを示した．今後の課題としては，物体検出や類似単語検索の高精度化，また特徴抽出器の再学習による，システムの更なるシーンの提示精度の向上などが挙げられる．

参考文献

[1] F. Schroff, *et al.*, “Facenet: A unified embedding for face recognition and clustering.”, CVPR, 2015.

[2] C. Szegedy, *et al.*, “Rethinking the Inception Architecture for Computer Vision.”, CVPR, 2016.

[3] D. Damen, *et al.*, “Scaling Egocentric Vision: The EPIC-KITCHENS Dataset”, ECCV, 2018

研究業績

[1] 祖父江 亮 等. “ソーシャルネットワーキングサービスに適した料理レシピ動画の生成”, WISS, 2018

[2] R. Sobue, *et al.*, “Cooking Video Summarization Guided By Matching with Step-By-Step Recipe Photos”, MVA, 2019.