5ZF-03

画像検索に基づく声優認識のための画像音声対収集支援システムの試作

鹿田 理央† 大囿 忠親‡ 新谷 虎松‡

†名古屋工業大学情報工学科 ‡名古屋工業大学大学院情報工学専攻

1 はじめに

映像作品中の登場人物の声優名を，その声を聞いた時に，即時かつ簡便に調べるためのツールが求められている．洋画やアニメのように，声優による吹き替えが用いられる場合，登場人物の映像からは声優を特定するための作業（エンドロールの確認など）が手間である．この手間を軽減することが可能ならば，利用者のみならず，声優や映像作品の作成者などのコンテンツ産業にとっても，新規ファンの発掘等のメリットがある．しかし，音声と声優名を紐付けるための公開されたデータベースが存在しないため，何らかの工夫が必要となる．本稿では，音声と声優名を紐付けるために，画像検索を利用することを想定し，そのために必要な画像と音声の対を効果的に収集するためのシステムの試作について述べる．

2 画像検索に基づく声優認識

本研究における声優認識とは，映像作品中の登場人物の画像（以降，画像）および登場人物の音声（以降，音声）から，その声優名を得ることである．

声優認識のためには，画像，音声，および声優名の組を蓄積することが必要である．音声と声優名を紐付けるための公開されたデータベースが存在しない．Web上には，画像と声優名を紐付けるための情報が存在する．映像作品中の登場人物を紹介するための情報が，非構造な状態で公開されている場合が多いからである．すなわち，Web上からは，画像と声優名の対を収集可能な状態であるが，音声と声優名の対に関しては，収集可能な状態にはない．

本研究では，音声からの声優認識において，画像検索を利用して，声優名を特定することにした．声優名を知りたいそのときに，映像作品中から得られるのは，画像および音声である．前述したように、画像と声優名の対を得ることが可能であるから，画像さえ得られれば声優認識が可能になる．しかし，声優名の対に関する情報が存在しないので，音声から声優名を特定するためには，音声から対応する画像を求め，それらの画像から声優名を特定することとした．

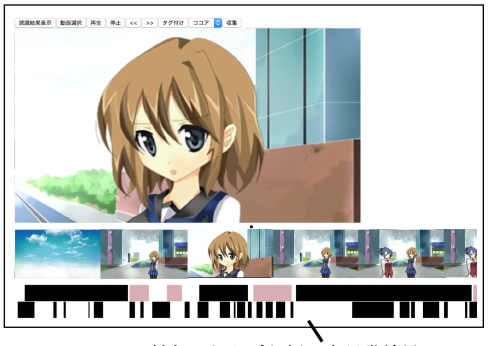

図 1: システム実行例

本研究では，画像，音声，および声優名の対を蓄積するために，画像と声優名の対，および画像と音声の対を蓄積することとした．画像と声優名の対に関しては，Web上での公開情報からの情報抽出により実現可能である．画像と音声の対に関しては，映像作品から新たに情報を抽出する必要があるため，画像と音声の対の取得に着目した．画像および音声の両方から声優名を特定することで，低品質の画像および音声からの声優認識が可能になると考えている．例えば，音声に関しては，BGMや効果音はノイズとして考えられる．

3 画像音声対収集支援システム

本研究における画像音声対とは，映像作品の声優認識のための話者認識器を作成するための学習用データである．本システムでは，映像作品から登場人物の顔画像と声を収集する．映像部分と声部分それぞれに対

Implementing a Image-Voice Pair Collection Support System for Voice Actress Recognition based on Image Retrieval

Rio SHIKADA†, Tadachika OZONO‡and Toramatsu SHINTANI‡

†Department of Computer Science, Nagota Institute of Technology. ‡Department of Computer Science, Graduate School of Engineering, Nagoya Institute of Technology.

してタグを付与し，タグ付け終了後，タグごとに音声と画像をまとめる．

本システムの実行例を図1に示す (動画，サムネイル部分は参考画像を利用 人物絵:「SUGAR STAR」http://sgst.x.fc2.com/sozai/sozai_base.htm 背景絵:「よく訓練された素材屋」(http://material.animehack.jp/index.html))．ユーザが動画を選択すると，初めて選択された動画であった場合，サーバにアップロードされる．以前選択された動画である場合，データベースに保存された動画に対するタグ情報を取り出す．タグ情報とは，以前に付与したタグ名，タグの開始，終了地点，音声，映像のどちらのタグかを示すフラグである．動画を15秒ごとに切りとり，サムネイル画像として動画の下部に表示する．また同時に，サーバで顔認識と声認識を行う．顔認識と声認識については次節で詳しく説明する．最後に，サーバから認識結果が返され，ユーザに提示される．サムネイルの下部の黒い帯部分が認識された不要部分である．

本システムでのタグ付け方法について説明する．サーバに動画のアップロードを行いクライアント部分に動画とサムネイル画像が表示されると操作が可能となる．本システムにはタグ付けボタンが存在し，ボタンをクリックするとクライアントの動画の現在の再生時間がタグの位置として記録される．1度目のクリックでタグの開始時間を，2度目のクリックで終了時間を記録し，その2度のクリックでタグ付けが完了となる．サムネイル下部の帯部分に付与されたタグが表示される．帯の色でタグが識別可能になる．

4 顔認識と声認識に基づく収集支援機能

現在，映像作品の多くは20分を超えており，さらに画像と音声の両方について収集する必要があるため，データ収集には時間と手間がかかる．そこで，本システムではユーザの負担を減らすために顔認識と声認識を行う．顔認識や声認識は映像中のどの部分に顔や声が存在するかを認識する機能である．ユーザは映像作品中の顔や声の存在しない部分を確認する必要がなくなるため，収集時間の短縮ができる．声の区間を検出する研究についてはすでになされている [1]．この研究は，レストランでの作業効率化を行うために，従業員の発話区間を推定し，従業員の現在行っている作業を推定することを目的としている．レストランではBGMがかかっていることも多く，また客の声なども含まれている中で発話区間の推定を行うため雑音処理は行われている．しかしこの研究では，従業員のそれぞれがマイクを持っており，ある程度の雑音については影響が無いレベルになっている．映像作品の音声は完全にBGMや効果音と登場人物の声が混ざっているため，雑音については取り除くことが困難である．そこで，本システムでは別の声認識の手法を用いる必要がある．

顔認識にはOpenCVを用いている．動画から画像を0.1秒ごとに切り出し，160×90pxにリサイズをし，各画像に対して顔認識が行われる．本システムで用いた顔認識では，前向きの顔しか検出されない．またリサイズにより解像度が低くなり，非常に小さい顔が認識されない．しかし，リサイズにより認識時間は4分の1以下になった．リサイズする前は画像の大きさが 640×360px であったが，この大きさでの認識時間は動画の時間の2倍程度かかってしまった．また，小さい顔については精度は下がったものの，解像度が下がることにより誤認識が減りリサイズをする利点の方が多いと考えられる．

声認識は，映像作品中の声はBGMや効果音よりも大きくないと聞こえなくなってしまうため，声の存在する部分はその周辺部分に比べて音声波形の振幅が大きくなるというヒューリスティクスを用いて，正の値の振幅の平均をとり，平均よりも大きい部分を声の部分としている．声を判別するための手法としてはlpc回帰分析を行いフォルマントを求める手法も用いたが，本システムで用いた手法と精度を試したところ，本システムで用いた手法の方がわずかであるが精度が高くなり，また処理時間も本システムの手法に比べて極めてかかってしまうため不採用とした．しかし，本システムの精度が極めて低くなる状況もある．

5 おわりに

本稿では，映像作品の声優を認識するための話者認識器作成を想定した画像音声対収集支援のためのシステムを試作した．利用者が単純な操作のみで収集を行えるようにすることで，時間も手間もかかる収集を容易に行うことができる．また，顔や声を認識する機能により収集に不要な部分をユーザに提示することで，ユーザがタグ付けの不要な部分を把握し，さらに短時間で収集を行えることを目指した．

参考文献

[1] 竹原正矩，加藤狩夢，田村哲嗣，天目隆平，蔵田武志，速水悟．“業務音声の発話区間検出による作業推定の改善”．電子情報通信学会論文誌 D，Vol.J97-D，No.10，pp.1563-1571，2014.