テレビ番組の字幕画面抽出と字幕認識 -精度改善と全自動化-

Extraction of caption images and recognition of captions in TV program
–Accuracy improvement and full automation –

D－11

宮崎 晶斗
Akito MIYAZAKI
芝浦工業大学
Shibaura Institute of Technology

高橋 正信
Masanobu TAKAHASHI
システム理工学部
Collage of Systems Engineering Science

1. 背景

多くのテレビ番組が地上波で放送されているが，視たくても時間が無く断念することがある．視るよりも時間をかけずに番組内容を把握できれば，視聴者にとって有益である．そこで，番組内容を短時間で把握するために，「字幕画面のみ」を利用者が視られるようにする機能の実現を考えた．字幕画面を視ることで短い時間でセリフを把握することができ，画像から雰囲気を把握することもできる．

これまでの研究[1]で字幕画面抽出機能を実現している．しかし，DVD にダビングした字幕あり，なし画面をそれぞれ撮影する必要があり，DVD の入れ替えを手動で行う必要があった．また，字幕あり，なし画像の差分画像を閾値処理して字幕領域を抽出しているが，抽出が不十分な場合があり，CM 画面の誤抽出が多くなるなどの問題があった．

2. 目的

本研究では，字幕画面抽出の精度を改善するとともに，DVD の入れ替えを不要とし，処理を自動化することを目的とする．また，字幕画面から字幕文字を認識し，発話者の情報とともにテキスト化する機能の実現も目的とした．

3. 手法

3.1 撮影

HD から DVD に字幕あり映像をダビングする．この作業のみ利用者が行う必要がある．DVD を再生プレーヤー(PowerDVD)で再生した画面を 1 秒間隔で自動的にスナップショットする．字幕あり画面から深層学習で字幕領域を認識することで，字幕なし画面の撮影を不要とした．

3.2 字幕画面の抽出

字幕あり画面から字幕領域を抽出する深層学習ネットワークにはセマンティックセグメンテーション用のモデルである U-Net を採用した．U-Net は撮影した画像(図 1(a))の入力に対して，字幕領域が白となる画像(図 1(b))を出力するように学習した．

(a)入力画像　　(b)出力画像

図 1　字幕領域抽出結果[2]

次に，入力画像中の字幕領域の部分(字幕画像)をグレースケール化し，閾値 127 で 2 値化することで字幕文字を抽出した画像(字幕文字画像)を作成する．なお，字幕文字の色(白，青，緑，黄)のうち緑については閾値処理で良好に 2 値化できなかったため，画素の色度が緑に相当する場合はグレースケール化せずに画素を白として字幕文字を抽出する．

次に，連続する字幕画面において字幕文字画像の差分が小さいものをグループ化し，そのうちの 1 つを最終的な字幕画面として抽出する．

3.3 テキスト化と字幕色認識

抽出した字幕画面の字幕画像中の文字は GoogleDrive の OCR 機能を用いて自動認識し，テキスト化する．文字色が緑の場合に誤認識が多いため，緑の場合は字幕文字が白，背景が黒の 2 値画像に変換してから文字認識する．

また，字幕文字の色は発話者ごとに異なるため，文字色を認識し，テキストと対応づけて保存することで発話者の情報も獲得する．文字色は白，青，黄，緑のうち，その色度の画素が字幕画像内に占める割合が最大の色とする．

4. 実験

テレビアニメ 7 番組 7 話からランダムに選択した 567 枚(CM73 枚含む)の画像の約 9 割を用いて U-Net を学習し，残り 1 割に対する評価が最も良かった学習回数のネットワークを採用した．精度評価には 4 番組 5 話の映像(CM を含む 138 分)を用いた．評価用 4 番組のうち 3 番組は学習に用いていない番組である．

抽出精度を表 1 に示す．実験の結果，全ての字幕画面を抽出することに成功した． CM を字幕画面と誤抽出する割合は従来の 12.7%から 0.53%へ大幅に低減できた．また，同じ字幕文字の字幕画面が合計で 117 枚誤抽出された．この誤抽出の主な原因としては，シーンが大きく変化して差分が大きくなった結果，グループ化を誤ったことが挙げられる．しかし，同じ字幕でもシーンが大きく異なるため，実用上は問題にはならないと考える．

表 1　字幕画面抽出精度

	再現率	CM 誤抽出率
従来研究[1]	100.0%(413/413)	12.7%(55/432)
本研究	100.0%(1932/1932)	0.53%(5/939)

テキスト化の評価指標には CER(文字誤り率)を用いた．CER は(挿入語数+置換語数+削除語数)/正解語数で計算され，0 に近いほど高精度といえる．実験の結果，CER は 0.074，文字色認識の誤り率は 1.2%となり，実用性の高い精度が得られたと考える．

5. まとめ

字幕あり映像を保存した DVD から字幕画面，および字幕文字の色とテキストを高精度で自動生成する機能を実現できた．今後の課題としては，CM の誤抽出のさらなる低減と字幕文字の抽出精度の改善が挙げられる．

[参考文献]

[1] 萩原，他:“テレビ番組の字幕画面抽出と字幕認識-精度改善-”，電子情報通信学会東京支部学生会研究発表会，115，2019.
[2] TBS，“まちカドまぞく”，第 7 話，2021 年 8 月 19 日放送.

Copyright © 2022 IEICE