E-034

番組映像とクローズドキャプションの相関性を利用した番組生成モデルの検討

A Study on the Generative Model of TV programs Using Correlation of Video and Closed Captions

三浦菊佳 † Kikuka Miura　山田一郎 † Ichiro Yamada　松井淳 † ‡ Atsushi Matsui　住吉英樹 † Hideki Sumiyoshi　八木伸行 † Nobuyuki Yagi

1. はじめに

NHK アーカイブスの設立をはじめ、これまで放送されてきた番組映像を二次利用するサービスが注目されている。保存されている大量の映像資産から必要な映像を効率的に取り出し活用するためには、人手をかけずに自動で番組内容を解析する必要がある。クローズドキャプションは、番組の内容を説明するナレーションを文字化したものであり、映像と共に用いることでコンテンツ内容を解析する重要な素材になり得る。

我々はこれまで、クローズドキャプションから映像の被写体となる名詞や動詞を抽出して、クローズドキャプションの提示時刻に同期した映像区間の内容を「ライオンが食べているシーン」などと推定してきた[1][2]。しかし、この手法では、クローズドキャプションに、被写体もしくは被写体の動作を表す単語そのものが出現しない場合には対応できないため、周辺単語の分布も利用して推定する必要がある。

そこで本論文では、文書が与えられたときにどの映像が相応しいかを示す確率を番組生成モデルと呼び、その構築手法を提案する。以下、提案手法と実際のテレビ番組を対象とした実験とその考察について報告する。

2. 番組生成モデルの生成

番組映像と対応するクローズドキャプション中の単語には、一定の相関があると考えられる。例えば、料理番組において「包丁」「みじん切り」「まな板」などの言葉からはまな板上の素材アップの映像が、「フライパン」「強火」「炒める」などからはガスレンジの前で出演者が料理をしている映像があると想像できる。

提案手法では、クラスタリングした番組映像に対応するクローズドキャプションから、分類されたクラスごとの単語の出現分布を基にして、ある文書ベクトル $\boldsymbol{x}$ が与えられたときの、クラス c の出現確率 $P(c|\boldsymbol{x})$ を求める。図 1 にその概要を示す。$P(c|\boldsymbol{x})$ を計算するために、Naïve Bayes 分類器の多項モデル[3]を用いる。

文書ベクトル $\boldsymbol{x}$ は、映像を基に切り出されているため、複数文から構成されることが多く、その文書に含まれる単語群により構成される。

$$\boldsymbol{x} = \{w_1, w_2, w_3, \ldots, w_n\} \tag{1}$$

各単語の生起は独立と仮定し、あるクラス c が与えられたときの文書ベクトル $\boldsymbol{x}$ の生起確率 $P(\boldsymbol{x}|c)$ を、文書ベクトル $\boldsymbol{x}$ に含まれる単語を利用して(2)式により求める。

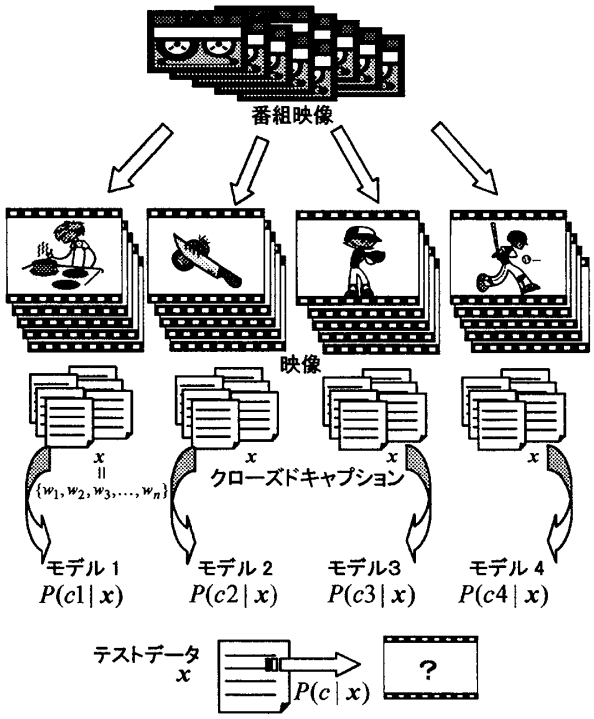

図 1.　番組生成モデル構築概要

$$P(\boldsymbol{x}|c) = P(|\boldsymbol{x}|)|\boldsymbol{x}|!\prod_i \frac{P(w_i|c)^{N(i,\boldsymbol{x})}}{N(i,\boldsymbol{x})!} \tag{2}$$

$N(i,\boldsymbol{x})$ は文書 $\boldsymbol{x}$ 内での単語 w_i の頻度を表す。$P(w_i|c)$ は、クラス c における単語 w_i の生起確率を示し、(3)式により算出する。

$$P(w_i|c) = \frac{\sum_{\boldsymbol{x}\in c} N(i,\boldsymbol{x}) + \delta}{\sum_i \sum_{\boldsymbol{x}\in c} N(i,\boldsymbol{x}) + \delta|V|} \tag{3}$$

ゼロ頻度問題を考慮して Laplace estimation によるスムージングを行い、そのパラメータとしてスムージング係数 δ を用いる。V はクラス c 中の単語の種類数を表す。

$P(c|\boldsymbol{x})$ は、ベイズの定理より以下の式で求められる。

$$P(c|\boldsymbol{x}) = \frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})} = \frac{P(c)P(\boldsymbol{x}|c)}{\sum_c P(c)P(\boldsymbol{x}|c)}$$

† NHK　放送技術研究所
‡ 早稲田大学理工学術院

$$= \frac{P(c)P(|x|)|x|!\prod_i \frac{P(w_i|c)^{N(i,x)}}{N(i,x)!}}{\sum_c P(c)P(|x|)|x|!\prod_i \frac{P(w_i|c)^{N(i,x)}}{N(i,x)!}}$$

$$= \frac{P(c)\prod_i P(w_i|c)^{N(i,x)}}{\sum_c P(c)\prod_i P(w_i|c)^{N(i,x)}}$$

$$\propto P(c)\prod_i P(w_i|c)^{N(i,x)} \quad (4)$$

(4)式の値により、テストデータとして文書ベクトル $\boldsymbol{x}$ が与えられたときの各クラスの生起確率が算出される。最終的に、(5)式により、最も高い値をとるクラスを文書ベクトル $\boldsymbol{x}$ に相応しい映像として選択する。

$$\hat{C} = \arg\max_C P(c)\prod_i P(w_i|c)^{N(i,x)} \quad (5)$$

3. 番組生成モデル構築実験

手法の有効性を評価するために、NHK の番組「おしゃれ工房」のメイクをテーマに扱う 12 番組を対象とした番組生成モデルの構築実験を行った。映像のクラスタリングでは、正面顔アップショット(C1)、正面顔バストショット(C2)、片目アップショット(C3)、片目と鼻全体が入るショット(C4)、それ以外(C5)の合計 5 クラスに手動で分類した。各ショットのイメージを図 2 に示す。なお、映像の特徴を用いたクラスタリング技術には様々な手法が提案されてきており[4]、特に人間の顔に関しては、高い精度で取り出すことができる[5][6]ため、この過程は将来的に自動で行えると思われる。

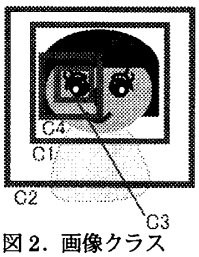

図 2．画像クラス

対象とする 12 番組のうち、11 番組を学習データ、残りの 1 番組をテストデータとして合計 12 回のクロスバリデーションによる評価を行った。今回は映像が属するクラスの変化点は既知として、クラスが変化するところまでをひとかたまりとした文書をテストデータとして入力しとし判定し、スムージング係数 δ は 0.04 とした。文書ベクトル生成処理では、対象とする単語を助詞、助動詞、記号を除くすべての単語とし、形態素解析には茶筌[7]を用いた。

構築したモデルにより判定された映像クラスの結果を実際の映像のクラスと比較して評価した結果を表 1 に示す。C1 から C4 までの適合率の平均は 39.7%、再現率の平均は 36.4%と改善の余地が残される。

表 1．実験結果

クラス	適合率	再現率
C1	29.5% (18/61)	22.5% (18/80)
C2	40.9% (27/66)	30.7% (27/88)
C3	60.0% (48/80)	64.9% (48/74)
C4	24.3% (17/70)	28.3% (17/60)
C5	61.2% (134/219)	69.1% (134/194)

片目アップショットの C3 は、「アイシャドウ」「アイライン」といった目元に扱う化粧品名や、「まゆ」「まつげ」「目尻」といったパーツの名称、「ぼかす」「引く」といった特定の動作を表す単語がこのクラスにおいて特徴的に出現していたため、比較的良い結果が得られた。C1、C2 では、「ファンデーション」など顔全体に塗るタイプの化粧品名や、「ツヤ」「ハイライト」といった肌の質感や顔の陰影を表す単語、「内側」「外側」といったおおまかな位置を表す単語が他クラスに比べて頻出しており、顔全体を写すときの特徴的な単語と考えられる。「チーク」「パウダー」などの単語は C1 よりも C2 での確率が高く、このような単語が使用される映像は人物全体の印象を捉えながら顔に粉をのせていくシーンであることがわかる。

判定結果には C1 と C2、C3 と C4 の分類誤りが多く、今後、特徴として使用する単語の選定や、構文的な特徴などの検討も行っていく必要がある。

4. まとめ

本論文では、番組映像とクローズドキャプションの相関性を利用した番組生成モデルを提案し、台本などの文書が与えられたときに相応しい映像ショットを推定する手法について検討した。映像により分類されたクラスごとにクローズドキャプションの語彙情報から Naïve Bayes 分類器によるモデルを構築し、12 番組を対象として実験を行った結果、一定の分別能力があることを示した。

今後、扱う単語の選別や、単語以外の特徴の検討、スムージング係数の推定などにより精度の向上を図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] 三浦菊佳, 山田一郎, 住吉英樹, 八木伸行, 奥村学, 徳永健伸 :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, 第 5 回情報科学技術フォーラム一般講演論文集第 2 分冊, E_017, pp.179-180 (2006)

[3] Andrew McCallum, Kamal Nigam : A comparison of event models for naive bayes text classication, In Proceedings of AAAI-98 Workshop on Learning for Text Categorization, pages 41-48 (1998)

[4] Richard O. Duda, Peter E. Hart, David G.Strok : Pattern Classification, Wiley-Interscience (2000)

[5] Stan Z. Li, Anil K. Jain : Handbook of Face Recognition, Springer (2005)

[6] 松井淳, Simon Clippingdale, 松本隆 :ベイズ的動画像顔検出における顔候補領域の逐次予測, 第 6 回情報科学技術フォーラム一般講演論文集 3K_5 (2007)

[7] 形態素解析システム茶筌 http://chasen.naist.jp/hiki/ChaSen/