

映像における印象操作の対象人物の検出と印象推定

田村 優幸[†] 新田 直子[†] 中村 和晃^{††} 馬場口 登[†]

[†] 大阪大学大学院工学研究科 〒565-0871 大阪府吹田市山田丘2-1

^{††} 東京理科大学工学部 〒125-8585 東京都葛飾区新宿6-3-1

E-mail: [†]{tamura,naoko,babaguchi}@nanase.comm.eng.osaka-u.ac.jp, ^{††}nakamura.kazuaki@rs.tus.ac.jp

あらまし 素材映像からの部分映像の抽出、並び替えなどによる映像編集により、映像中の特定の物体に対して意図的に印象を操作することが一般に行われている。本研究では、特定の人物に対し positive もしくは negative な印象を与えるよう、専門家が意図的に編集していると考えられる映像から、印象操作の対象人物、およびその印象を推定する手法を提案する。人物の与える印象に影響する要素として、外観や言動などが想定され、これらは特定のフレーム中の対象人物領域の画像特徴や発話内容のみでなく、対象人物以外の領域や他フレームとの関係性にも関連すると考えられる。よって、特定のフレームに対し、その付近における全ての人物の発話内容、もしくは関連する他フレームを同時に入力し、Attention 機構を持つ物体検出モデルを利用し、印象操作の対象人物、およびその印象の推定を試みる。本稿では、登場人物ごとに役割が既知であり、各役割を印象付けるため、専門家による意図的な演出がなされていると考えられる映画を対象とした実験により、画像やテキストなどの各メディアが印象操作に与える影響について検証する。

キーワード 印象推定、物体検出、印象操作、マルチメディア処理、Attention

1. はじめに

素材映像からの部分映像の抽出、並び替えなどによる映像編集により、映像中の特定の物体に対して意図的に印象を操作することが一般に行われている。こうした映像に対して、物体が視聴者に与える印象を推定した場合、その印象の偏りから意図的な編集が行われたことを検知できると考えられる。本研究では、映像から印象操作の施された物体を検出し、その物体が視聴者に与える印象を推定することを目的とする。

画像から特定の物体を検出するために、これまで DPM (Deformable Part Model) [1] のように特定の領域に対し、領域の画像特徴に基づき物体クラスを推定し、高い信頼度で特定の物体クラスと推定された領域を物体領域として検出する手法が多く提案されてきた。近年は、YOLO (You Only Look Once) [2] のように、領域検出と物体クラス推定を回帰問題として設定し、同時に実現する手法も提案されている。また、対象領域のみでなく周辺領域の画像特徴も推定精度の向上に重要であるため、自然言語処理において近年注目を集めている、文章中の単語間の関係を考慮するための Attention 機構を持つ Transformer を画像における物体検出に適用することにより、この回帰問題を解決する DETR (DEtection TRansformer) [3] も提案されている。また DETR を拡張したものとして、画像の説明文に応じた物体領域を抽出するため、テキストと画像の関係も考慮した MDETR (Modulated DEtection TRansformer) [4] や、映像から複数のフレーム列の関係性を考慮して物体を検出する TransVOD (Transformer Video Object Detection) [5] なども提案されている。

一方、画像の印象に関連した既存研究としては、画像全体

の印象推定を目的とし、与えられた画像に対し、画像特徴に基づき、positive, neutral, negative といった極性クラスや、happy, sad, fear などの感情クラスを推定するものがほとんどである。映像に対する印象推定では、物体検出と同様に、映像を構成する複数の情報の統合により推定精度が向上でき、例えば Pereira [7] らは、ニュース映像を対象に、キャスターの顔を含むフレーム列から得られる画像特徴、及び音声から得られるテキスト特徴と音響特徴に基づきキャスターの印象が positive か negative かを推定した。また、複数の人物が出現する映像において、与えられた人物領域の画像特徴に加え、周辺の画像特徴や発話内容を考慮し、友人や敵といった人間関係を推定する研究 [8], [9] は、画像中の特定の領域に対して positive/negative といった印象を推定する研究と考えられる。

本研究では、特定の人物に対して positive または negative な印象操作が行われた映像を対象に、映像中の任意のフレームから印象操作が行われた人物の矩形領域の検出とその人物が視聴者に与える印象の推定を同時に実現することを考える。映像を構成する特定フレームにおける人物の印象は、単一フレームの画像情報のみでなく、他フレームとの関係性や発話内容とも関連すると想定される。そこで、他フレームの画像領域や発話内容間の関係性を考慮するための Attention 機構を持つ物体検出モデルを適用できる。本論文では印象操作が行われた映像として映画をデータセットとして利用し、映像における印象操作の検出に対する、画像や発話内容など各情報の有用性を評価する。

2. 提案手法

本研究では、映像中の任意のフレーム F_l から、positive または negative な印象を与える J 人の人物の矩形領域

$\hat{b}_j \in [0, 1]^4 (j = 1, \dots, J)$ およびそれぞれの人物が positive または negative な印象にあたるクラスのいずれかに属する確率 $\hat{p}_j \in [0, 1]^2 (j = 1, \dots, J)$ の組 $\hat{y}_j = (\hat{p}_j, \hat{b}_j)$ を推定することを目的とする。ただし、 $\hat{b}_j \in [0, 1]^4$ は、検出された人物の矩形領域の中心の座標および幅と高さを表すベクトルであり、画像全体に対する相対的な位置を 0 ~ 1 の値で表す。また、クラス $\hat{p}_j \in [0, 1]^2$ は推定対象の各クラスの確率を表すベクトルである。

これは、画像から物体の矩形領域とそのクラスを推定する物体検出と同様の問題設定である。人物の印象は、対象人物の外観やその周辺領域との関係性と関連すると考えられるため、Attention 機構を持つ物体検出モデルが適用できる。しかし、映像を構成する特定フレームにおける人物の印象は、単一フレームの画像情報のみでなく、発話内容や他フレームとも関連すると想定される。そこで、発話内容を考慮する場合には、任意のフレーム F_l の画像特徴に加え、周辺の単語数 v からなる発話内容 $T_{l,k} (k = 1, \dots, v)$ から得られるテキスト特徴を共に入力とし、他フレームとの関係を考慮する場合には、任意のフレーム F_l と前後のフレーム $F_{l-n}, \dots, F_{l+n} (n \neq 0)$ のフレーム間の関係を考慮する機構を追加した上で、Attention 機構を持つ物体検出モデルを学習する。

以下では、まず単一フレームを入力とし、自然言語処理によく用いられる Transformer を採用した、Attention 機構を持つ物体検出モデルである DETR をベースモデルとして説明し、続いてテキスト特徴、及び他フレームを入力に加えたモデルについて述べる。

2.1 DETR [3]

図 1 に DETR の概要図を示す。DETR は、 N 個の物体の矩形領域とそのクラスを推定する。ただし、 N は画像から検出されうる物体の上限数とする。DETR では、まず、3 チャンネルかつ $h_0 \times w_0$ の入力画像 $F_l \in R^{3 \times h_0 \times w_0}$ を CNN に入力し、画像特徴マップ $f_l^{Imap} \in R^{c \times h \times w}$ を生成する。この画像特徴マップに位置情報を付与して Transformer を適用し、画像中の領域間の関係を考慮しながら、物体検出を行う。Transformer のエンコーダは特微量列を入力とするため、画像特徴マップを $h \times w$ 個の d 次元画像特微量列 $f_l^I \in R^{d \times h \times w}$ に変換する。これをエンコーダに入力することにより、画像中の各部分領域に対し、周辺領域も考慮した画像特微量列 $f_l^{I,enc} \in R^{d \times h \times w}$ を得る。

デコーダは N 個の Object query に対し、エンコーダから出力される画像特微量との関係性を考慮した query 特微量 $q_l \in R^{N \times d}$ を出力する。最後に、この query 特微量を全結合層から成る検出層に入力することにより、推定対象のフレーム F_l 内の印象操作の対象人物の矩形領域 $\hat{b}_j \in [0, 1]^4 (j = 1, 2, \dots, N)$ および各クラスに属する確率 $\hat{p}_j \in [0, 1]^3 (j = 1, 2, \dots, N)$ をそれぞれ推定する。DETRにおいては最大 N 個の物体を検出するが、推定対象の画像には常に N 個の物体が含まれているとは限らない。そこで推定対象のクラスに、positive/negative に検出可能な物体が存在しないことを表すクラス ϕ を加え、計 3 クラスの確率を推定する。

学習データとしては、各フレームに存在する印象操作の対象人物の矩形領域の中心の座標および幅と高さを表す $b_i \in R^4$ 、およびその印象を表すクラス $c_i \in \{\text{positive}, \text{negative}, \phi\}$ の組 $y_i = (c_i, b_i)$ を用い、この推定結果に対し以下のように算出される Hungarian Loss を最小化するよう学習する。ここで、各フレームに存在する印象操作対象人物が N 人未満である場合、不足分はクラス ϕ で埋める。

まず、正解データ y_i と推定器により検出された \hat{y}_j を下記のように定義される $\mathcal{L}_{match}(y_i, \hat{y}_j)$ が最小化されるよう対応付ける。

$$\mathcal{L}_{match}(y_i, \hat{y}_j) = -\mathbb{1}_{\{c_i \neq \phi\}} \hat{p}_j + \mathbb{1}_{\{c_i \neq \phi\}} \mathcal{L}_{box}(b_i, \hat{b}_j) \quad (1)$$

ただし、 \hat{P}_j は j 番目の検出結果の各クラス確率 \hat{p}_j の内、クラス c_i の確率である。また、 $\mathcal{L}_{box}(b_i, \hat{b}_j)$ は j 番目の検出結果の矩形領域 \hat{b}_j と b_i の IoU と L1 距離により以下のように定義される。

$$\mathcal{L}_{box}(b_i, \hat{b}_j) = \lambda_{iou} \mathcal{L}_{iou}(b_i, \hat{b}_j) + \lambda_{L1} \|b_i - \hat{b}_j\|_1 \quad (2)$$

ただし、 $\lambda_{iou}, \lambda_{L1} \in R$ はハイパーパラメータであり、IoU と L1 距離を重みづけをして線形結合するために用いられる。

このように、検出結果 $\hat{y}_j = (\hat{p}_j, \hat{b}_j)$ のうち、正解データ $y_i = (c_i, b_i)$ と対応付けられたものを $\hat{y}_{\sigma(i)} = (\hat{p}_{\sigma(i)}, \hat{b}_{\sigma(i)})$ と表し、以下のように定義される Hungarian Loss を最小化するように DETR のパラメータを学習する。

$$\mathcal{L}_{Hungarian}(y, \hat{y}) = \sum_{i=1}^N [-\log \hat{p}_{\sigma(i)} + \mathbb{1}_{\{c_i \neq \phi\}} \mathcal{L}_{box}(b_i, \hat{b}_{\sigma(i)})] \quad (3)$$

2.2 発話内容を考慮した MDETR [4]

DETR ではエンコーダにおいて、画像特徴マップの領域間の関係性を考慮する。DETRにおいて発話内容を考慮する場合、単に画像特徴量にテキスト特徴量を結合し、エンコーダへの入力とすることにより、画像領域間、単語間、および画像領域と単語間、それぞれの関係性が考慮されると考えられる。そこで、図 2 に示す MDETR を利用する。本研究では、テキスト特徴量を抽出し、画像特徴量に結合する以外の処理は DETR と同一となる。

まず、テキスト特徴抽出器を用い、推定対象のフレーム F_l の周辺の v 語の単語列からなる発話内容 $T_{l,k} (k = 1, \dots, v)$ から得られる v' 個のトークンに対し、 d_{text} 次元テキスト特徴量 $t_l \in R^{v' \times d_{text}}$ を抽出する。

さらに、こうして得られたテキスト特徴を、画像特徴量 $f_l \in R^{d \times h \times w}$ と結合して DETR のエンコーダに入力するため、全結合層を用いて、画像特徴量と同じ d 次元に変換する。結合された特徴量列 $f_l^{comb} \in R^{(hw+g) \times d}$ を DETR のエンコーダに入力することにより、画像領域間、単語間、及び画像領域と単語間、それぞれの関係性を考慮した、マルチモーダルな特徴量列 $f_l^{comb,enc} \in R^{(hw+g) \times d}$ を生成する。

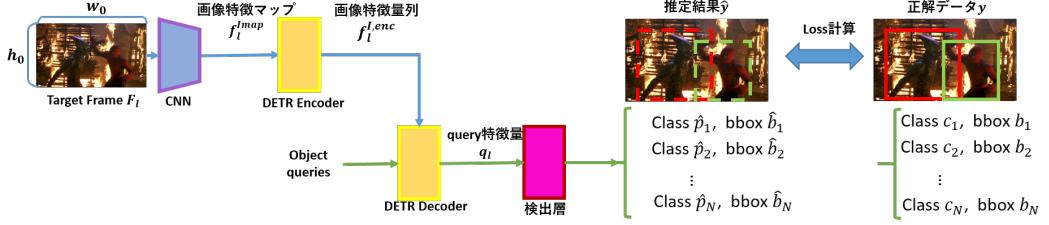


図 1 DETR の概要

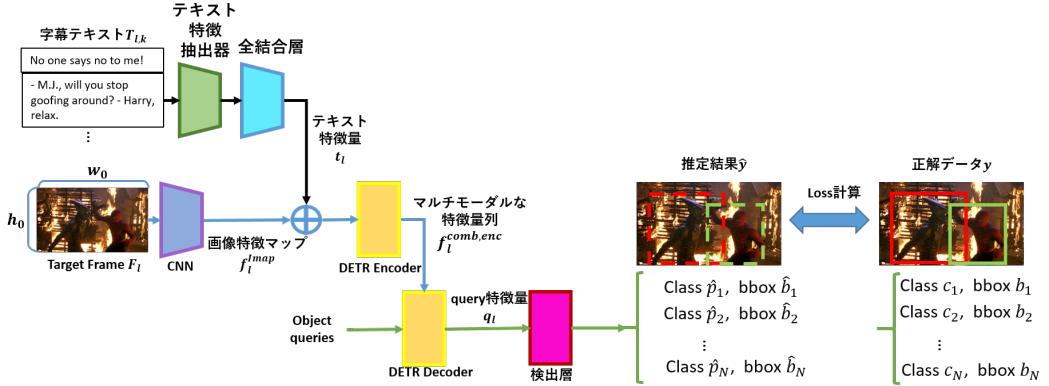


図 2 フレームの画像特徴と発話内容のテキスト特徴を用いたネットワークの概要

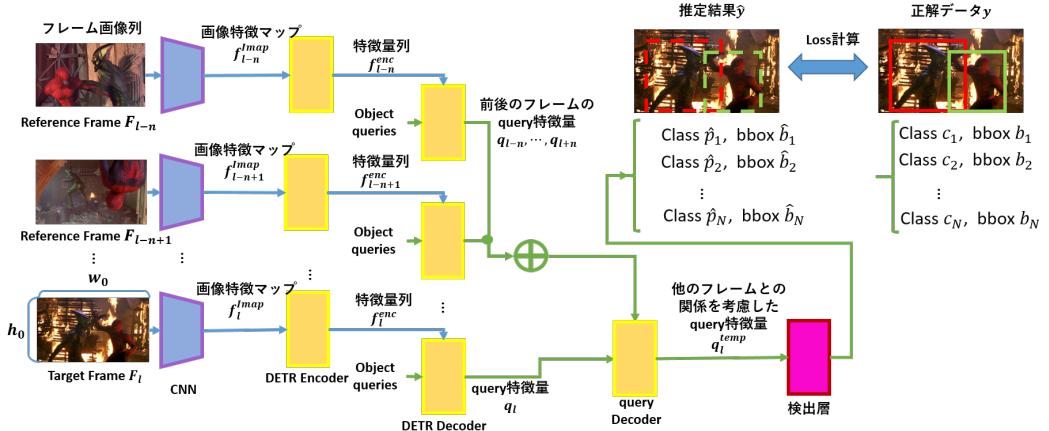


図 3 他フレームとの関係を考慮するネットワークの概要

2.3 他フレームとの関係を考慮した人物検出と印象推定

DETRにおいて関連する他フレームとの関係を考慮する場合、発話内容を考慮する場合と同様に、推定対象のフレームの画像特徴量に関するフレームの画像特徴量を結合し、DETRのエンコーダに入力する手法が有効であると考えられるが、計算量やメモリの面でコストが大きい。そこで対案として、各フレームから抽出される、検出時に用いられるquery特徴量間の関係を考慮しながら新たなquery特徴量を算出するため、DETRに対し、図3に示すようにDETRデコーダと同一の構造をしたqueryデコーダを追加する。

推定対象のフレーム \$F_l\$ の関連するフレーム \$F_{l-n}, \dots, F_{l+n} \in R^{3 \times h_0 \times w_0} (n \neq 0)\$ について、2.1節と同様の処理を行い、query特徴量 \$q_{l-n}, \dots, q_{l+n} \in R^{N \times d} (n \neq 0)\$ を生成する。推定対象のフレーム \$F_l\$ のquery特徴量 \$q_l \in R^{N \times d}\$ と関連するフレーム \$F_{l-n}, \dots, F_{l+n} (n \neq 0)\$ のquery特徴量

\$q_{l-n}, \dots, q_{l+n} \in R^{N \times d} (n \neq 0)\$ との間の関係性をqueryデコーダで考慮し、最終的なquery特徴量 \$q_l^{temp} \in R^{N \times d}\$ を生成する。queryデコーダ以外の部分についてはDETRと同様の処理とする。

3. 評価実験

本章では、実験に用いるデータセットの作成について記述し、その後作成したデータセットを用い、単一フレームのみを用いた推定実験、発話内容を考慮した推定実験、他フレームとの関係を考慮した推定実験の3つを行う。

3.1 データセットの作成

本研究には専門家によって特定の人物に対して印象操作が行われている映像がデータセットとして必要になる。こうした映像の例として政治的な話題に関するニュース映像や広告映像などが考えられる。しかし、こうした映像は収集や正解となる印

象操作の対象人物の矩形領域およびその印象のラベルの付与が困難である。

そこで本研究では映画に着目する。映画には人物領域などのメタデータが付与されたデータセット[10]が公開されており、データの収集が容易である。また、映画は登場人物毎に役割が決められており、その役割を強調する演出がなされている。特に、アクション映画では、主人公(Hero)に対して positiveな印象操作、敵役(Villain)に対して negativeな印象操作を施すよう演出がなされていると考えられる。

1,100本の映画に対する MovieNet データセット[10]は、タイトルやジャンル、あらすじや字幕に加え、ショットのキーフレーム、及び主要な 10 名程度の登場人物に対し、登場フレーム、及びフレーム内の該当する矩形領域をメタデータとして含む。そこで、アクション映画の中でもメタデータに字幕を含む 57 本に対して、wikipedia[11]等のあらすじや登場人物に関する記述から、手動で主人公に相当する登場人物に positive、敵役に相当する登場人物に negative をラベル付けし、映画内でこれらの登場人物に一貫して該当する印象操作が施されているという前提のもと、フレーム内の各登場人物の矩形領域に positive、negative とラベル付けした。

この結果、68名の登場人物に対して positive、78名の登場人物に対して negative が付与され、いずれかのラベルを持つ登場人物が登場する 33,435 フレームに対してラベルが付与された。図 4 に主人公および敵役の登場頻度を表すグラフを示す。グラフは横軸が一つの映画における主人公または敵役の登場回数を、縦軸が映画の本数を表し、右上には総登場数を記載している。グラフより、主人公は多い映画では 1,000 ショットを越えて登場するのに対して、敵役はほとんどの映画において 300 ショット以下しか登場せず、総登場数は 3 倍以上の差があることが分かる。実験に用いるデータでは、クラス間でのデータ量の不均衡を避けるため、主人公と敵役の登場頻度が出来るかぎり等しくなる様にフレームを選出した。実験では、47本の映画の 9,033 フレームを学習データ、7本の映画の 692 フレームを検証データ、9本の映画の 2,664 フレームをテストデータとした。最後にデータセットに含まれるフレーム上に主人公または敵役を表す矩形領域とラベルを可視化したものを図 5 に示す。各人物に矩形領域とラベルを付与できていることが確認できる。

3.2 実験と考察

実験ではフレームの画像サイズ $h_0 \times w_0$ に対して、画像特徴マップのサイズ h, w はそれぞれ $h, w = \frac{h_0}{32}, \frac{w_0}{32}$ とした。また、画像特徴列の次元は圧縮前後で $c = 2048$ から $d = 256$ に変化させた。さらに Object query の数 $N = 100$ 、DETR のエンコーダ・デコーダは共に 6 層、Loss 計算の際に用いられるハイパーパラメータはそれぞれ $\lambda_{iou} = 2, \lambda_{L1} = 5$ とした。なお、推定精度の担保と学習時間の短縮を目的に、利用可能な事前学習済みの重みが存在する場合は利用した。また人物領域の検出では、クエリ特徴量から得られる候補のうち、尤度が閾値を超えた矩形領域を検出が成功したと判定する。閾値は DETR の作成者の設定に倣い、0.7 とした。

実験結果の評価においては、検出した矩形領域と正解データ

の矩形領域の IoU 値が閾値以上で、かつ印象推定結果が正解データのラベルと一致したものを正しい検出結果と判定する。なお、正解データに存在しない人物領域を検出したとき誤検出、正解データに存在する人物領域が検出されなかったとき未検出と判定する。また、IoU とは物体検出での評価手法であり、図 6 に示すように、正解データと検出結果の矩形領域がどれほど重なっているかを表す。図 7 から IoU 値の閾値は 0.5~0.7 程度で十分と考えられる。

まず、ベースモデルである、単一フレームのみを入力とする DETR を画像における物体検出タスクで事前学習した重みを初期値とし、160 エポック学習したときの推定結果を表 1 に示す。IoU は 0.5~0.7 程度で十分と考えると、適合率は 40% 程度、再現率は 50% 程度となった。

次に、発話内容を入力に加える。MovieNet では、各映画のショットに発話内容の字幕テキストがメタデータとして付与されている。ここから、推定対象のフレームの周辺のショットから単語数が最大 20 語となる様に発話内容を選択し、学習に利用した。選択された字幕テキストの例を図 8 に示す。RoBERTa[12] のエンコーダをテキスト特微量抽出器とし、画像に対して説明文に応じた物体を検出するタスクで事前学習した MDETR の重みを初期値とし、2.2 節で紹介した推定器を 90 エポック学習した結果を表 2 に示す。表 1 の結果と比較すると、誤検出を大きく減少させることにより、再現率を 50% 程度に保ったまま、適合率が 45% 程度に向上したことが分かる。また、発話内容を考慮したとき F 値も 1% 程度向上している。

図 9 に発話内容を考慮することにより誤検出が改善された例を示す。人物領域が小さい、人物が見切れている、人物が後ろを向いている、人物の周囲に人が多い、画面が暗い、といった要因により不明瞭な人物に対する誤検出が減少する傾向が見られた。また、最上段の例について、向けられた Attention を可視化したものが図 10 である。図では明るい色をしている部分ほど強く Attention が向けられている。発話内容を入力に加えたことにより、テキスト部分に対して一様に Attention が向けられる様になり、その分、画像部分に対する Attention が減少し、人物領域周辺で分散するようになった。このため、多くの誤検出を含め、全体の検出数が減少したと考えられる。これは本実験において、発話内容を考慮することに対して期待していた効果とは言い難いため、発話内容に対して一様に Attention が向かない様な工夫が必要である。

また、図 11 の例では、'devil' のような negative な印象を与えると考えられる発話内容が入力されたことにより、推定された印象が positive に変更された。この様に発話内容が必ずしも推定精度の向上に寄与してはいない。これはデータセットに印象推定に寄与する発話内容が少なく、学習が十分でないことが原因に挙げられる。こうした問題には、データセットの増量が有効であると考えられる。

最後に、推定対象のフレームに関連する他フレームを入力に加える。3.1 節で作成したデータセットの各フレームに対して、その直前のショットの 1 フレームを関連するフレームとして選択する。選択された直前のショットのフレームの例を図 12 に示す。

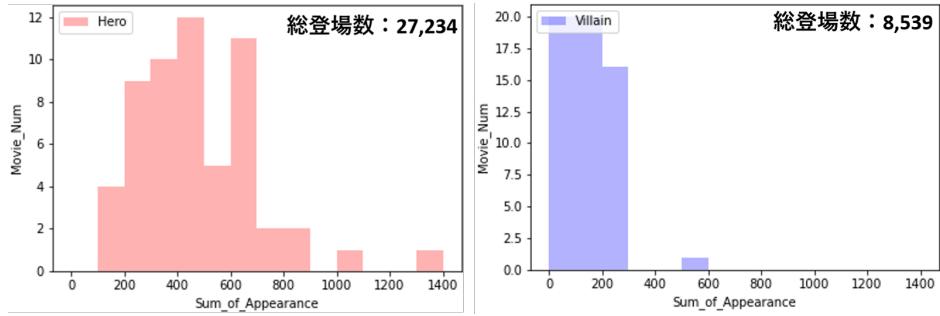


図 4 主人公および敵役の登場頻度

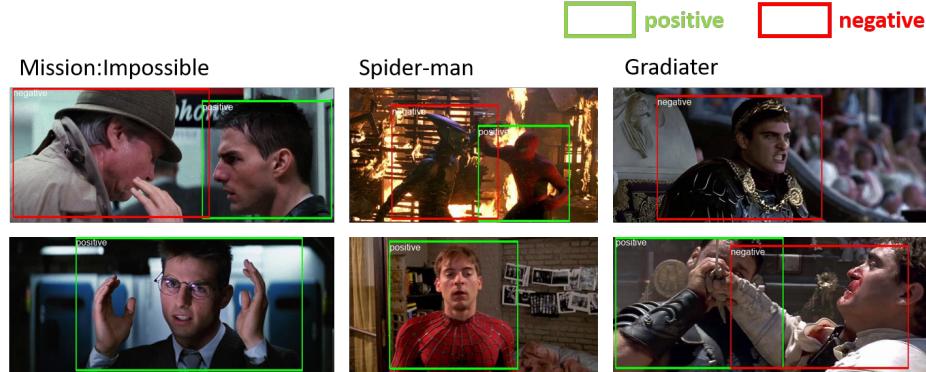
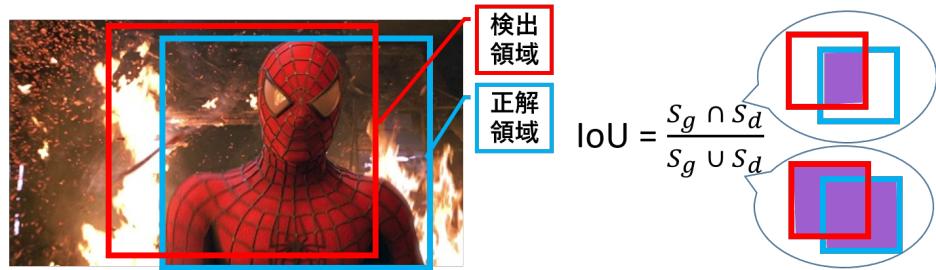


図 5 データセットの例



正解領域の面積： S_g 検出領域の面積： S_d

図 6 IoU の定義



図 7 IoU の例

す。本実験では、各フレームに対する query 特徴量を生成する部分までは、単一フレームの画像特徴に基づく推定実験と同じ処理になる。そこで、画像特徴マップを生成する CNN, DETR エンコーダ、DETR デコーダは、単一フレームのみを入力とする DETR を 160 エポック学習した重みで固定し、query デコーダおよび検出層のみを学習する。また、query デコーダは

6 層とした。

以上の条件で 2.3 節で紹介した推定器を 110 エポック学習した結果を表 3 に示す。表 1 の結果と比較すると、ラベルの一致が少量増加し、再現率が 0.5%程度、適合率が 1%程度上昇し、F 値も向上したことが分かる。

直前のショットの 1 フレームを考慮した推定では、図 13 の

表 1 単一フレームの画像特徴に基づく推定結果

Intersection over Union	0.5	0.6	0.7	0.8	0.9
未検出	312	351	433	639	1279
誤検出	882	921	1003	1209	1849
ラベル不一致	1118	1104	1063	983	693
ラベル一致	1602	1577	1536	1410	1060
適合率 (%)	44.48 (1602 /3602)	43.78 (1577 /3602)	42.64 (1536 /3602)	39.14 (1410 /3602)	29.43 (1060 /3602)
再現率 (%)	52.84 (1602 /3032)	52.01 (1577 /3032)	50.66 (1536 /3032)	46.51 (1410 /3032)	34.96 (1060 /3032)
F 値 (%)	48.3	47.54	46.31	42.51	31.96

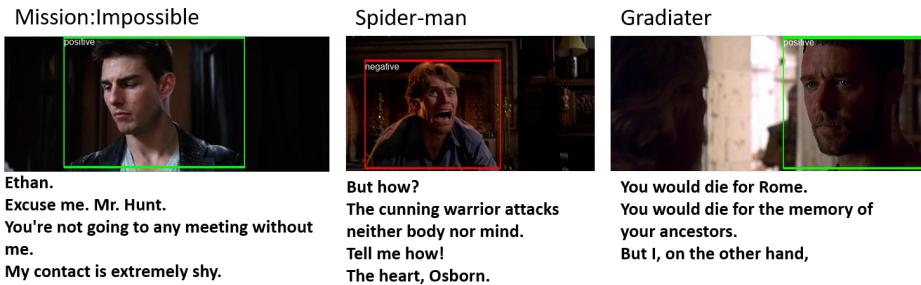


図 8 発話内容の例

表 2 発話内容を考慮した推定結果

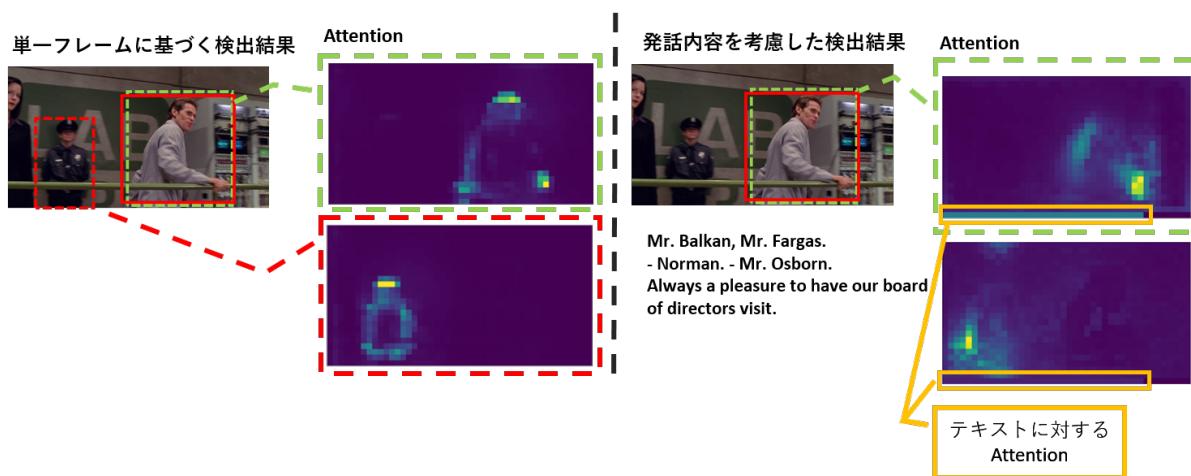
Intersection over Union	0.5	0.6	0.7	0.8	0.9
未検出	462	501	575	728	1332
誤検出	608	647	721	874	1478
ラベル不一致	1027	1010	984	919	661
ラベル一致	1543	1521	1473	1385	1039
適合率 (%)	48.55 (1543 /3178)	47.86 (1521 /3178)	46.35 (1473 /3178)	43.58 (1385 /3178)	32.69 (1039 /3178)
再現率 (%)	50.89 (1543 /3032)	50.16 (1521 /3032)	48.58 (1473 /3032)	45.68 (1385 /3032)	34.27 (1039 /3032)
F 値 (%)	49.69	48.99	47.44	44.61	33.46

表 3 他フレームとの関係を考慮した推定結果

Intersection over Union	0.5	0.6	0.7	0.8	0.9
未検出	344	382	457	648	1288
誤検出	853	891	966	1157	1797
ラベル不一致	1070	1049	1022	940	676
ラベル一致	1618	1601	1553	1444	1068
適合率 (%)	45.69 (1618 /3541)	45.21 (1601 /3541)	43.86 (1553 /3541)	40.78 (1444 /3541)	30.16 (1068 /3541)
再現率 (%)	53.36 (1618 /3032)	52.8 (1601 /3032)	51.22 (1553 /3032)	47.63 (1444 /3032)	35.22 (1068 /3032)
F 値 (%)	49.23	48.71	47.25	43.94	32.5

様に、複数フレームに同一人物が登場する場合に、ラベル一致が増加する例が見られた。これは直前のショットのフレームに

同一人物が存在するという情報を取得し、推定を補助していると考えられる。本研究では直前のショットの 1 フレームのみを



入力したが、複数のフレームを入力することにより、更なる精度の向上が期待される。

また本研究では、関連するフレームとして、どのようなフレームを選択すべきであるかという検証が不十分であり、今後の課題としたい。

4. ま と め

本研究では、特定の人物に対し positive もしくは negative な印象を与えるよう、専門家が意図的に編集していると考えられる映像を対象に、Attention 機構を持つ物体検出モデルを利

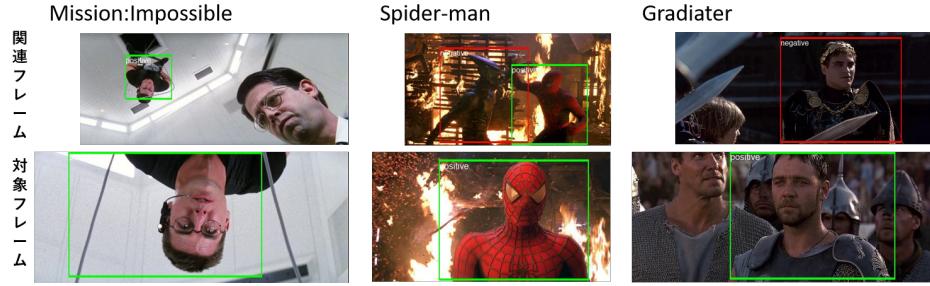


図 12 直前のショットの 1 フレームの例



図 13 直前のショットの 1 フレームを考慮してラベル不一致が改善した例

用し、印象操作の対象人物、およびその印象を推定する手法を提案した。57 本の映画から得られるフレームに加え、発話内容や他フレームとの関係を考慮して、人物検出・印象推定器を学習した。発話内容から得られるテキスト特徴量を利用することにより、検出精度が向上したが、発話内容に対する学習が十分に行えたとは言い難い。一方、推定対象のフレームの直前のショットの 1 フレームとの関係を考慮することにより、推定精度が向上することを確認した。本研究では、テキストに対して一様に Attention が向けられており、これを重要な単語ほどより強く Attention が向けられるように改善する必要がある。また、データセットの增量や関連するフレームを増加させた実験、関連するフレームの選択に関する検証などが今後の課題として挙げられる。

本研究の一部は、JST CREST JPMJCR20D3、科学研究費補助金基盤（C）19K12019 の助成による。

文 献

- [1] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A Discriminatively Trained, Multiscale, Deformable Part Model,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, 2008.
- [2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788, 2016.
- [3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-End Object Detection with Transformers,” European Conference on Computer Vision, pp. 213–229, 2020.
- [4] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, “MDETR - Modulated Detection for End-to-End Multi-Modal Understanding,” IEEE International Conference on Computer Vision, pp. 1780–1790, 2021.
- [5] L. He, Q. Zhou, X. Li, L. Niu, G. Cheng, X. Li, W. Liu, Y. Tong, L. Ma, and L. Zhang, “End-to-End Video Object Detection with Spatial-Temporal Transformers,” arXiv

preprint, abs/2105.10920 , 10 pages, 2021.

- [6] R. Rothe, R. Timofte, and L. Van Gool, “Some Like It Hot-Visual Guidance for Preference Prediction,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 5553–5561, 2016.
- [7] M. H. R. Pereira, F. L. C. Pádua, A. C. M. Pereira, F. Benevenuto, and D. H. Dalip, “Fusing Audio, Textual, and Visual Features for Sentiment Analysis of News Videos,” AAAI Conference on Web and Social Media, 4 pages, 2016.
- [8] X. Liu, W. Liu, M. Zhang, J. Chen, L. Gao, C. Yan, and T. Mei, “Social Relation Recognition from Videos via Multi-Scale Spatial-Temporal Reasoning,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 3566–3574, 2019.
- [9] A. Kukleva, M. Tapaswi, and I. Laptev, “Learning Interactions and Relationships between Movie Characters,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 9849–9858, 2020.
- [10] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin, “Movienet: A Holistic Dataset for Movie Understanding,” European Conference on Computer Vision, pp. 709–727, 2020.
- [11] “Wikipedia,” <https://ja.wikipedia.org/wiki/>
- [12] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, “Roberta: A Robustly Optimized Bert Pretraining Approach,” arXiv preprint, abs/1907.11692, 22 pages, 2019.