アクティブカメラを使った遠隔教育システムの試作*

2G－2

－画像と音声による発話者検出－

田中英治　倉立尚明　福井和広

(株) 東芝 関西研究所

1　はじめに

人間のコミュニケーション手段として、音声や文字などの他に相互理解のためにノンバーバルな情報も重要な役割を果たしている [1]。例えば、会話をしている時に相手が話し始める前にこちらを向くことで相手が話し出そうとしていることが分かる。このノンバーバルな情報の多くは、画像情報から得られる。会話の順番や発言権を調整する役目の動作は、「調整子」と呼ばれるが、コンピュータを用いたシステムにおいても、画像処理によってこの調整子を検出することで円滑な対話を支援できることが予想される。

本研究はATM回線を用いた遠隔教育実験†での講師と生徒のface-to-faceコミュニケーションの質を向上させることを目的とする。そのために、表示される相手の顔を見ながらマイクを持つという発話の前段階を画像で検出し、音声の検出と組み合わせて発話者を適切にズーム表示させるシステムを構築した。

2　遠隔教育におけるコミュニケーション

遠隔教育を円滑に行なうためには講師と生徒が実際に同じ場所にいるかのような自然さをシステムで実現する必要がある。例えば、会話の時に対面することは相手の表情などの様子を観察できるので会話の流れを自然にするために重要と考えられる。遠隔教育の講義中で、この自然さを実現するには、講師側の画面で、質問時には質問者をズームアップすることで対面している状況を作り出し、講義中は生徒全体の様子が分かるように全景を映すのが望ましいと考えられる。

本研究で用いたシステムでは、そのようなカメラワークを自動的に行なうために、カメラの向きやズームを計算機制御できるアクティブカメラと音声検出手段を用いて発話者の検出を行なう [2]。

3　発話者検出システム

3.1　システムの構成

システム全体の構成を、図1に示す。

アクティブカメラによって捉えた映像を講師側に送るとともに、グラフィクスワークステーションIndy(R4600 PC)で画像処理し、マイクと顔領域の抽出を行なう。マイク領域と顔領域の重心座標が画像中心になるようにSun SPARC Station 2でアクティブカメラを制御する。

*Trial experiment of the active-camera system for education between remote places –active detection of speaker by image and voice–

Eiji Tanaka , Takaaki Kuratate , Kazuhiro Fukui

TOSHIBA Kansai Research Laboratories

†NTTの「マルチメディア通信の共同利用実験」

マイクから入った音声は講師側に送られるとともに、音声検出を行ないズーム比率のコントロールを行なう。

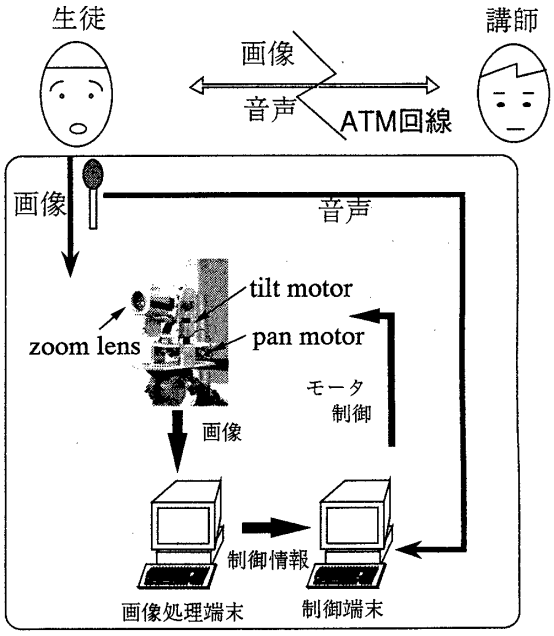

図 1: システム構成

自作したアクティブカメラは雲台に付けられたサーボモータによってパン (水平方向)、チルト (垂直方向) でカメラの向きを制御できる。またレンズにつけられたサーボモータによってズーム比率を制御できる。

3.2　発話者検出の手順

発話者はマイクを持って、講師の映っている画面に顔を向けて話し始めるので、このシステムでも同様の手順で発話者を検出する。まず、マイクのカバーの色を元にマイク領域を見つけ、その近傍で顔を検出する。次に音声の入力と同時に、顔領域とマイク領域のいずれもが画面に適切に収まるようにフレーミングしながらズームアップする。その手順を図示すると図2のようになる。

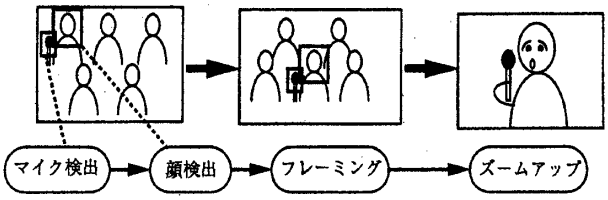

図 2: 発話者検出の手順

3.3 マイク領域の抽出

今回のシステムではマイクの抽出を容易にするためにマイクのカバーの色を予め決めておき、画像中でそのカバーと同じ色の領域のうち最大のものをマイクの領域として検出する。

3.4 顔領域の抽出

アクティブカメラで捉えた入力画像から顔領域を抽出するための手法として部分空間法を用いる。

まず、予め種々の顔画像を、各画素の輝度値で作られる画素空間に写像する。その分布に応じて、「顔」の特徴空間における固有ベクトルをKL展開によって求める。そうして得られた固有ベクトルの組を「顔」の特性を示す辞書として登録しておく。辞書を作るのに用いた画像の平均画像を図3中央に示す。

入力画像から任意の大きさと位置で抽出した画像と、辞書として登録された各固有ベクトルとの類似度が、予め設定したしきい値を超えた領域について、顔領域と判定する[3]。図3に示すような入力画像(左)から類似度の高い領域を抽出した結果が抽出画像(右)である。

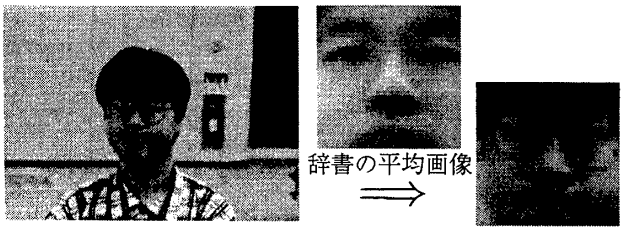

図 3: 顔領域の抽出

3.5 発話の検出

次に発話のタイミングを音声認識により判断してアクティブカメラをズームアップさせる。

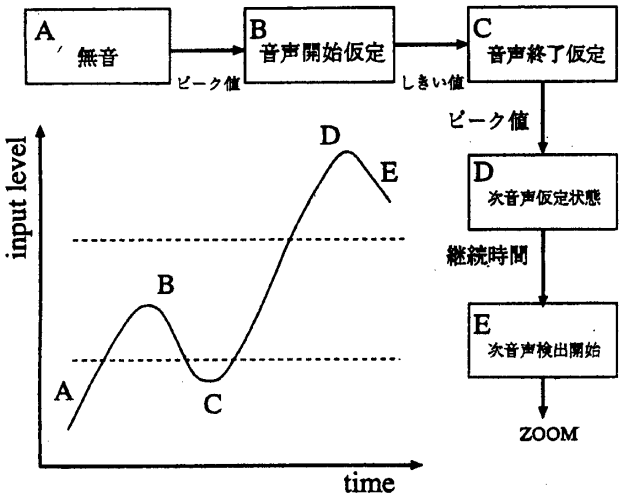

図 4: 発話検出

この発話検出では、音声の開始を判定し、カメラへズーム動作のトリガー信号を出力する。このため、音声の入力レベルの推移に応じて、図4に示すような処理を施し、最終的に2回目の音声入力の上昇を判定した時点で、ズーム動作へのトリガー信号を出力する。各状態は入力レベルのピーク値や継続時間が設定したしきい値を越えた時点で推移する。

音声が無くなって5秒を超えた時点で図2の右の図のようなズームアップ動作をやめ、図2の左の図のような会場全体を写す状態に戻る。

3.6 アクティブカメラの方向制御

質問者を注視した場合にカメラの視野に顔領域が収まるようにアクティブカメラの撮影方向を制御する。

入力画像中心を (x_c, y_c) とし、抽出した顔領域の重心を (x_g, y_g) とすると、カメラのpan, tilt方向の移動制御量 m_p, m_t およびカメラのズーム率 $(zoom)$ を以下のように定める。

$$m_p = f_p(zoom) \times |x_g - x_c| \quad (1)$$

$$m_t = f_t(zoom) \times |y_g - y_c| \quad (2)$$

$$zoom = f_z(scale) \quad (3)$$

f_p, f_t はズーム率 $(zoom)$、f_z は抽出した顔領域の大きさ $(scale)$ に関する1次関数である。

4 結果

カメラから約2mの場所でカメラの視野内に人間が着座している場合に、着座位置をランダムに移動させて100回の試行を行ない、その結果、質問者を検出した回数を表1に示す。ここで質問者を検出したというのは、ズームしてフレーミングした時に顔の輪郭が画像中に収まり、マイクのカバーも全体を検出した状態を指す。

マイクのみ	顔のみ	マイクと顔の両方
54	32	74

表 1: 発話者の検出回数

マイク領域と顔領域を併用することで発話者の検出率が向上することが分かった。

マイク領域は色のみに依存するので、同じ色の場所を誤検出する可能性があり、また顔領域は顔パタンに似た部分を誤検出する可能性があるが、発話者の含まれる画像ではお互いが近傍の領域にあるという仮定によって検出率を上げることができるためである。

この結果により当システムは発話者の検出に有効であり、遠隔教育を始めとする遠隔地どうしでの1対多あるいは多対多でのコミュニケーションでの対話性の向上を支援できることと考えられる。

5 今後の課題

誤検出をより少なくするために、今後の課題としては

- 顔領域の検出に色検出を併用する
- 顔領域の探索スケールを動的に変化させる
- 検出する色を入力画像のマイクの色によって適応的に決める

等の対策が考えられる。また、人物追跡の速度を向上させ、より実用的に使えるものとしていく予定である。

参考文献

[1] 黒川隆夫. ノンバーバルインタフェース. オーム社, 1994.
[2] 渡辺睦 福井和広. ヒューマンインタフェースにおける顔のセンシング. 電気学会, 1994.
[3] 飯島泰蔵. パターン認識理論. 森北出版, 1989.