G_010

# 映像コンテンツと関連文書の連携によるシーン検索システム

## An audio-visual information retrieval system using related text documents

寺尾 真† 越仲 孝文† 安藤 真一† 磯谷 亮輔† 奥村 明俊†
Makoto TERAO Takafumi KOSHINAKA Shinichi ANDO Ryosuke ISOTANI Akitoshi OKUMURA

## 1. はじめに

近年、大量の映像コンテンツが流通するようになり、それらを効率的に閲覧するための検索技術の重要性が増している。映像コンテンツを検索する方法の一つとして、音声認識によるインデキシングが挙げられるが、誤認識や未知語の存在が検索精度の低下につながっている。

従来、このような問題を解決するための研究がいくつか行われている[1][2]。[1]では、未知語を検索するために、検索クエリと認識結果とを音素などのサブワード単位のDP マッチングによって照合している。[2]では、固有名詞の認識誤りを訂正するために WEB 文書を利用している。すなわち、認識結果中のいくつかの単語を用いて検索された WEB 文書から固有名詞を抽出し、認識結果の低信頼度箇所と音響的に照合することにより、認識結果を修正する。これらの手法に共通している点は、ある単語が音声データに含まれているかどうかを音響的に判定している点である。そのため、雑音下やくだけた発声のように音響的に認識困難な場合に精度が低下する。また、音響的な照合を行うには単語の読み情報を得る必要がある。

そこで本稿では、クエリが音声認識の未知語であっても音声データとの音響的な照合をすることなく、検索対象に関連する文書を用いた言語的な処理によって適切なシーンを検索可能な手法を提案し、その有効性を示す。

## 2. 映像コンテンツのシーン検索

一般に、映像コンテンツには様々なトピック(話題)が含まれているため、これらトピック単位でコンテンツを検索・閲覧できることが有用と考えられる。そこで我々は、トピックを単位としたシーン検索の実現を目指している。図1にその処理の流れを示す。まず、映像コンテンツ中の音声を認識し発話内容をテキスト化する。次に、得られた認識結果テキストに対して意味内容に基づいたテキスト分割を行うことで、映像コンテンツを複数の意味的にまとまったトピックに分割する[3]。最後に、検索クエリと認識結果とを照合してトピックを検索する。本稿では、このようなトピック単位の検索を実現するためには必ずしもクエリと音声とを直接照合する必要がない点に注目し、あるトピック全体の中にクエリが含まれているかどうかを判定する手法を提案する。

## 3. 関連文書を用いたシーン検索

提案法では、検索対象となる映像コンテンツに関連する文書群 $D$ を用意する。そして、クエリ $q$ に対するトピック $i$ の検索スコア $score(q,i)$ を次のように計算する。

$$score(q,i) = \alpha \times s(q,i) + \sum_{j \in D} f(i,j) \times s(q,j) \quad (1)$$

† NEC メディア情報研究所

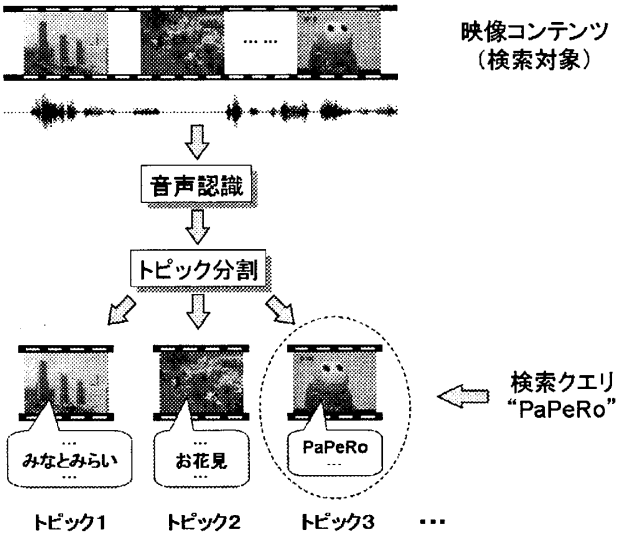



図1. トピックを単位としたシーン検索システム

ここで、$s(q,i)$ はクエリ $q$ とトピック $i$ の認識結果テキストとを照合したスコア、$s(q,j)$ はクエリ $q$ と関連文書 $j$ とを照合したスコアである。これらはテキスト中のクエリの出現数等を考慮して計算しても良いが、本稿では単純にクエリが 1 回でも出現したら $s$=1、出現しなければ $s$=0 とした。また、$f(i,j)$ は関連文書 $j$ のスコアをどれだけ重視するかを表す。本稿では次のように定義した。

$$f(i,j) = \begin{cases} d_{ij} & \cdots \ d_{ij} \geq t \\ 0 & \cdots \ d_{ij} < t \end{cases} \quad (2)$$

ここで、$d_{ij}$ は認識結果テキスト $i$ と関連文書 $j$ とのコサイン類似度、$t$ は認識結果にどれだけ類似した文書まで考慮するかを定める閾値を表す。すなわち提案法では、トピック $i$ の検索スコアを計算する際に、クエリ $q$ に対する関連文書 $j$ のスコアを、トピック $i$ と文書 $j$ との類似度に応じて加算する。なお、$\alpha$ はクエリに対するトピック $i$ 自身のスコアをどれだけ重視するかを表す重みである。

図2で、提案法が誤認識による検索精度の低下を抑える仕組みを説明する。図2(a)は、PaPeRo というロボットに関するトピックの認識結果から “PaPeRo” が脱落した場合である。結果、クエリ “PaPeRo” で検索した場合、式(1)の第1項は0になるが、認識結果に類似した文書は PaPeRo に関連した文書であり文書中に “PaPeRo” が出現する可能性が高いと考えられるため、式(1)の第 2 項が大きな値となり、このトピックは検索結果の上位に順位付けられる。このとき、認識結果により類似した文書に “PaPeRo” が出現するほどより高いスコアになることで、正確なスコア付けが可能となっている。また、図2(b)はこれとは逆に PaPeRo と無関係なトピックの認識結果に “PaPeRo” が湧き出した場合である。結果、式(1)の第 1 項でスコアがつく

が、認識結果に類似した文書中に“PaPeRo”が出現する可能性は低いため第 2 項は小さな値となり、このトピックは検索結果の上位には現れない。

## 4. 評価実験

提案法を用いた検索精度の評価を行う。評価データはニュース番組 48 時間分とし、正解のトピック境界を用いて 1215 個のトピックに分割して検索対象とした。これらには BGM 等の雑音やアナウンサー以外の自由発声が比較的多く含まれており、平均の音声認識率は約 66%だった。関連文書には同時期の新聞記事 3316 記事を用いた。検索クエリはこれらのデータに出現する単語の中から選び、未知語を 7 種類(出現するトピック数は 85 個)、既知語を 29 種類(605 個)とした。また、認識結果と関連文書のコサイン類似度は自立語を用いて計算し、$\alpha$ は実験的に 3 とした。検索されたトピック中でクエリが発話されている場合に検索成功とし、式(2)の類似度の閾値 $t$ を 0.05～0.25 まで変化させて評価した。

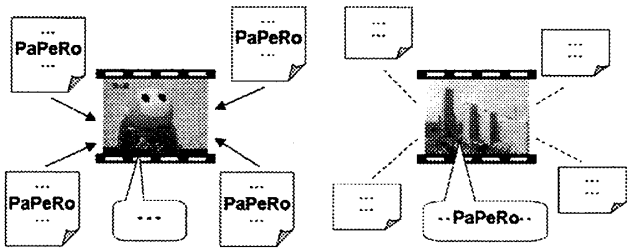
まず、未知語によるシーン検索精度を図 3 に示す。各閾値ごとに検索結果の上位 1, 2, 3, 5, 7, 10, 15, 20, 25, 30, 40, 50, 70, 100 位以内の再現率(横軸)、適合率(縦軸)をプロットしている。未知語であるため認識結果には現れずそのままでは検索できないが、提案法により閾値 0.1～0.15 で再現率 50%において適合率 50%程度の結果が得られている。閾値を大きくする、すなわち認識結果により類似した文書のみを用いて式(1)の第 2 項を計算した場合、使用する類似文書の数が減るため再現率の上限が低下しているが、同じ再現率で比較した場合は適合率が改善する傾向がみられる。

次に、既知語によるシーン検索精度を図 4 に示す。図中に baseline としてプロットした点は、式(1)の第 1 項のみを用いた場合、すなわちクエリが認識結果に存在するトピックをそのまま検索結果とした場合の検索精度である。提案法では上位の検索結果は適合率が高く、適切な順位付けが出来ている。また、baseline と比較すると、同じ再現率において閾値が大きい場合にはほぼ同等の適合率が得られているが、未知語を検索可能な小さな閾値では適合率が低下している。提案法は既知語の誤認識に対しても効果があると期待し、実際、クエリによっては湧き出し誤りを抑制できた例も見られたが、一方で、クエリが発話されていないトピックを誤って検索してしまう影響も大きかったと考えられる。ただし、提案法が検索するトピックは、クエリ自体が発話されていなくてもクエリに関連した内容を含むと考えられ、この点についてどの程度満足できる検索結果が得られているかという観点での評価も今後必要である。

## 5. おわりに

映像コンテンツをトピック単位で検索するために、音声認識結果に類似する文書のスコアを類似度に応じて加算する手法を提案した。その結果、クエリと音声データとの音響的な照合をしなくても未知語による検索が可能であることを示した。

今後の課題としては、映像コンテンツのトピック単位への分割を自動で行った場合[3]の検索精度への影響、および認識精度と検索精度との関係等の調査が挙げられる。




(a) 脱落誤り　　(b) 湧き出し誤り

図 2. 誤認識に頑健なシーンの検索

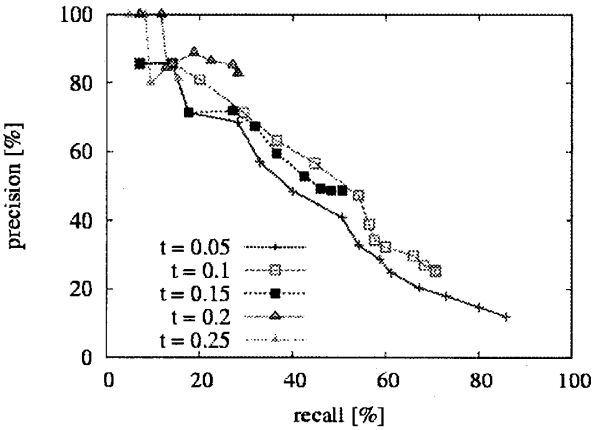



図 3. 未知語によるシーン検索精度

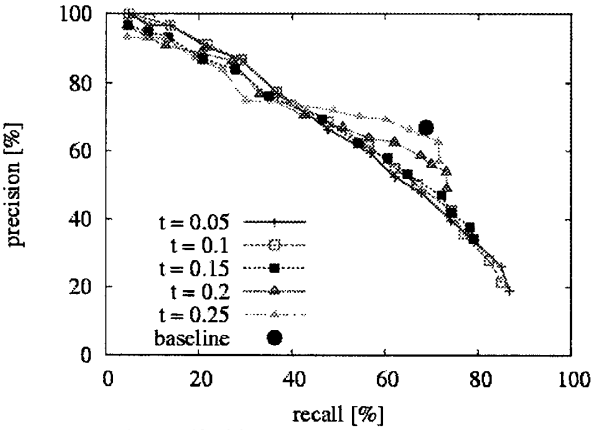



図 4. 既知語によるシーン検索精度

## 参考文献

[1] 大竹, 岩田, 伊藤, 小嶋, 石亀, 田中, 李, “スポッティング区間の再認識に基づく音声検索性能の向上,” 日本音響学会講演論文集, pp. 23-24, 2006.3.

[2] 西崎, 伊藤, 関口, 中川, “WEB 文書を利用した音声認識誤りの訂正方法の検討,” 情処研報 2003-SLP-49, pp.181-186, 2003.

[3] T.Koshinaka et al., “An HMM-based text segmentation method using variational Bayes approach and its application to LVCSR for broadcast news,” ICASSP2005, pp.485-488, 2005.