話者方向推定機能を持つパノラマカメラのハードウエア実現について

Hardware Realization of Panoramic Camera with Direction of Speaker Estimation Function

上杉 徹 † 川村 尚生 †† 清水 忠昭 †† 菅原 一孔 ††
Toru Uesugi † Takao Kawamura †† Tadaaki Shimizu †† Kazunori Sugahara ††
† 鳥取大学大学院　工学研究科　　†† 鳥取大学　工学部

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のが通常である．このため，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などの問題がある．こ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 ビデオカメラを用いて撮影した映像から生成したパノラマ映像と，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を特定する手法が必要となる．話者を特定するには，各々のビデオカメラの映像から話者の顔領域を抽出することが必要となるが，1 つのカメラからの映像内に複数の顔画像が映っていることが考えられる．このような場合には，話者だけを特定することは容易でないため，本稿では，話者を含んで複数の顔領域を抽出することを考える．特定領域を抽出する方法として，通常の動的輪郭モデル [1] を用いた場合には単一の領域しか抽出できない．本研究では，抽出対象が複数であるため，先に我々が提案した分裂する動的輪郭モデル [2] を適用し，画像から複数の特定領域を抽出することを考える．

本研究では，音声による話者方向推定から 3 つのカメラのうち話者を含んでいるカメラを選択し，選ばれた入力映像に対して分裂する動的輪郭モデルを適用し，複数の顔領域を抽出する処理を FPGA を用いて，その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，FPGA ボード，デジタルビデオエンコーダ，モニタから構成されている．

処理の流れは以下の通りである．まず，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，FPGA ボード上のメモリに蓄積する．そして FPGA 上の各回路において，入力された映像データを基に，パノラマ映像生成と分裂する動的輪郭モデルの処理を行い，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．システムを構成している各装置を図 1 に示す．

パノラマ映像生成と分裂する動的輪郭モデルによる複数領域抽出は，FPGA 上に実装した．この FPGA を搭載した FPGA ボードのユーザー I/O ポートと，デコーダ，エンコーダ，メモリを実装したビデオ信号入出力ボードのコネクタを接続することで，FPGA の内部回路がアクセス可能となる装置を開発し，実験を行った．開発したシステム全体の構成を図 2 に示す．

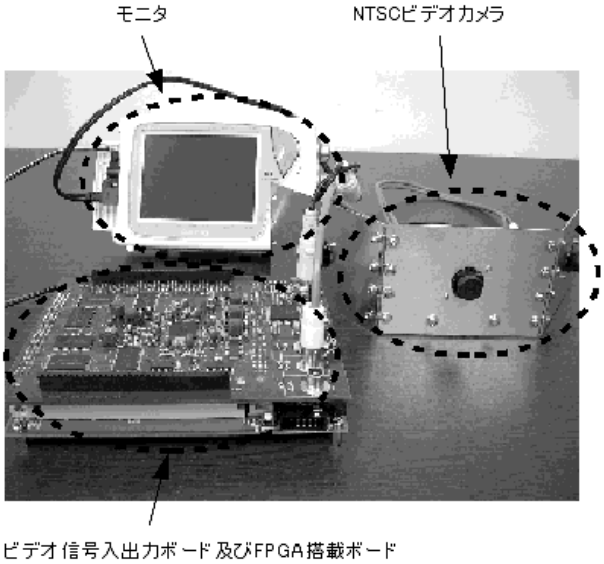


図 1: 使用した装置

3　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図 3 のような単純な方向推定モデル [3] を用いる．

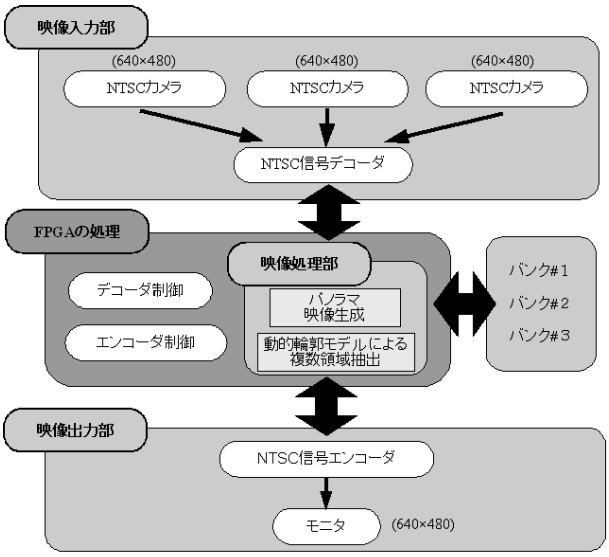


図 2: 全体の構成図

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．

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

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

$$\Delta\phi = \frac{2\pi f\tau}{Fs} \tag{2}$$

で与えられる．

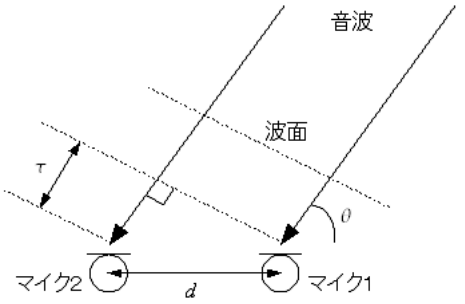


図 3: 方向推定モデル

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．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 3 の θ 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 3 に示した遅延量 τ 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．

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

式 (2) および式 (3) を，音源方向 θ について解くと，

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

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．

4 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．その様子を図 4 に示す．

映像サイズの縮小方法として，まず，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．

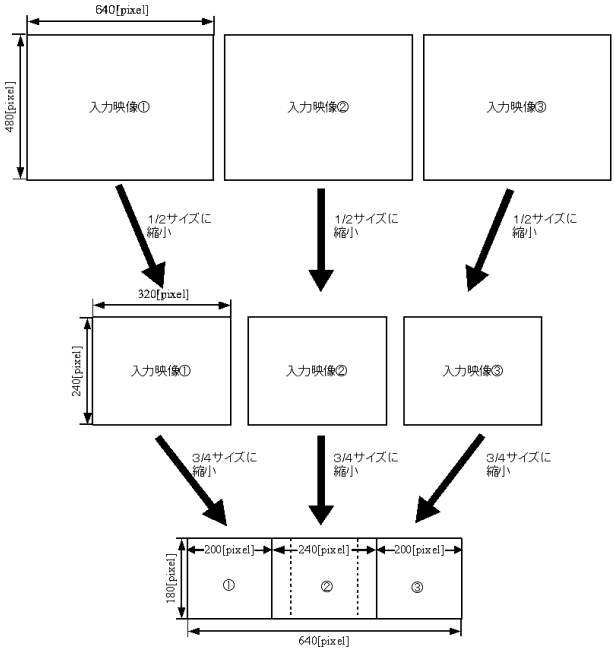


図 4: パノラマ映像を生成する様子

5　複数領域抽出手法

5.1　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．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．

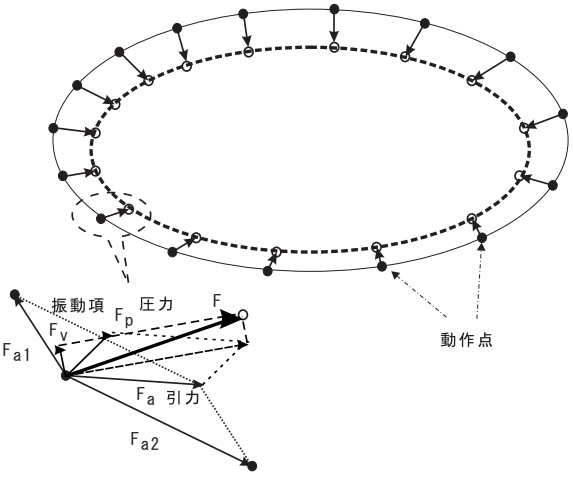
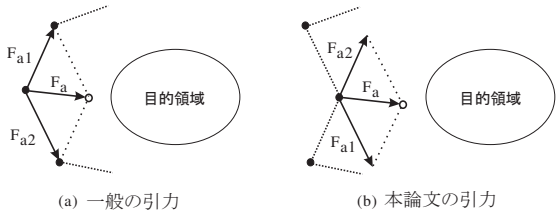


図 5: 動的輪郭モデルによる閉曲線の収縮

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，**5.2** 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，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．

5.2　引力

引力は，図 6 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．



(a) 一般の引力　　(b) 本論文の引力

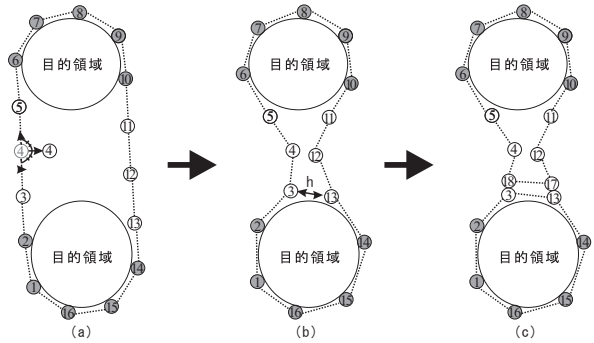
図 6: 引力

5.3　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 7(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 7(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 7(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．



(a)　(b)　(c)

図 7: 分裂する動的輪郭モデル

6　出力映像生成処理

出力映像生成処理は以下の流れで行う．

1. 音声による話者方向推定から 3 つのカメラのうち話者を含んでいるカメラを選択する．
2. 3 つの入力映像のうち，選択された 320×240[pixel] の映像に対して分裂する動的輪郭モデルを適用し，そのカメラに映っている話者の顔領域を抽出する．但し，映っている人物が話者一人だけでない場合，話者を含んで複数の顔領域を抽出する．
3. 3 つの入力映像からパノラマ映像を生成する．
4. 顔領域を抽出した結果，20 点ある動作点の x 座標と y 座標の平均値を計算する．
5. 計算された平均値の座標点を中心として，話者を含んでいる領域を拡大する．その際，出力映像の映像サイズが 640×480[pixel] であることを考慮して，640×180[pixel] のサイズであるパノラマ映像から 213×80[pixel] の話者を含んだ領域を取り出す．
6. その領域を 3 倍の 639×240[pixel] に拡大表示し，それを話者映像とする．
7. パノラマ映像と話者映像から出力映像を合成し，モニタに出力する．

合成する出力映像の仕様を，図 8 に示す．

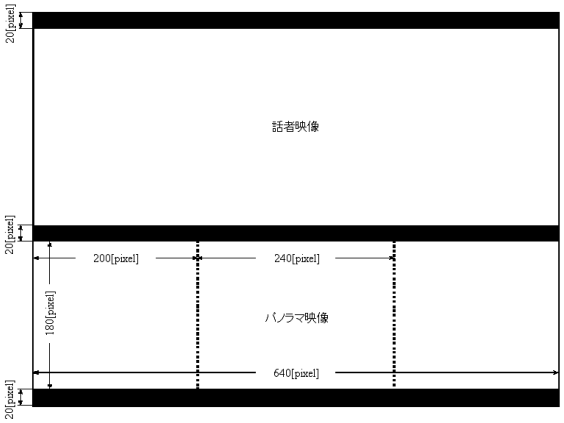


図 8: 出力映像の仕様

7　映像合成処理実験

映像合成処理の実験結果を図 9 に示す．図 9 に見られる小さい四角い点は収縮している動作点を表している．この実験において，分裂する動的輪郭モデルが，複数の顔領域を抽出していることが確認できる．また，分裂する動的輪郭モデルを適用し，複数の顔領域を抽出する処理を組み込んだ FPGA の回路規模を表 1 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，**3** で述べた手法を用いて音声により切り換えている．

表 1: 回路規模

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

図 9: 映像合成処理実験結果

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，分裂する動的輪郭モデルを適用し，複数の顔領域を抽出した．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．

参考文献

[1] 菅原 一孔, 新地 俊幹, 小西 亮介 : 振動項を持つ動的輪郭モデル，電子情報通信学会論文誌 D-Ⅱ, Vol.J80-D-Ⅱ, No.12, pp.3232-3235 (1997)

[2] 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)

[3] 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)