# 話者方向推定およびパノラマ映像生成機能を持つテレビ会議用ビデオカメラの開発

上杉徹*，川村尚生，菅原一孔（鳥取大学）

Video Camera for Teleconferecing with Direction of Speaker Estimation and a Panoramic Image Generation Function

Toru Uesugi*, Takao Kawamura, Kazunori Sugahara (Tottori University)

**Abstract**

In most of the current teleconferencing systems, a conference room is taken in one fixed video camera. For that reason, a face image of speakers becomes small. To take both full view images of a conference room and face images of speakers, it is required to develop a multi camera system with human operations. In this paper, we develop a new video camera for a teleconferencing without human operations in above mentioned situations. The proposed video camera is constructed with 3 NTSC video cameras and outputs the images which composed with panoramic images generated with images from these video cameras and the facial images of speakers. In this paper, Active Contour Models for panoramic images is adopted and directions of speakers are estimated by voice and hardware realization the system is mentioned.

**キーワード**：話者方向推定，動的輪郭モデル，パノラマ映像
(Direction of Speaker Estimation，Active Contour Models，Panoramic image )

## 1. はじめに

最近では，ネットワークの通信速度の高速化や高性能の映像撮影装置の低価格化に伴い，テレビ会議が開催される場面が増えてきている．ところが一般のテレビ会議では，1 台のビデオカメラで会場全体を撮影するため，撮影された映像から話者の表情を読み取ることが困難である．また，逆に話者の表情を読み取ろうとして話者が大きく映るように撮影すると，会場の狭い範囲しか撮影できなくなるため会場の雰囲気を知ることが困難になるなどの問題がある．これらの問題を解決するため，複数台のビデオカメラで会場全体の映像と話者の映像を別々に撮影する場合もあるが，複数のビデオカメラを用意するにはコストがかかる上にその切替操作には人手がかかるなどの問題が出てくる．そこで，我々は会議会場を方向の異なる 3 つのビデオカメラで撮影し，会議会場全体の広範囲の映像を表示するためのパノラマ映像と，話者を映している映像をカメラ内部で合成したものを映像信号として出力するテレビ会議用ビデオカメラを開発している．このような装置を構成する際には，パノラマ映像の合成手法と話者方向推定手法とそれに伴ってパノラマ映像から複数の話者領域を抽出する手法が重要な機能として挙げられる．話者領域を抽出する方法としては，動的輪郭モデルを用い，抽出対象が複数であるため，従来の動的輪郭モデルの閉曲線を分裂させることにより，画像から複数の話者領域を抽出する．

本稿では，音声による話者方向推定処理をした結果として選ばれたカメラ画像に対して動的輪郭モデルを適用し，複数の顔画像領域を抽出する．そして，話者の拡大映像とパノラマ映像を表示する処理を FPGA を用いてハードウエア実現を試みた．

## 2. システムの構成

システムは，3 つの NTSC ビデオカメラを用いた映像撮影装置，NTSC ビデオデコーダ LSI，メモリ，デジタルビデオエンコーダを実装したビデオ信号入出力ボード，FPGA ボード，モニタから構成されている．構成している各装置を図 1 に示す．そして，FPGA の仕様を表 1 に示す．

表 1　FPGA の仕様
Table 1.　FPGA specifications

| | |
|---|---|
| 開発元 | ALTERA Co.Ltd. |
| 型番 | EP1C20F400C7 |
| ロジックエレメント数 | 20,060 |
| 最大 I/O ポート数 | 301 |
| パッケージ | 400-Pin FineLine BGA |
| サイズ | 21 × 21 [mm] |
| 動作周波数 | 48[MHz] |

メモリは 3 バンク実装している．そして，各バンクは 1 アドレスに 24 ビットデータの保存が可能であり，RGB 各 8 ビットの 1 画素分のカラーデータを 1 つのアドレスに保存することができる．そして，実時間で映像処理を行うために 3 つのバンクによるバンク切り替えを行っている．これにより，１つのバンクにアクセスが集中することで起こる処理の待ち時間を効率よく防ぐことができる．各部での 1 フレーム分の処理が終了するのに同期してバンク切り替

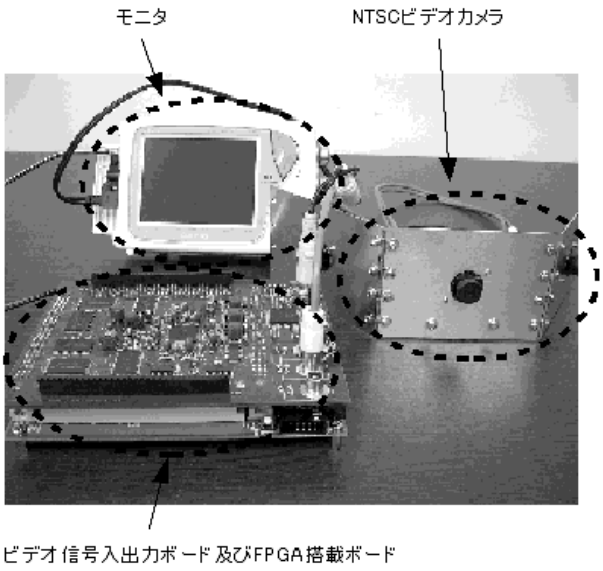


図 1　装置
Fig. 1.　System configuration

えを行う．バンク切り替えを行っている様子を図 2 に示す．

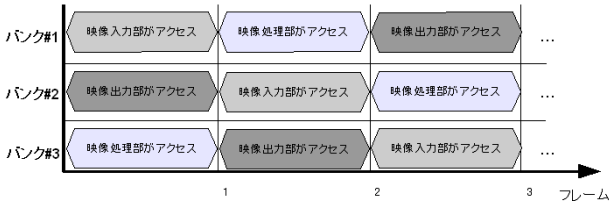


図 2　バンク切り替え
Fig. 2.　Bank exchanges

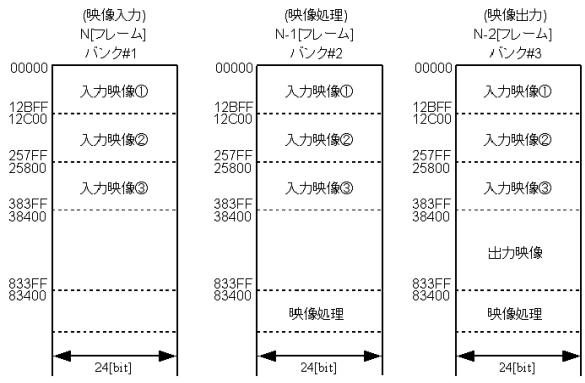


図 3　メモリマップ
Fig. 3.　Memory map

図 2 に示した処理の流れは以下のとおりである．バンク #1 には映像入力部がアクセスし，NTSC ビデオカメラから送られてきた NTSC アナログ映像信号をデコーダによってデジタルの RGB 各 8bit のデジタル映像信号に変換し，ビデオ信号入出力ボード上のメモリに蓄積する．同時に，バンク#2 には映像出力部がアクセスし，FPGA 上の各回路で処理を行ったデータをエンコーダで NTSC アナログ映像信号に変換しモニタに表示する．バンク#3 には映像処理部がアクセスし，FPGA 上の各回路においてメモリに蓄積したデータを基にパノラマ映像生成の処理を行う．そして，1 フレーム分の処理が全てのバンクで終了したら，次に，バンク#1 には映像処理部，バンク#2 には映像入力部，バンク#3 には映像出力部がアクセスし処理するというように，バンク切り替えを行いながら処理を行う．

それぞれのメモリバンクには，320×240[pixel] の撮影映像が 3 枚，および 640×480[pixel] の出力映像が 1 枚格納されており，それぞれの映像ごとに格納アドレスが定められている．メモリマップを図 3 に示す．

## 3.　パノラマ映像生成

パノラマ映像生成では，会議会場全体の広範囲の映像を表示するために，3 つの NTSC ビデオカメラで撮影した入力映像を基にパノラマ状の映像を生成する．しかし，撮影した入力映像の映像サイズ (640×480[pixel]) のままでは，パノラマ映像を生成する際，想定している出力映像の映像サイズ (640×480[pixel]) より大きくなり，パノラマ映像を生成することはできない．そこで，入力映像を 3/8 サイズの 240×180[pixel] に縮小し一部分を重ねて合成し，640×180[pixel] で表示することにした．

映像サイズの縮小方法として，撮影した入力映像をメモリに書き込む時に，1 画素ずつ飛ばして書き込むことで，入力映像の映像サイズを 1/2 に縮小する．さらに，入力映像を 3/4 サイズにするために，パノラマ映像を生成する際に，縮小前の映像から画素を縦横方向にそれぞれ 4 画素に 1 画素間引くという手法を用いた．またパノラマ映像の生成方法は，入力映像データをメモリの空き領域にコピーすることによって行われる．まず，映像サイズを縮小しながらメモリの決められたアドレスに 3 つの映像データの横 1 行分をコピーする．続いて 2 行目，3 行目とコピーするが，縦方向も 4 回に 1 回行を飛ばして縮小しながら全映像をコピーしてパノラマ映像を生成する．その様子を図 4 に示す．

本研究で開発したシステムでは，テレビ会議会場を 3 台の NTSC ビデオカメラで撮影した撮影映像から最適と思われる結合位置を自動的に検出する．これにもとづき，パノラマ映像を生成して，その映像をテレビ会議会場の全体映像としている．

## 4.　複数領域抽出手法

**〈4•1〉　動的輪郭モデルの動作点に働く 4 つの力**　動的輪郭モデルは仮想的な閉曲線上にある複数の動作点に圧力，引力，反力および振動項と呼ばれる 4 つの力が働くことにより，閉曲線が収縮し領域を抽出する手法である．しかし，この方法は 1 枚の画像から 2 人以上の顔領域を抽出するなどの，抽出対象が複数となる場合には有効ではない．

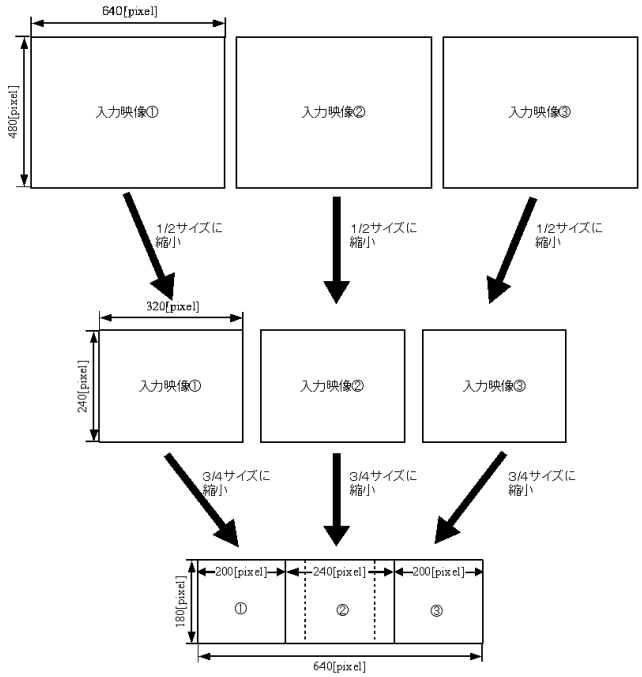


図 4　補正前のパノラマ映像を生成する様子
Fig. 4.　Panoramic image generation before correction

そこで，閉曲線を分裂させることにより画像から複数の特定領域を抽出する．図 5 に動作点に働く様々な力により閉曲線が収縮する様子を示す．

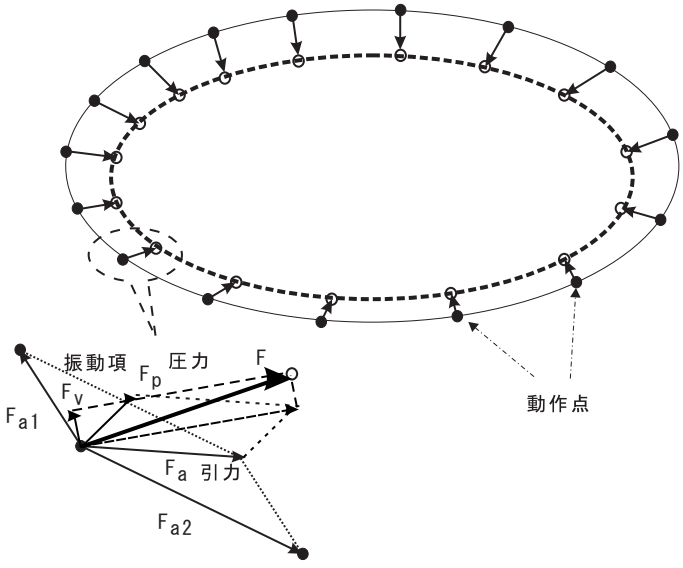


図 5　動的輪郭モデルによる閉曲線の収縮
Fig. 5.　Retraction of closed curve using active contour models

圧力 $F_p$ は，着目している動作点と隣り合う 2 つの動作点で作る角の 2 等分線方向に働く力であり，その大きさは一定の値 $K_p$ を持つ．引力 $F_a$ は，隣り合う 2 つの動作点間に働く力 $F_{a1}$ と $F_{a2}$ の合力であり，その間の距離に比例した大きさを持つ．引力については，〈4･2〉で詳しく説明する．振動項 $F_v$ は圧力と引力の合力に対し直角方向に働く力であり，収縮のたびにその方向を反転する．これにより，動作点は左右に振動しながら収縮動作を行うため，収縮途中でノイズに引っかかった場合でもノイズをよけて通ることができる．なお，この振動項の大きさは一定の値を持つ．また，反力 $F_r$ は図 6 に示すように，動作点が対象の画像領域に接した際に働く力であり，圧力 $F_p$，引力 $F_a$，振動項 $F_v$ の合力の抽出領域に対する法線方向成分を打ち消す働きを持つ．ただし，反力はその大きさに閾値を持ち，ある一定以上の値の法線成分は打ち消すことができないものとする．

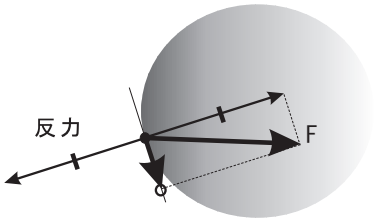


図 6　反力
Fig. 6.　Repulsion

**〈4･2〉 引　　力**　引力は，図 7 に示すように，閉曲線が内側に凸となった場合，その動作点に加わる力の向きが反転し，隣り合う動作点へと引き寄せられていた力が押し出される力となる動作をするものである．これによって，圧力と引力による閉曲線の収縮が可能となり，凹凸な目的領域の抽出が可能となる．さらには，引力のみの収縮の場合においても，閉曲線のバランスを保つことが可能となる．また圧力の値 $K_p$ は，注目した動作点が隣り合う 2 つの動作点より内側に凸となる位置へ進入するためのきっかけを作るものとして，ごく僅かで良い．このように，この引力を用いることで，比較的単純な仕組みで動的輪郭モデルを表現することができる．

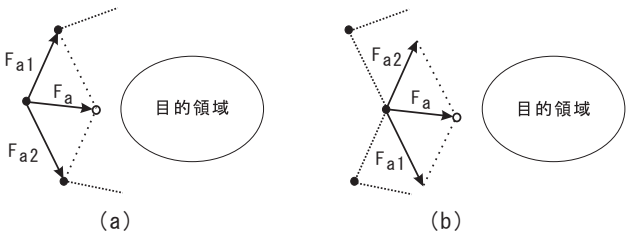


図 7　(a) は一般の引力，(b) は本論文の引力
Fig. 7.　(a) is general attraction，(b) is attraction in this paper

**〈4•3〉 分裂する動的輪郭モデル**　画像中に目的領域が唯一つである場合，従来の動的輪郭モデルによって抽出することができる．しかしながら，目的領域が複数ある場合，従来の動的輪郭モデルは，収縮するのみで分裂する性質を持たないため，それぞれの領域を抽出することができない．ここでは，分裂する性質を持った動的輪郭モデルが，複数領域を抽出する様子を示す．

図 8(a) に示すように，閉曲線が収縮を続ける過程で，動作点に働く引力がつり合った時，圧力を考慮することで，着目した動作点が隣り合う 2 つの動作点より内側に凸となる位置へ進入することが可能となる．

これらの力の働きにより閉曲線は，図 8(b) に示す状態となる．この時，着目した動作点間の距離 $h$ がパラメータとして与える一定の値よりも小さくなった場合，図 8(c) に示すように動的輪郭モデルの閉曲線を分裂させる．このとき，分裂の対象となる 2 つの着目した動作点上に新たに動作点を生成させる．その後，目的領域がまだ複数ある場合，同様に分裂を繰り返し，分裂した閉曲線はそれぞれの目的領域に向かって収縮を続ける．このように，圧力と引力によって閉曲線の収縮が可能となり，分裂する性質を持たせることで複数の目的領域の抽出が可能となる．

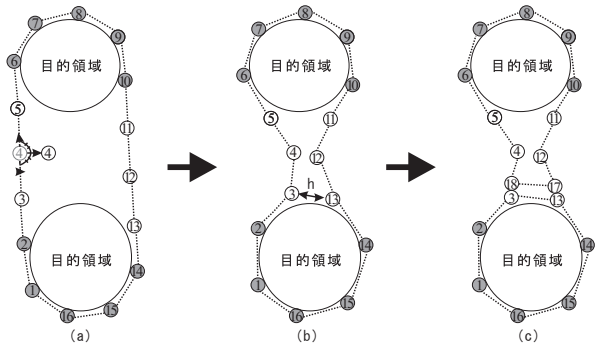


図 8　分裂する動的輪郭モデル
Fig. 8.　Split active contour models

## 5.　話者方向推定処理

話者方向推定処理の手法としては，様々なものがあるが，本システムでは推定方向はカメラの数に相当する 3 方向としている．このため，あまり高い角度分解能は要求されず，また FPGA 上の処理のため，計算量が比較的少ない手法が望ましいという点から以下の単純な手法を用いる．

まず，サンプリング周波数 $Fs$ で受音される理想的な正弦波は式 (1) のように表される．

$$\sin\left(2\pi f\frac{t}{Fs}\right) \cdots\cdots (1)$$

式 (1) において時間 $\tau$ だけ遅れることにより生じる位相差 $\Delta\phi$ は

$$\Delta\phi = \frac{2\pi f\tau}{Fs} \cdots\cdots (2)$$

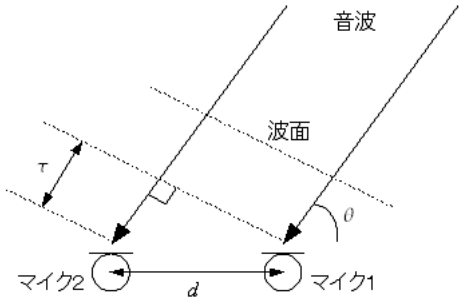


図 9　方向推定モデル
Fig. 9.　Direction of estimation model

で与えられる．

空間的に配置された 2 本のマイクロホンで音波を受音すると，各音波の間には時間差が生じる．図 9 を用いてこのことを説明する．ただし，ここでは自由空間において音波が音速 $c$ で到来している状況を仮定する．この音波を間隔 $d$ で並べた 2 つのマイクロホンで受音する．

図 9 の $\theta$ 方向から到来した音波は，まずマイク 1 において受音される．次に，音波は図 9 に示した遅延量 $\tau$ だけ遅れてマイク 2 に到達する．この関係を式で表すと式 (3) のようになる．

$$\tau = \frac{d\cdot\cos(\theta\times\frac{\pi}{180})\cdot Fs}{c} \cdots\cdots (3)$$

式 (2) および式 (3) を，音源方向 $\theta$ について解くと，

$$\theta = \cos^{-1}\left(\frac{\Delta\phi c}{2\pi fd}\right)\times\frac{180}{\pi} \cdots\cdots (4)$$

の関係を得る．この式 (4) の計算結果から，話者方向の推定を行っている．

## 6.　出力映像生成処理

パノラマ映像からの話者領域抽出と出力映像生成処理では以下のことを行う．

（1）　NTSC ビデオカメラで撮影した 320×240[pixel] の 3 つの入力映像の 1 つに対して分裂する動的輪郭モデルを適用する．

（2）　話者領域抽出結果を基に話者を拡大表示し，話者映像を生成する．

（3）　3 つの入力映像からパノラマ映像を生成する．

（4）　話者映像とパノラマ映像から出力映像を合成する．

合成する出力映像の仕様を，図 10 に示す．

出力映像は，話者を拡大した映像である「話者映像」，および会議会場全体の映像である 640×180[pixel] の「パノラマ映像」を，640×480[pixel] サイズに合成したものである．

拡大表示の処理の流れは以下の通りである．まず，320×240[pixel] の入力映像に対して分裂する動的輪郭モデルを適用し，話者領域を抽出している 20 点ある動作点の $x$ 座標と $y$ 座標の平均値を計算する．そして，計算された平均値の座標点を中心として，話者映像を拡大するのだが，出力映像の映像サイズ (640×480[pixel]) を考慮して，3/8 サ

イズの 240×180[pixel] に縮小した入力映像の 213×80[pixel] の画像を取り出し，その 3 倍である 639×240[pixel] を話者映像としてモニタに出力している．

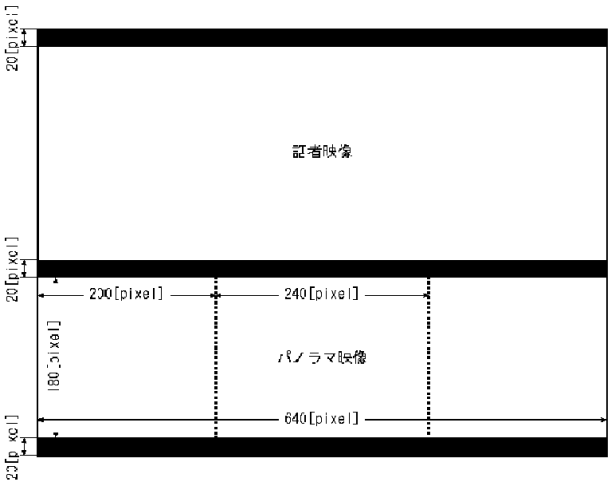


図 10　出力映像の仕様
Fig. 10.　Output image

## 7. 実　　験

図 11 に分裂する動的輪郭モデルが，パノラマ映像から複数の話者領域を抽出する様子を示す．分裂する動的輪郭モデルが，2 人の顔領域を抽出している様子が確認できる．また，パノラマ映像に対して分裂する動的輪郭モデルを適用し，複数の顔画像領域を抽出する処理を組み込んだ FPGA の回路規模を表 2 に示す．なお，開発した回路は，使用可能なロジックエレメント数の約 67 %で実現できた．映像処理部の 1 フレームの処理にかかるクロック数は，約 5,919,890 で約 123[ms] である．また，3 つの入力映像のうち，どの映像に動的輪郭モデルを適用するのかは，5 で述べた手法を用いて音声により切り換えている．

表 2　回路規模
Table 2.　Circuit size

| 処理回路名 | 消費ロジックエレメント数 |
|---|---|
| デコーダ制御処理部 | 274 |
| 映像処理部 | 12467 |
| エンコーダ制御処理部 | 370 |
| メモリ等 | 252 |
| 合計 | 13363 |

## 8. おわりに

本稿では，テレビ会議用ビデオカメラの開発を目的に，会議風景のパノラマ映像と，話者の拡大映像を表示するシステムのハードウエア実現を試みた．本システムでは，パノラマ映像に対して分裂する動的輪郭モデルを適用し，複数の顔画像を抽出した．また，話者方向を推定する処理のハードウエア実現を試みた．

図 11　実験結果
Fig. 11.　Experimental result

## 参考文献


（1） 李 咏梅，桂 竜二，菅原 一孔，小西 亮介：周波数・位相推定手法に基づく複数音声の到来方向推定について，電子情報通信学会論文誌 A, Vol.J86-A, No.3, pp.320-325 (2003)

（2） 米本 良，上杉 徹，川村 尚生，菅原 一孔：パノラマ映像生成機能を持つテレビ会議用ビデオカメラの開発に関する研究，電気学会電子回路研究会資料, Vol. ECT-06-119, pp.31-35 (2006)

（3） 三秋 俊雄，川村 尚生，菅原 一孔：複数領域抽出のための新しい引力項を持つ動的輪郭モデルのハードウェア実現について，情報科学技術レターズ, pp.261-264(2006)

（4） 長谷川 裕恭：VHDL によるハードウェア設計入門，CQ 出版社 (2002)

（5） 森岡 澄夫：HDL による高性能ディジタル回路設計，CQ 出版社 (2002)