

視線情報と話者情報とを組み合わせた動画への動的字幕配置手法

赤堀 渉* 平井 辰典† 森島 繁生‡

概要. 本研究では、鑑賞者の冗長な視線移動を減少させるための新たな字幕の配置方法を提案する。従来広く使用される画面下部中央に固定して配置される字幕では、鑑賞者は映像内容の理解のために視線を字幕へ逸らさなければならず、映像を集中して鑑賞することができなかった。この問題に対し、従来研究では、映像の視覚的注意の集まりやすい領域（以降、関心領域と呼ぶ。）の付近に字幕を配置し、鑑賞者の冗長な視線移動を減少させる方法が提案されてきた。しかし、従来手法では、字幕と映像の関心領域とが重複してしまい、鑑賞者にとって映像が不自然に見える場合があった。そこで我々は、複数人の鑑賞者の視線追跡結果から関心領域を推定し、推定された関心領域の外部に字幕を配置することで、字幕と映像の関心領域との重複を避けることを可能にする。さらに、映像中の画像と音声の情報から発話者を検出し、関心領域の外部かつ検出された発話者の付近に字幕を配置することで、鑑賞者の鑑賞体験を低下させることなく映像中の発話者を容易に認識することを可能にする。

1 はじめに

現在、外国語の映像の発話内容の理解を補助するため、映像中に字幕を配置する方法が広く普及されている。従来広く使用される字幕は、映像中の下部中央の固定された位置に配置される。しかし、鑑賞者は映像中の注視したい領域と字幕の領域との間で視線を頻繁に移動させなければならず、映像を集中して鑑賞することができない。したがって、鑑賞者の冗長な視線移動を減少させ、映像内容と発話内容をより効率的に取得できる字幕配置方法が必要とされる。このような背景から、発話者やユーザ個人の視線付近に字幕を配置する方法 [2][3] が提案されているものの、字幕と映像の関心領域とが重複して映像の見た目が不自然になる場合がある。

そこで本研究では、視線情報と話者情報に基づき字幕を配置することで、映像の見た目の自然さを保ちつつ鑑賞者の冗長な視線移動を減少させる方法を提案する。提案手法の概要を図1に示す。提案手法では、まず、入力映像に対する複数人の視線追跡結果を用いて、映像中の関心領域を計算する。この関心領域の外部に字幕を配置することにより、多くのユーザが注視したい領域と字幕とが重複する状況を避けることができる。次に、映像の画像と音の情報から話者検出を行い、それにより得られる発話者の位置と関心領域に基づいて字幕を配置する。字幕を話者付近に配置することで、鑑賞者は容易に映像中の発話者を認識することができる。

2 提案手法

提案手法では、音声付き映像とその字幕ファイル、複数人の視線追跡結果を入力として、字幕の位置が異なる映像を出力する。ここで、字幕ファイルは、発話情報と時間情報を含む字幕セグメントで構成される SRT フォーマットの文書ファイルとする。時間情報は、各々の字幕が画面に出現する時間と消失する時間を示している。前処理として、字幕ファイルの時間情報に基づき、映像を発話区間に分割しておく。我々の手法は、1) 関心領域の推定、2) 話者検出、3) 字幕配置の大きく三つの構成に分けられる。

2.1 関心領域の推定

映像の見た目を自然に保つため、字幕と鑑賞者の視覚的注意の集まる領域とが重複しないように関心領域を定義しなければならない。そこで我々は、映像鑑賞時の複数人の視線追跡結果から、映像中の関心領域を算出する。以降、視線追跡の手順と関心領域の取得方法を述べる。

2.1.1 視線追跡結果の取得

我々は、“ローマの休日”（英語音声）という映画から約 90 秒の映像を選択した。大学関係者の中から集められた、日本語を母語とし英語の基本的な知識を保有する 6 名の実験参加者（男 5 名、女 1 名、20 代）に、41 インチのディスプレイに表示された映像を見てもらった。参加者はディスプレイから約 2.0m 離れた位置に座り、視線測定装置（EMR-9）を参加者の頭部に装着した状態でキャリブレーションを実行し、視線追跡結果を 60Hz で記録した。

Copyright is held by the author(s).

* 早稲田大学

† 駒澤大学

‡ 早稲田大学理工学術院総合研究所

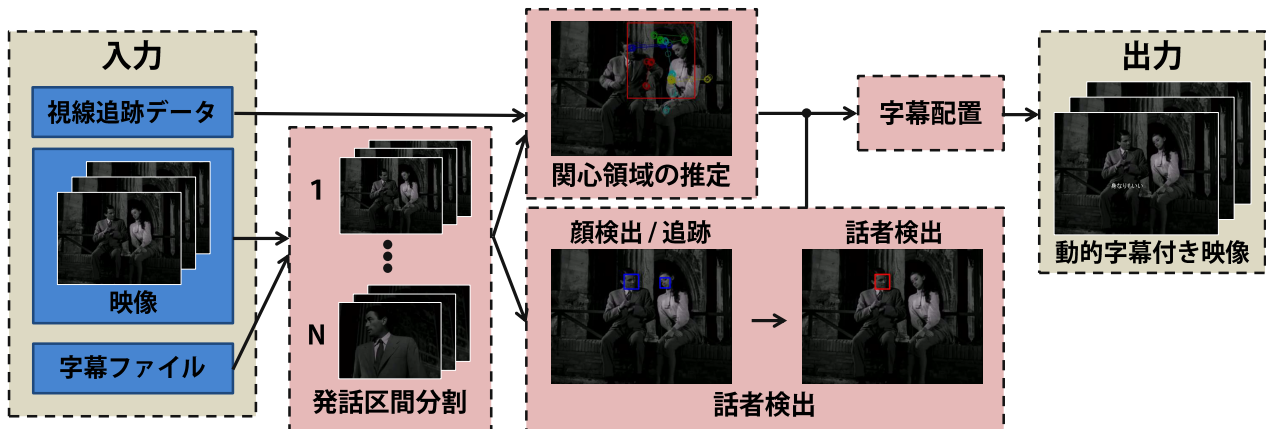


図 1. 提案手法の概要.

2.1.2 関心領域の推定

鑑賞者の関心領域は、字幕が表示される時間においても映像内容に依存して変化する。Katti ら [3] は、字幕位置を関心領域の変化に追従させた場合、鑑賞者に悪い印象を与えることを明らかにしている。そこで我々は、字幕表示時間における複数人の視線追跡結果を用いて、関心領域を一つの矩形として定義する。各発話区間において、複数人の鑑賞者の視線位置の平均 μ と標準偏差 σ を計算し、 $\mu \pm 2\sigma$ で囲まれる矩形領域で関心領域を定義する。¹

2.2 話者検出

鑑賞者が動的字幕により映像中の非発話者を発話者であると誤解することを避けるため、話者検出を行う必要がある。まず、dlib 内に実装されている顔検出と物体追跡の手法を用いて顔検出及び顔追跡を行う。そして、追跡結果から Hu らの話者検出アルゴリズム [2] に基づき話者を検出する。このアルゴリズムは、1) 口唇の動き、2) 人物の位置関係、3) 顔追跡時間と発話時間との相関、4) 音量と口領域との相関の4つの特徴から成り立つ。今回使用した映像における話者検出精度は、57.14% であった。

2.3 字幕配置

2.3.1 字幕配置の候補領域の定義

関心領域の付近であれば、いずれの場所も字幕位置の良い候補点になり得る。Hu ら [2] は、効率良く字幕位置を計算するため、話者の顔領域の周辺の8つの候補点（話者の顔の左上、上、右上、左下、下、右下、左、右）を定義し、その中から字幕位置を最適化している。しかし、映像中の関心領域が大きい場合、関心領域の左右に横文字の字幕を配置する十分な領域がなくなる。さらに、鑑賞者の視線ストレスを軽減させるために連続する字幕間で字幕位置を大きく変化させるべきではないことと、多くの鑑賞者は画面下部に配置される字幕に慣れていることを踏まえて、関心領域の下部領域に字幕を配置する。

2.3.2 字幕配置

推定された関心領域と話者検出の結果に基づき字幕位置を決定する。字幕領域の中心の x 座標は、話者存在時は話者の顔の中心の x 座標とし、話者不在時は関心領域の中心の x 座標とする。字幕領域の y 座標は、関心領域の下端と字幕領域の上端とが重なる位置とする。

3 まとめと今後の予定

我々は、字幕付き映像の鑑賞体験を向上させるため、視線情報と話者情報とを組み合わせた新たな字幕配置手法を提案した。提案手法では、複数人の視線追跡結果から映像中の関心領域を推定し、推定された関心領域の外部かつ話者の顔付近に字幕を配置する。今後は、提案手法の有効性を示すため、ユーザの視線追跡結果の分析と主観評価の調査を行う予定である。

謝辞

本研究の一部は JST CREST の支援を受けた。

参考文献

- [1] W. Akahori, T. Hirai, S. Kawamura, and S. Morishima. Region-of-Interest-Based Subtitle Placement Using Eye-Tracking Data of Multiple Viewers. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, pp. 123–128. ACM, 2016.
- [2] Y. Hu, J. Kautz, Y. Yu, and W. Wang. Speaker-following video subtitles. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(2):32, 2015.
- [3] H. Katti, A. K. Rajagopal, M. Kankanhalli, and R. Kalpathi. Online estimation of evolving human visual interest. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(1):8, 2014.

¹ 関心領域の推定方法の詳細は [1] を参照されたい。