H-013

映像とクローズドキャプションの解析に基づく自動番組要約に関する検討

A study of automated TV program summarization based on analysis of video and closed caption

河合 吉彦 †　住吉 英樹 †　藤井 真人 †
Yoshihiko Kawai　Hideki Sumiyoshi　Mahito Fujii

1 まえがき

映像要約は，所望の映像を効率的に検索するための有効な技術のひとつである．これまで我々は，電子番組表に記載される番組概要テキストと，番組音声の書き起こしであるクローズドキャプション（closed caption: CC）との対応付けに基づいて，テレビ番組を自動要約する手法を検討してきた [1]．しかし，この手法は，電子番組表が入手できない過去の放送番組などには適用できないという問題があった．また，電子番組表や CC といったテキスト解析のみに基づく手法であるため，映像における特徴が考慮されないという問題もあった．そこで本稿では，映像と CC の統合的な解析に基づいて，番組映像を自動要約する手法を提案する。電子番組表などの外部のデータは利用しない．提案手法では，番組映像の各ショットから算出した画像特徴量と，CC から抽出した形態素から，それらの出現頻度や分布を解析し，映像的にも言語的にも特徴的なシーンを抽出し連結することによって要約映像を生成する．実験では，実際の放送映像に対して提案手法を適用し，その有効性を検証する．

2 映像と CC の解析に基づく番組要約

提案手法では，映像と CC の解析に基づいて，番組を特徴付けるようなショットを抽出し，連結することによって要約映像を生成する．ここでショットとは，一台のカメラで連続的に撮影されたフレーム列を指す．

提案手法の概要を図 1 に示す．まず始めに，ショット境界検出処理によって番組映像をショット単位に分割する．次に，各ショットに対応する CC と映像のそれぞれを解析する．CC の解析処理では，まず形態素解析処理を実施し，自然言語である CC を形態素に分割する．その後，記号や助詞などの重要性の低い形態素を不要語リストを用いて取り除いた後，形態素の出現傾向に基づいてショットの重要度を算出する．もう一方の映像の解析処理では，カメラ動きに基づいて各ショットから複数のキーフレーム画像を抽出した後，キーフレームから視覚単語（visual word）と呼ばれる特徴量 [2] を抽出する．次に，CC と同様に，視覚単語の出現傾向に基づいてショットの重要度を算出する。最後に，CC と映像から算出された重要度を統合し，重要度の高いショットから順に連結していくことによって，目標とする映像長の要約映像を生成する．

以降では，キーフレームおよび視覚単語の抽出方法と，重要度の算出方法について詳細を説明する．なお，ショット境界検出処理，および形態素解析処理については，既存手法 [3, 4] がそのまま利用できるため，詳細は説明しない．

2.1 キーフレームおよび視覚単語の抽出

まず，ショットからのキーフレームの抽出手順について説明する．ショット内の全フレームを解析すると計算コストが非常に高くなるため，提案手法では，ショットから複数の代表フレームを抽出して解析する．ショット全体の特徴を捉えるため，本手法ではフレーム間の変化量に基づいてキーフレームの抽出位置を決定する．つまり，映像が激しく変化する区間では，キーフレームを抽出する時間間隔を短くし，映像に変化が少ない区間では抽出間隔を長くする．あるショットに含まれるフレーム列を $f_1, f_2, ..., f_n$ とし，このショットから合計 K 枚のキーフレームを抽出する場合を考える．このとき，式 (1) の条件を満たす最初のフレーム f_i を k 番目のキーフレームとして抽出する．

$$D(f_i) \geq \frac{D(f_n) * (k-1)}{K-1}, \quad (1 \leq k \leq K) \qquad (1)$$

ここで，$D(f_i)$ は，フレーム f_1 から f_i までの隣接フレーム間の差分累積和を表し，次式によって定義される．

$$D(f_i) = \frac{1}{HW} \sum_{j=1}^{i-1} \sum_{y=1}^{H} \sum_{x=1}^{W} |f_j(x,y) - f_{j+1}(x,y)| \qquad (2)$$

$f_j(x,y)$ は，フレーム f_j の座標 (x,y) における画素値を表し，H および W はフレーム画像の高さと幅を表す．

次に，抽出された K 枚のキーフレームから視覚単語を抽出する．具体的には，SURF（speeded up robust features）[5] を利用して，各キーフレームから特徴点を検出し，特徴点周辺の画像領域から特徴記述子を算出する．この特徴記述子を視覚単語として利用する．ただし，算出された特徴記述子をそのまま利用すると非常にスパースな分布となってしまうため，番組全体で特徴記述子をクラスタリングし，視覚単語の種類数をある程度限定する．特徴記述子のクラスタリングには k 平均法を利用する．

2.2 ショットの重要度の算出

映像の視覚単語に基づく重要度と，CC の形態素に基づく重要度の算出方法について説明する．いずれの重要度も同一の計算式を利用して算出する．情報検索においては，語の出現頻度（term frequency: TF）と，語が出現したショットの割合（inverse dcument frequency: IDF）に基づいて重要度を算出する TF-IDF 法が広く利用されている．同様のアイデアに基づいて，番組の i 番

† NHK 放送技術研究所

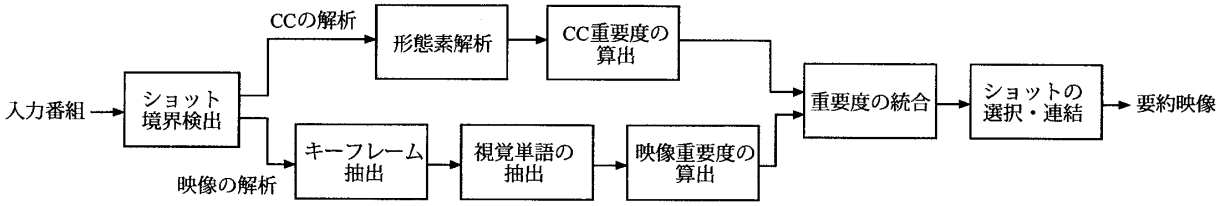

図 1　提案手法の概要

目のショットにおける語 w のスコア $\sigma_{i,w}$ を，以下の式 [6] を利用して算出する．

$$\sigma_{i,w} = \frac{(k_1+1)n_w}{k_1\left(1-k_2+k_2\frac{L_i}{AL}\right)+n_w} \cdot S_w \qquad (3)$$

ここで，語 w は視覚単語，あるいは形態素に対応する．n_w は番組における w の出現頻度を表す．また，L_i は i 番目のショットに出現する語の総数を表し，AL は番組の各ショットにおける語の平均数を表す．k_1 および k_2 は調整のためのパラメータである．式 (3) における S_w の項は次のように算出される．

$$S_w = \log\sum_{k=1}^{M} n_{k,w} - \sum_{i=1}^{M} \frac{n_{i,w}}{\sum_{j=1}^{M} n_{j,w}} \log \frac{n_{i,w}}{\sum_{j=1}^{M} n_{j,w}} \qquad (4)$$

M は番組におけるショットの総数を表し，$n_{k,w}$ は k 番目のショットにおける語 w の出現頻度を表す．S_w は，語 w のエントロピーの増減を反転した値を表している．

式 (3) における k_1 は，語の出現頻度 n_w の変化に対する $\sigma_{i,w}$ の感度を調整するパラメータであり，値を小さくするほど変化に鈍感となる．また，k_2 は正規化の程度を調整するパラメータであり，値を小さくするほど正規化の影響が減少する．提案手法では，文献 [6] の予備実験結果を参考に，$k_1 = 2.0$，$k_2 = 0.75$ に設定する．

次に，重要度の統合について説明する．i 番目のショットの重要度 σ_i は，視覚単語に基づく重要度 $\sigma_{i,w}^{vw}$ と，形態素に基づく重要度 $\sigma_{i,w}^{cc}$ の重み付き和によって定義する．

$$\sigma_i = \frac{a_1 \sum_{w \in shot_i^{vw}} \sigma_{i,w}^{vw} + a_2 \sum_{w \in shot_i^{cc}} \sigma_{i,w}^{cc}}{a_1 + a_2} \qquad (5)$$

最後に，算出された重要度に基づいてショットを選択し連結していく．ショット長が長すぎる場合は，ショットの前後を切り取ることにより，適切な長さに調整する．本手法では，実際に放送された番組スポット映像の解析結果を参考に，各ショットの長さを最長で 3 秒とした．算出された重要度が高いショットから順に選択，連結していくことにより，目的の長さの要約映像を生成する．

3　実験

実際の放送番組を対象に実験を実施した．実験には，自然ドキュメンタリ番組「地球・ふしぎ大自然 ナミブ砂漠 歩き続けるキリンたち」を利用した．番組の長さは 43 分であり，フレームサイズは 320×240 ピクセルである．本実験では，各ショットのキーフレーム数を $K = 5$，式 (5) におけるパラメータを $a_1 = 0.5$，$a_2 = 0.5$ に設定した．また，生成する要約映像の長さは 30 秒とした．

実験の結果，提案手法によって 30 秒の要約映像が自動的に生成された．出力された要約映像の内容は，砂漠のロングショットから始まり，砂漠を歩くキリンの群れ，水を飲む様子などとなっており，番組の概要がある程度理解できる内容となっていた．しかし，説明用のコンピュータグラフィックス（CG）などの不要と思われるシーンも含まれていた．また，実際に放送された 30 秒の放送スポットの映像内容を調査したところ，砂漠の空撮ショット，砂漠を歩くキリンの群れ，首で争う様子などとなっており，提案手法による要約映像と概ね類似した映像内容となっていた．

4　あとがき

本稿では，番組映像とクローズドキャプションの解析に基づいて，テレビ番組を自動的に要約する手法を提案した．提案手法では，映像における視覚単語とクローズドキャプションにおける形態素の出現傾向から各ショットの重要度を算出し，要約映像に使用するためのショットを選択した．これにより，映像的にも言語的にも特徴的なシーンを選択した．実際の放送映像に対する実験では，提案手法によって 43 分の番組映像から 30 秒の要約映像が自動生成されることを確認した．今後は，様々な番組に対する実験を実施し，手法の有効性を定量的に検証したい．また，映像とクローズドキャプションの統合の有効性を検証したい．

参考文献

[1] 河合, 住吉, 八木, “電子番組表における紹介文を利用した番組紹介映像の自動生成手法,” 信学論 (D), vol. J91-D, no. 8, pp. 2157–2165, 2008.

[2] G. Csurka, C. Bray, C. Dance and L. Fan, “Visual categorization with bags of keypoints,” in Proc. ECCV Workshop on Statistical Learning in Computer Vision, pp. 59–74, 2004.

[3] 河合, 住吉, 八木, “逐次的な特徴算出によるディゾルブ, フェードを含むショット境界の高速検出手法”, 信学論 (D), vol. J91-D, no. 10, pp. 2529–2539, 2008.

[4] T. Kudo and Y. Matsumoto, “Japanese dependency analysis using cascaded chunking”, in Proc. Conf. Natural Language Learning, pp. 63–69, 2002.

[5] H. Bay, T. Tuytelaars and L.V. Gool, “SURF: Speeded up robust features,” in Proc. ECCV, vol. 3951, pp. 404–417, 2006.

[6] K.S. Jones, S. Walker and S.E. Robertson, “A probabilistic model of information retrieval: development and status,” Inform. Process. Manag., vol. 36, no. 6, pp. 779–840, 2000.