I_065

AdaBoostを利用したスポット候補映像区間の抽出手法

A Method to Extract Video Sections for a TV Program Trailer using AdaBoost

河合 吉彦 †　山田 一郎 †　住吉 英樹 †　八木 伸行 †
Yoshihiko Kawai　Ichiro Yamada　Hideki Sumiyoshi　Nobuyuki Yagi

1 まえがき

大量の映像データから，目的の映像を効率的に探し出すための有効な技術のひとつとして映像の要約がある．放送局に蓄えられた大量の番組アーカイブ映像に対しても，番組スポットのような要約映像を付加しておくことができれば，見たい番組を選び出す際の有効な手がかりとなり，映像資産のより有効な活用が可能になると考えられる．しかしながら，すべての映像データに対して人手で番組スポットを作成することは困難であるため，それらを自動生成する手法が必要である．

我々は，電子番組表（EPG:Electronic Program Guide）に記載される番組紹介文（EPG テキストと呼ぶ）を利用した番組スポットの自動生成を検討している [1]．提案手法では，次のような考え方に基づいた番組スポットの生成を試みる．EPG テキストは，番組内の魅力的なシーンや見どころとなるシーンについて，その内容を紹介する文字情報である．EPG テキストでは，具体的な数値を提示するなど，映像内容を魅力的に伝えるための特徴的な文章表現が使用されている．そこで提案手法では，ナレーションの書きおこしであるクローズドキャプション（CC: Closed Caption）において EPG テキストのような表現が用いられる箇所には，魅力的なシーンが含まれているという仮定に基づき，スポットで利用する候補映像の抽出を試みる．抽出した映像は，カット長の調整や並び替えなどの編集を加え番組スポットとする．

本稿では，その第一段階として取り組んでいるスポット候補映像区間の抽出について述べる．提案手法では，まず学習用に収集した EPG テキストから AdaBoost [2] によって EPG 文の特徴を学習する．次に，学習結果に基づいて番組 CC の中から EPG 文と同様の特徴を持つ CC 文を抽出する．最後に，対応する映像区間を，CC に記述されたタイムコード情報に基づいて決定する．

2 AdaBoost を利用したスポット候補抽出

2.1 手法の概要

提案手法の概要を図 1 に示す．学習手続きでは，まず収集した EPG 文と CC 文に対して形態素解析と固有表現抽出を実施する．次に AdaBoost を利用して，EPG 文と CC 文を分類するための弱識別器を学習する．抽出手続きでは，学習された弱識別器を用いて，番組 CC 内の各文がどの程度 EPG 文らしいかを判定する．EPG 文らしさを表す値としては，各弱識別器に付与されている信頼度の和を用いる．番組 CC 内のすべての CC 文について EPG 文らしさを算出し，値が高いものから順に決められた文数だけを選択する．最後に，選択された文に対応する映像区間を決定し，これらを番組スポットの候補映像区間とする．

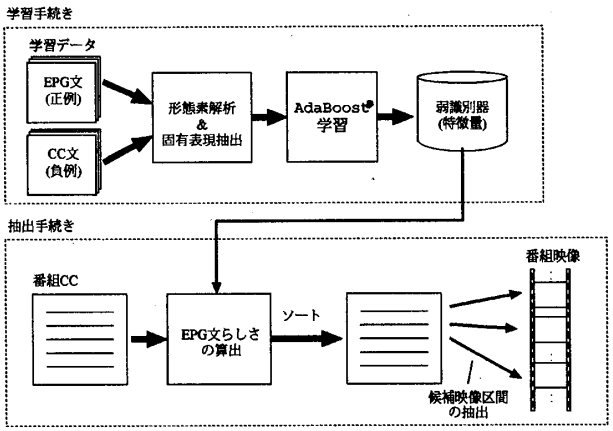

図 1　提案手法の概要

2.2 AdaBoost による学習

学習の目的は，EPG 文とそれ以外の文を分類する識別器を得ることである．学習データには，正例として EPG 文を，負例として番組 CC 文を使用する．なお，番組 CC には正例と識別されるべき文が含まれる可能性があるが，その割合は非常に小さいため影響はほとんどないと考える．識別には以下の特徴量を使用する．これらの特徴量は，特徴抽出の容易性と学習処理における計算負荷を考慮して選択した．

(a) 形態素数が閾値以上，もしくは閾値以下
(b) ある品詞が含まれる，もしくは含まれない
(c) ある索引語が含まれる，もしくは含まれない
(d) ある固有表現が含まれる，もしくは含まれない

ここで索引語は形態素と品詞の組とする．また，固有表現として IREX で定義 [3] された組織名，人名，地名，固有物名，日付表現，時間表現，金額表現，割合表現の 8 種類を使用する．

学習アルゴリズムを表 1 に示す．AdaBoost は，精度の低い弱識別器を複数組み合わせることにより，全体として精度の高い識別器を構築するものである．表 1 は T 個の弱識別器を組み合わせる場合である．ループ内の処理について説明する．まず，学習データの重み $w_{t,i}$ を正規化した後，すべての弱識別器 $h_j(x)$ について識別性能 ε_j を算出する．ε_j は誤識別された学習データの重み和で表される．また $h_j(x)$ は以下のように定義される．

$$h_j(x) = \begin{cases} 1 & if \quad s_j f_j(x) > s_j \theta_j \\ 0 & otherwise \end{cases} \tag{1}$$

$f_j(x)$ は j 番目の特徴量を表し，θ_j は閾値を表す．s_j は不等号の向きを制御するための値で $s_j = \{-1, 1\}$ である．

† NHK 放送技術研究所

表1 AdaBoost学習アルゴリズム

- 学習データ:$(x_1, y_1), ..., (x_n, y_n)$,
  x_i が正例なら $y_i = 1$，負例なら $y_i = 0$
- 弱識別器の候補:$h_1, ..., h_m$
- 重みを初期化，$w_{1,i} \leftarrow 1/n$
- $t = 1$ から T まで
  1. $w_{t,i}$を正規化，$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$
  2. $\varepsilon_j = \sum_i w_i |h_j(x_i) - y_i|$, $(j = 1..m)$ を計算
  3. ε_j の最小値 ε_t と弱識別器 h_t を選択
  4. $w_{t+1,i} \leftarrow w_{t,i} \varepsilon^{\alpha_t |h_t(x_i) - y_i|}$，$\alpha_t = \log \frac{1-\varepsilon_t}{\varepsilon_t}$

表2 学習された弱識別器の例

f_t	s_t	θ_t	α_t
名詞–固有名詞–人名–名	1	0	0.82
ます，助動詞	-1	1	0.68
です，助動詞	-1	1	0.67
名詞–サ変接続	1	0	0.63
形態素数	-1	11	0.59
助詞–格助詞–一般	1	0	0.46
地名表現	1	0	0.42
:	:	:	:
迫る，動詞–自立	1	0	0.04
なぜ，副詞–助詞類接続	1	0	0.03
テーマ，名詞–一般	1	0	0.02
:	:	:	:

次に，弱識別器の中から，ε_j が最小となる $h_t(x)$ を選択する．最後に，$h_t(x)$ が誤識別する学習データの重みを大きくした後，次の弱識別器の探索を実施する．

最終的な識別器 $h(x)$ は以下のように表される．

$$h(x) = \frac{\sum_{t=1}^{T} \alpha_t h_t(x)}{\sum_{t=1}^{T} \alpha_t} \quad (2)$$

α_t は，弱識別器 h_t の識別性能 ε_t から算出でき，h_t の信頼度を表す．提案手法による識別器は，α_t の和により，入力された文 x が正例である度合いを出力する．

3 評価実験

提案手法を用いて，番組スポットの候補映像区間の抽出実験を実施した．識別器の学習には，正例として約500番組分のEPGテキストから取得した10,000文を，負例として約500番組分の番組CCから取得した200万文からランダムに選択した10,000文を使用した．なお，句点や括弧などの記号はEPG文の本質的な特徴ではないと判断し，これらを取り除いた上で学習に利用した．候補とした特徴量の総数は約23,000種類であり，その中から500個の弱識別器を学習した．

学習された弱識別器の例を表2に示す．人名や地名表現などがEPG文の特徴と学習されている．また，EPGでは「ます」「です」は含まれない場合が多いという特徴もある．その他には，「迫る」や「なぜ」，「テーマ」などの表現が学習されている．

実際に放送された45分の自然番組のCCに対して，

表3 EPG文らしさの判定結果の例

CC文	$h(x)$
脂肪やたんぱく質の多いアリのさなぎは子育て中に栄養が必要な母グマには欠かせない食べ物。	0.5682
明治から昭和の初めにかけての最盛期には銅の年間生産量は1万トンを超え日本一を誇りました。	0.5648
銅山の開発で壊滅的な打撃を受けた足尾の森。	0.5601
足尾の山で生き物たちの不思議な行動が観察されました。	0.5581
足尾の山では草地が少しずつ回復するとともに生き物たちの姿も見られるようになりました。	0.5554
:	:
おいしそうですね。	0.3815
どうしてこのようになってしまったのでしょうか？。	0.3502

EPGらしさ $h(x)$ を算出した．番組のテーマは足尾銅山である．$h(x)$ が高いものから順に並び替えた結果を表3に示す．「最盛期」「壊滅的」などの表現や，地名表現，数値表現などの弱識別器によって，EPGテキストで比較的よく見られる表現の文が上位に配された．

通常，テレビ放送における番組スポットは15秒から30秒程度で構成される．そこで，映像の合計長が30秒となるよう表3の上位の文から順に対応する映像区間の抽出を試みた．その結果，上位5文に対応する合計27秒の映像が得られた．映像内容は，親子のクマ，足尾銅山の空撮，木が枯れた20年前の銅山，緑一面の現在の銅山などであった．実際に放送された番組スポットとカット単位で映像内容を比較したところ再現率60%（3/5），適合率50%（3/6）という結果が得られた．

4 あとがき

本稿では，EPGテキストが番組中の魅力的なシーンについて記述していることに着目し，番組CCの中からEPG文と同様の特徴を有する文を抽出することによって，番組映像から番組スポットのための映像区間を抽出する手法を提案した．EPG文が持つ特徴はAdaBoostによって学習した．評価実験では，EPGに比較的よく見られる表現の文を番組CC内から良好に抽出することができた．また，抽出されたCC文に対応する映像区間を調べたところ，スポットで使用するのに適切なシーンを抽出することができた．今後は，実験に使用する番組数を増やし，手法の有効性を詳しく検証したい．また抽出された映像の編集手法を検討したい．

参考文献

[1] 河合, 山田, 住吉, 八木, "EPGテキストとクローズドキャプション情報を利用した番組スポット候補映像区間の抽出手法," 信学技報, Vol.106, No.98, pp25-30, 2006

[2] Y.Freund and R.E.Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Computer and System Science, Vol.55, No.1, pp.119-139, 1996

[3] Information Retrieval and Extraction Exercise（IREX），"NEルール，定義（バージョン990214），" http://nlp.cs.nyu.edu/irex/NE/