Triplet Loss を用いた調理レシピ動画要約におけるキーフレーム抽出

EP14074 祖父江亮　　指導教授：山下隆義

1. はじめに

調理動画の要約は，調理の詳細な手順を効率よく把握するために有用である．しかしながら，大量の調理要約動画を手動で作成するのは人的コストがかかる．そこでレシピサイトなどの手順画像を利用して自動で動画を要約する方法が考えられる．本研究では Deep Convolutional Neural Network(DCNN) で出力される特徴ベクトルを用いて動画要約に適したキーフレームを選択する方法を提案する．

2. DCNN による特徴量抽出

図 1 に示すように，DCNN を学習する際，基準画像 (anchor) と同じクラスの画像 (positive) はユークリッド距離が小さく，異なるクラス (negative) の画像は大きくなるような特徴ベクトルを抽出するための誤差関数として Triplet Loss[1] がある．Triplet Loss を用いて DCNN を学習することで，画像同士の類似度を上げることができるため，人物同定に用いられている [2].

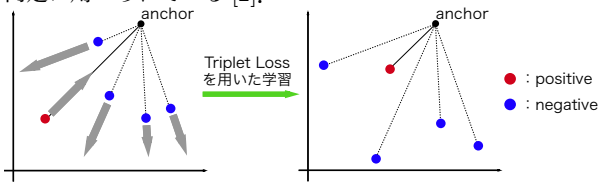

図 1：Triplet Loss の学習方法

3. 提案手法

キーフレーム抽出による動画生成モデルを図 2 に示す．入力動画は要約させる動画，レシピ手順画像は入力動画と同じ料理名のレシピの手順画像である．

まず，入力動画フレーム及びレシピ手順画像を DCNN に入力して画像ごとの特徴ベクトルを得る．DCNN のネットワーク構造には，InceptionV3 を用いる．

ネットワークの学習時に出力される中間層の特徴ベクトルが同一調理手順の場合は近く，異なる調理手順の場合は遠くなるように Triplet Loss を用いて学習する．本手法では，手順画像を anchor とし，同一調理手順のフレーム画像を positive，異なる調理手順のフレーム画像を negative とする．これにより，同一の調理手順で似た特徴ベクトルを抽出し，調理動画の最適なキーフレーム選択が可能となる．

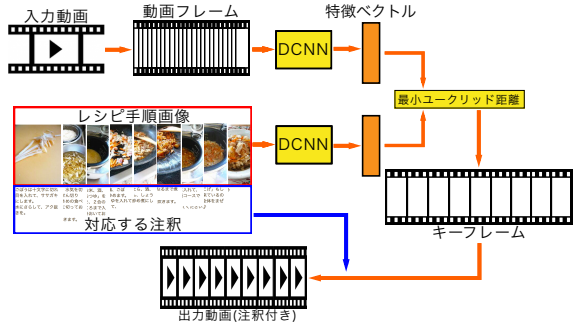

図 2：動画要約生成モデル

推論時は入力動画の各フレームに対する特徴ベクトルを抽出し，手順画像の特徴ベクトルと最もユークリッド距離が近いフレームをキーフレームとして選択する．これを全ての手順画像について行う．そして，キーフレーム毎にキーフレームとその前後のフレームから動画クリップを作成し，動画クリップを連結させて要約動画を作成する．

4. 評価実験

提案手法によるキーフレーム抽出の精度を評価する．

4.1. 実験概要

本実験では，手順画像として 3 枚から 13 枚の手順画像が含まれるレシピ 115 個を楽天レシピから，動画 9 本を動画投稿サイトから取得し，表 1 のデータセットを構築した．料理の種類は 9 種類である．各手順画像を anchor とした時の positive は，同料理名の動画のフレームから指定する．手順画像とキーフレームのバリエーションを増やすため，指定した positive フレームの前後 30 フレームのうち，6 フレームごとに追加の positive を選び，ランダムに anchor と組み合わせて学習する．

テストはハンバーグの調理を対象として評価を行う．手順画像ごとに動画からキーフレームを正しく選択した正答率を求める．また，比較手法には Triplet Loss による学習を行っていない GoogLeNet を用いる．GoogLeNet は一般物体検出に用いられるネットワークである．

表 1：データセットの手順画像及び動画数

train		test	
手順画像	動画	手順画像	動画
4573	10	18	1

4.2. 実験結果

各評価の正答率を表 2 に示す．表 2 により提案手法が比較手法の正答率を約 27.8% 上回り，キーフレームの抽出において Triplet Loss による学習が有効であることがわかる．

表 2：提案手法と比較手法による正答率

	提案手法	GoogLeNet
正答率 (%)	**61.1**	33.3

ハンバーグを作成する際の 5 つの手順において，選択されたキーフレームを図 3 に示す．左から 1 番目，2 番目，3 番目，5 番目の手順画像は手順画像と同一の調理手順のキーフレームを選択しているが，4 番目は調理手順が異なるキーフレームを選択している．

表 2 より，入力動画は料理全体を撮影したものが多く，他の物体が映り込んでしまうことが考えられるため，視点変化や大きさへの対応をすることによってさらなる精度向上が期待できる．また，さらなる学習サンプルの追加，positive の選択方法の改善などが必要である．

また，評価実験によって選択されたキーフレームを元に作成した要約動画の一部を図 4 に示す．動画内には左下に字幕が付加されているが，字幕の内容と動画の内容に差異があることが確認できる．これはレシピサイトの注釈をそのまま用いているためであり，動画内容に忠実な注釈の作成を実現することは今後の課題である．

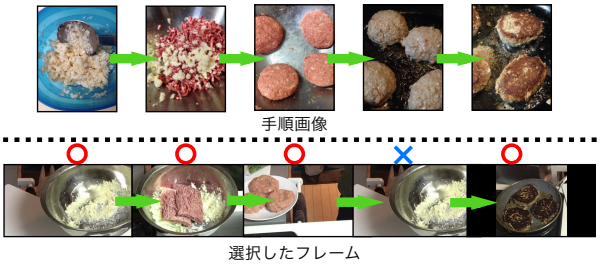

図 3：選択したキーフレームの例

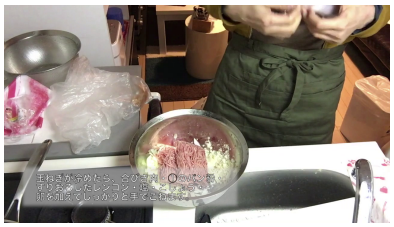

図 4：評価実験で作成した要約動画

5. おわりに

本研究では，Triplet Loss を用いた DCNN による調理レシピ動画の要約作成モデルを提案した．今後は動画内の調理部分の把握による正答率向上や，学習サンプルの改善，画像の情報から動画要約に付加させる注釈の生成を行う．

参考文献

[1] F. Schroff, *et al.*, “Facenet: A unified embedding for face recognition and clustering”, CVPR, pp. 815-823, 2015.

[2] Z. Zhou, *et al.*, “See the Forest for the Trees: Joint Spatial and Temporal Recurrent Neural Networks for Video-based Person Re-identification”, CVPR, pp. 4747-4756, 2017.