E_017

クローズドキャプションを対象とした被写体の動作推定

A Study on Detecting Behavior of Video Objects Using Closed Captions

三浦菊佳* 山田一郎* 住吉英樹* 八木伸行*
Kikuka Miura Ichiro Yamada Hideki Sumiyoshi Nobuyuki Yagi
奥村学† 徳永健伸‡
Manabu Okumura Takenobu Tokunaga

1. はじめに

近年、放送局では大量の番組を蓄積する環境が整備されてきた。これらの番組には貴重な映像が含まれているため、教育用途や他の番組での素材としての再利用が望まれる。番組の効果的な二次利用のためには、番組の時間軸上のどの区間に何が描かれているかを示すセグメントメタデータを、時間と労力をかけずに自動的に付与する技術が求められる。

これまで我々は、字幕放送用の字幕テキストデータ（以後、クローズドキャプションと呼ぶ）を利用して映像中の被写体を特定する手法を提案してきた [1]。さらに動画の特徴を生かし、「ライオンが走っているシーン」など、被写体の動作も指定して映像を特定することを目指し、研究を進めている。

被写体の動作推定に関する研究には、柴田らのクローズドキャプションと画像の特徴を用いた手法[2]がある。この手法では料理番組を対象として、料理の作業区間を取り出し、「切る」「炒める」など各区間で行われる被写体の動作を推定する。動作の遷移が限定されている特徴を利用してHMM により推定を行っているが、遷移が不確定な場合に応用することは困難と考えられる。また、画像解析によって動画から被写体の動作を抽出する手法[3][4][5]も提案されているが、現状では限定された動作のみを対象としており、被写体の全ての動作へ適用とすることは難しい。

本稿では、クローズドキャプションを利用して映像に映る被写体がどのような動作を行っているか推定する手法を提案する。被写体の動作を表現するときの言葉の特徴に着目し、クローズドキャプションに出現する全ての動詞を対象とし、各動詞が被写体の動作を表すか否かを判定する。以下に、クローズドキャプションに出現する動詞の特徴を分析した予備調査と、各動詞が映像中の被写体の動作を表すか否かを判定する実験について報告する。

2. 予備調査

クローズドキャプションは、ナレーションをもとに作成されたテキストであり、映像内容を特定する有益な情報である。しかし、動作のシーンを抽出する場合、単純にその動詞に時間対応する映像を抜き出すだけでは不十分である。たとえば、「ある日新芽を食べた後　群れで変わったことが起きました。」に対応する映像では、「食べる」という被写体の動作は見られなかった。

詳細に検討するために、「食べる」という動詞に着目し、クローズドキャプションに「食べる」が出現したとき、その文に対応する映像カット（カメラの切り替わり点から次の切り替わり点までの映像区間）に被写体の食べる動作が含まれる割合を調査した。動物や自然を紹介する NHK の「地球！ふしぎ大自然」（1 番組 42 分 30 秒）18 番組を対象とした結果を表 1 に示す。

表1. クローズドキャプション中の動詞「食べる」に対応する映像内容

動作を含む	動作を含まない	合計
104	69	173

表 1 では、動詞「食べる」が出現しても、対応する映像カットの約 40% (=69/173) で被写体が食べる動作をしておらず、クローズドキャプション中の動作を表す単語の有無だけでは、被写体の動作を推定できないことがわかる。

番組では、クローズドキャプションのみからでは映像内容を推定できないような場合もある。そこで、人間はどの程度、クローズドキャプションのみから映像中の被写体の動作を推定できるか、動詞「食べる」を対象として実験を行った。この実験では、被験者に動詞「食べる」を含むクローズドキャプション一文を提示し、対応する映像カットに被写体の食べる動作があるか否かを判定した。結果を表 2 に示す。この値は、言語情報のみから自動抽出する際の上限値と考えられる。

表2. 人間による判定結果

適合率	再現率	F 値
84.2% (85/101)	81.7% (85/104)	0.829

表 2 の結果から、人間は映像なしでも言語上の何らかの特徴を捉えて精度良く被写体の動作を推定していることがわかる。さらに、同様の実験を番組全体のテキストを提示して行った。その結果、適合率と再現率の調和平均を示す F 値は 0.843 であった。動詞を含む文の前後を提示しても表 2 の結果と大きな差は見られず、動作を表す動詞を含む一文のみからでも、被写体が動作する様子を表現する特徴を有効に捉えることができると考えられる。

例えば、映像中の被写体の動作を表現する文の特徴には「食べています／食べている」のような現在進行形が考えられる。前述と同じ 18 番組を対象として、「食べています／食べている」がクローズドキャプションに現れていたときに、対応する映像カットに被写体の食べる動作が含まれているか否かを調査した結果を表 3 に示す。

表3. 動詞の現在進行形が被写体の動作を表現する割合

適合率	再現率	F 値
89.3% (25/28)	24.0% (25/104)	0.378

*NHK 放送技術研究所
† 東京工業大学精密工学研究所
‡ 東京工業大学大学院情報理工学研究科

適合率の結果より、現在進行形の動詞は高い確率で被写体の動作を表現することがわかる。しかし、再現率の結果より、他の特徴も抽出する必要があると考えられる。

3. 機械学習を用いた被写体の動作推定

前章では、被写体の動作を表す動詞は、現在進行形以外の特徴もあることを述べた。本章では、現在進行形以外の特徴も考慮するため、動詞「食べる」がクローズドキャプションに出現するときに、対応する映像カットに被写体の食べる動作が含まれるか否か、機械学習を用いて判定する処理について説明する。

3.1 判定処理手順

判定対象とする動詞「食べる」が出現する文から、表4に示す文の特徴を抽出する。野呂らはブログテキストを対象として各テキストがイベントを表すか判定する手法[6]を提案しており、そこで使用された特徴を表4の①～⑤の参考としている。また、動詞を修飾する単語は映像を説明する文の手がかりになると考えられるために、⑥の特徴を設ける。この特徴では、動詞を修飾する単語の品詞が副詞である場合と、「今」「昔」など品詞が名詞でも副詞的機能をもつ場合のみを対象とし、各単語を「現在」「過去」「習慣」「その他」のカテゴリーに分類する。それらの単語により修飾されない場合は「修飾無し」とする。カテゴリー分類には、分類語彙表[7]を用いる。なお、前章で述べたように一文のみからの推定でも有効であると考えられるため、機械学習で利用する素性は一文のみから抽出できる特徴を利用している。

表4. 判定で利用する特徴

① 対象動詞と同じ文節に引用を表す助詞が存在するか否か
② 対象動詞が過去形か否か
③ 対象動詞が未然形か否か
④ 対象動詞に別の動詞が後続して複合動詞となっているか否か
⑤ 対象動詞が現在進行形か否か
⑥ 対象動詞を修飾する単語の種類:「現在」「過去」「習慣」「その他」「修飾無し」

クローズドキャプション中の動詞「食べる」に対応する映像カットに食べる動作があるか否かの2値を人手で付与したものを学習データとし、表4の特徴を素性とした機械学習を行う。学習にはQuinlanのC4.5決定木学習アルゴリズムを用いる。学習の結果生成される決定木から、動詞「食べる」を含む文に対応する映像カットが、被写体の食べる動作であるか否かを判定することができる。

3.2 実験結果および考察

2章の予備調査で用いた、「地球！ふしぎ大自然」18番組を対象として実験を行った。18番組すべてに対し人手により正解を与え、このうち9番組を学習データ、残りをテストデータとし、クロスバリデーションによる2回の試行を行った。決定木学習による判定の評価結果を表5に示す。

表5では、表3の現在進行形のみで抽出した手法よりも再現率が高く、他の特徴も判定の手がかりとして有効であることが分かる。生成された決定木より、「食べ続けています」「食べ始めました」など動詞「食べる」に別の動詞が後続して複合動詞により表現される場合や、動詞が未然形でない場合、「毎日」「通常」などの日常の習慣を表す語がない場合などが映像中の被写体の動作を表すと判定されていた。このような特徴が判定の手がかりになると考えられる。

表5. 実験結果

適合率	再現率	F値
65.7% (71/108)	68.3% (71/104)	0.670

F値も表3の結果と比べて高い値が得られたが、人間による判定結果との差は大きく改善の余地が残されている。例えば、「ようやく1枚食べました。」では、食べ終わった完了を示す文なのか、食べていたという過去を示す文なのか、今回の素性では判断できなかったために抽出できていない。完了と過去を区別することは難しいが、完了は動作を表す重要な手がかりなると考えられる。また、「マンタはプランクトンを食べます。」「マンタは大きな口をあけエラで海水をこしてプランクトンを食べます。」はどちらも今回の手法では映像に動作は含まれていないと判定されるが、人間による予備実験では、後者の文は映像に動作が映っていると判定されており、実際の映像も食べる動作であった。このように被写体の描写が詳細に表現されている文では、たとえ習性を説明している文でも映像に動作が含まれることが考えられ、精度向上のため新たな特徴を考慮していかなければならない。

4. まとめ

本論文では、クローズドキャプションに含まれる「食べる」という動詞を対象として、対応する映像カットに被写体の「食べる」動作があるか否かを判定する手法の検討を行った。対象動詞の付属語や修飾語を素性とした機械学習により、適合率65.7%、再現率68.3%という結果が得られ、一定の分別能力があることを示した。今後、機械学習で利用する特徴を検討するとともに、「食べる」以外の動作を表す動詞や、動作主の特定も行っていく予定である。

参考文献

[1] 三浦菊佳, 山田一郎, 住吉英樹, 八木伸行 :クローズドキャプションを利用した映像主被写体の推定手法, 情報処理学会研究報告, 2006-NL-171, pp.1-6 (2006)
[2] 柴田知秀, 黒橋禎夫 :言語情報と映像情報の統合による作業教示映像の構造的理解, 第2回デジタルコンテンツシンポジウム, 1-1 (2006)
[3] グエン ホウ バッ, 篠田浩一, 古井貞熙 :動的特徴量を用いたHMMによる連続動作認識, 電子情報通信学会総合大会, D-12-120, pp.286 (2004)
[4] 大和淳司, 大谷淳, 石井健一郎 :隠れマルコフモデルを用いた動画像からの人物の行動認識, 電子情報通信学会論文誌, Vol.J76-D-II No.12, pp.2556-2563 (1993)
[5] A. Amir et. al., "IBM ResearchTRECVID 2004 Video Retrieval System", NIST TRECvid2004.
[6] 野呂太一, 乾孝司, 高村大也, 奥村学 :イベントの生起時間帯判定, 情報処理学会研究報告, 2005-NL-170, pp.7-14 (2005)
[7] 国立国語研究所 :分類語彙表 増補改訂版 (2004)