K-60　ニュース映像に対する発話内容と人物問い合わせシステム

Inquiry system of the utterance contents and unknown persons in news videos

井上 徹† 藤本 雅清† 山本 夏夫† 有木 康雄† 熊野 雅仁† 堂下 修司†
Toru Inoue Masakiyo Fujimoto Natsuo Yamamoto Yasuo Ariki Masahito Kumano Syuji Doshita

1. はじめに

近年、インターネットの普及により、我々は様々な情報源から情報を得ることが可能となった．しかし，このような環境においても，我々が情報を得る情報源としては，依然としてテレビニュース等に依るところが大きい．現在のテレビ放送は放送局側が一方的に視聴者へ情報を送っている．そのため，視聴者にとって知らない情報があった場合に、視聴者がテレビに直接問い合わせることで，情報を得ることが可能な対話型テレビの出現が望まれる．そこで，本稿では放送映像に対して対話検索を行うシステムの構築について述べる．

2. 発話内容問い合わせシステム

このシステムは，ニュース映像中に，ユーザーの知らない単語が出て来た場合に「○○って何ですか」と問い合わせることよって，その単語について検索を行うシステムである．

2.1 システム概要

このシステムは大きく図1に示すように，以下のようなモジュールから構成されている．

- ニュース映像制御部 (スクリーン 1)
- 音声入力部 (マイクロフォンアレイ)
- 音声認識部
- 検索結果表示部 (スクリーン 2)

マイクロフォンアレイ，スクリーン2枚を用いている．スクリーン1枚につき1台のマシンとマイクロフォンアレイに1台のマシンを接続し，合計3台のPCをTCP/IPで接続することによって各マシン間のデータのやり取りを行っている．

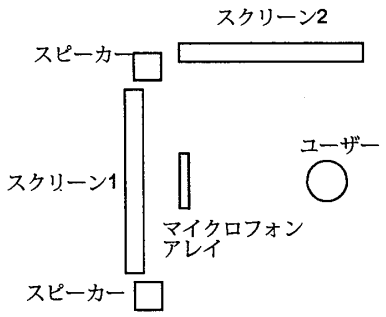

図 1: 発話内容問い合わせシステムの構成

†龍谷大学理工学部

2.2 システムのフロー

システムのフローを図2に示す．システムはスクリーン1上でニュース映像の再生や停止，再開を行っいる．ユーザーの発話はマイクロフォンアレイに入力され，音声認識が行われる．認識結果があらかじめ登録しておいたキーワード辞書に含まれる場合，ニュース映像を停止する．含まれない場合は，ビデオの再生を続ける．ニュース映像を停止した場合，サーバーはマイクロソフトエージェントを表示し，「検索結果は○○です．右のスクリーンを御覧下さい」と発話する．スクリーン2にはWebから音声認識した結果で検索したWebページを表示する．ユーザーが続きを見たい場合，「ビデオの続きを見せて下さい」と発話することにより，音声入力，音声認識が同様の手順で行われ，ニュース映像が再開される．

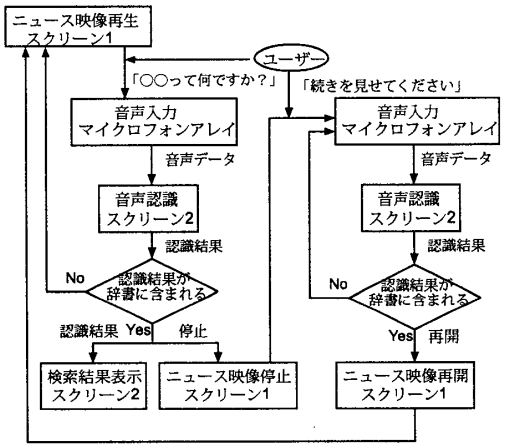

図 2: 発話内容問い合わせシステムのフロー

3. 顔問い合わせシステム

このシステムは，ニュース映像中に，ユーザーの知らない人物が出て来た場合に，「この人は誰ですか」と問い合わせることによって，問い合わせた人物について検索を行うシステムである．

3.1 システム概要

このシステムは大きく，図3に示すように以下のようなモジュールで構成されている．

- ニュース映像制御部 (スクリーン 1)
- 音声入力部 (マイクロフォンアレイ)
- 音声認識部
- 指指示部 (モーショントラッカー)
- 顔切り出し部
- 顔認識部

- 検索結果表示部

マイクロフォンアレイ，モーショントラッカー，スクリーンを用いたものである．スクリーンに1台のマシンとマイクロフォンアレイに1台のマシン，モーショントラッカーに1台のマシン，顔切り出し・認識用のマシン1台を接続し，合計4台のPCをTCP/IPで接続することによって各マシン間のデータのやり取りを行っている．

図 3: 顔問い合わせシステム

3.2 システムのフロー

システムのフローを図4に示す．システムはスクリーン1上でニュース映像の再生や停止，切り出した顔の表示，並びに検索結果の表示を行っている．ユーザーの発話はマイクロフォンアレイに入力され，音声認識が行われる．あらかじめキーワード辞書にキーワードを登録しておき，認識結果がキーワード辞書に含まれる場合，キーワードを抽出する．また同時にユーザーが人物の顔を指で指し示すと，モーショントラッカーが指示座標を取り込む．一定時間静止すれば何らかの目標物を指したと判断する．この2つのキーワードと座標値が同時に発生した場合に，ニュース映像を停止する．その際に，停止したニュース映像の現フレームを切り出し，画像処理マシンで顔の切り出し，顔の認識を行う．切り出された顔画像並びに認識結果による検索結果がスクリーン1に表示される．

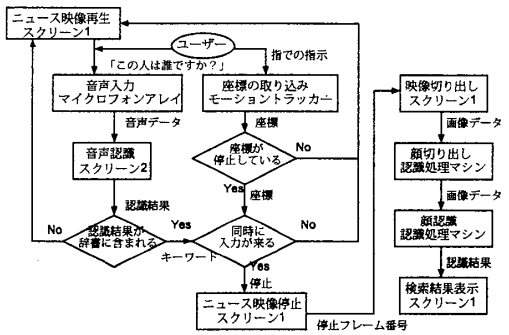

図 4: 顔問い合わせシステムのフロー

4. 実験

男性被験者5名により，NHKのニュース番組中に出現する10人の有名人に対して人物の問い合わせ実験を行った．1人の有名人に対して，それぞれ3回ずつ実験を行っている（合計150回試行）．顔認識は切り出された1枚の画像に，目標とする人物のみが出現するという前提で行っている．

4.1 ハンズフリー音声認識

本研究におけるハンズフリー音声認識では，話者方向を推定し，ビームフォーミングを行う．次に，話者方向の時間的安定性に基づいてユーザー発話区間を検出し，音声認識を行う [1]．音声認識における音響モデルには，話者独立な monophone HMM(5状態3ループ，各状態12混合分布，41音素) を用いた．HMMの学習には，日本音響学会新聞記事読み上げ音声コーパスのうち，男性話者137人分の21782発話を用いている．音声認識結果は，人物に関する質問であると判定された場合，つまり音声認識によるトリガーが発生すれば正解であるとした．実験の結果，93.33%(140/150) の音声認識結果が得られた．

4.2 顔認識

顔領域検出及び，顔認識を実現するために，主成分分析に基づく部分空間法 [1] を用いた．認識用の顔モデルは，NHKのニュース番組から切り出した有名人10人であり，各人物ごとに学習を行っている．学習データの数は各人物10枚画像を用いている．

まず，切り出された1枚画像に対する顔領域検出の精度について評価する．顔領域検出は，検出された領域に人物の目，鼻，口の全てが含まれていれば正解であるとした．実験の結果，顔領域検出精度は 60.00%(84/140) であった．また，明らかに顔とは異なる領域を切り出した事例数は2であった．

次に，切り出された顔領域画像に対する認識実験を行った．実験の結果，76.19%(64/84) の顔認識精度が得られた．顔認識精度の誤りの原因として，顔領域検出の場合と同様に，顔の方向の問題が挙げられる．また，認識用モデルは10枚の顔画像のみで学習していたため，学習量が十分でなかった．今後，大量学習データを用いてモデルを学習し，実験を行う必要がある．

5. まとめ

本稿では、対話型テレビの一つの機能として，ハンズフリー音声認識を用いた発話内容の検索方法と映像中に出現する人物を指先指示により情報検索する方法について検討を行った．被験者5名による実験の結果，音声認識結果 93.3%，顔領域検出精度 60.00%，顔認識率 76.19%が得られた．今後，顔領域検出，認識精度の改善について、モーショントラッカーを用いず指先検出を行い、指示座標の取り込みについて検討を行う予定である．

参考文献

[1] 藤本 雅清, 山本 夏夫, 有木 康雄, 熊野 雅仁: “マルチモーダルインタラクションによるニュース映像中の人物認識と検索” 人工知能学会研究会資料 SIG-SLUD-A201, pp.7-13, 2002-06