ISS-P-058

作業工程の言語情報を利用した作業映像のシーン分割

林　純平†　　古田　諒佑††　　谷口　行信†
† 東京理科大学　　†† 東京大学

1. はじめに

製造業や飲食店といった業種において，効率よく技術や技能を伝承する手段として，作業の様子を撮影した映像を用いることがある．しかし，作業工程が多い場合には映像が長くなってしまうため，技能伝承の効率が下がってしまう．非熟練者が見たいシーンをすぐに検索できれば効率が上がるが，そのためにはあらかじめ作業映像を作業工程のシーンごとに分割する必要がある．

一般の映像に比べ，作業映像は見えの変化が少ないため，映像情報のみを用いる従来の手法[1][2]では適切にシーン分割できないことがある．そこで本稿では，マニュアルなど作業工程についての言語情報を利用した作業映像のシーン分割手法を検討する．

2. 画像認識と自然言語処理の融合

本稿では，フレーム画像と作業工程の言語情報に対して Cross-View Triplet Loss[3]を用いた距離学習を行うことで，言語情報を考慮してシーン分割する方法を提案する．

作業映像と作業工程について，1つの作業工程には必ず連続となる複数フレームから構成されるシーンが対応し，シーンに含まれないフレームは存在しないことを前提とする．この前提下では，作業映像を作業工程ごとに分割することは，工程のキャプションとフレームの1対多マッチングと等価である．したがって，フレームとキャプションから抽出した特徴量を同一空間に埋め込むことができれば，特徴量を重みとした2部グラフの最大マッチングとしてシーン分割を実現できる．

3. 実験

図1に示すように，フレームとキャプションから ResNet50 と BERT を用いて特徴量を抽出し，全結合層により埋め込みを行った．大規模データセットである MSCOCO2017 を用いて特徴量埋め込みの事前学習を行い，次に作業映像データセットである YouCook2 の学習データ(約57時間分)を用いて同様に学習した．

YouCook2 の検証データ(約20時間分)について特徴量を埋め込み，フレームとキャプションのマッチングを行った．マッチングの手法として，①1つのフレームに複数のキャプションがマッチしないよう1対多に制約した DP マッチング，②各フレームについて最近傍の特徴量を持つキャプションを探索するマッチング，の2種類を検証した．映像を工程数で等分するものを③線形分割としてベースラインに定め，2種類のマッチングと比較した．

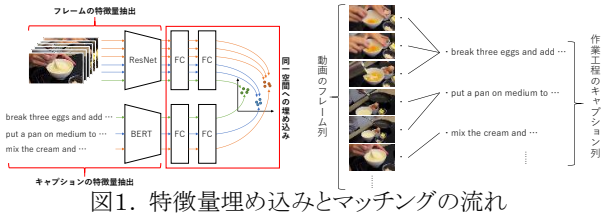

図1．特徴量埋め込みとマッチングの流れ

表1．手法ごとの精度の比較

DP マッチング	最近傍探索	線形分割
0.362	0.173	0.419

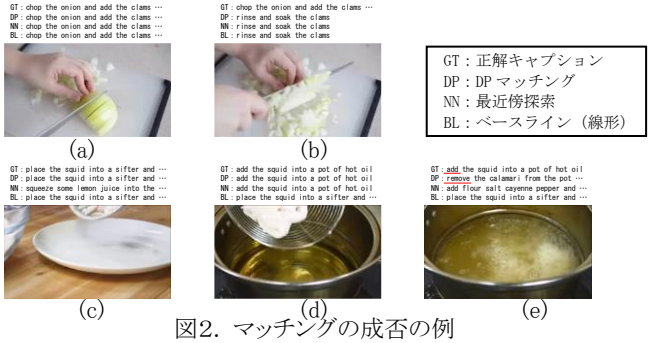

図2．マッチングの成否の例

正しいキャプションが対応するフレームの割合を精度として評価し，各手法を比較したものが表1である．DP マッチングは最近傍探索によるマッチングより精度が高かったが，ベースラインを上回らなかった．

被写体やキャプションの特徴により，マッチングの成否が分かれた．図2(a)のように，画像に物体がはっきり写っている場合や，動作がわかりやすい場合は適切にマッチしやすい一方で，図2(b)のように，物体が原形をとどめていない場合には適切にマッチできない傾向にあった．また，図2(c)(d)(e)のように，1フレームの画像のみからでは動作を判断できない場合に，誤ったキャプションをマッチすることが多かった．以上より，時系列情報を特徴量に含むことができれば，精度が改善すると考えられる．

5. まとめ

実験の結果，ベースラインとして設定した線形分割の精度を上回らなかったため，手法の改良が必要である．今後の課題として，時系列情報を考慮できるモデルを用いた実験を行うことがあげられる．

参考文献

[1] J. Mas & G. Fernandez, “Video Shot Boundary Detection Based on Color Histogram”, TRECVID, 2003.

[2] W. Tong, et al., “CNN-Based Shot Boundary Detection and Video Annotation”, IEEE International Symposium on BMSB, 2015.

[3] X. Gu, et al., “Multi-Modal and Multi-Domain Embedding Learning for Fashion Retrieval and Analysis”, IEEE Transactions on Multimedia 2018.