2Z-4

多人数参加型会議システムの映像表示法

冨野　剛† 　山田　貴弘† 　井上　亮文† 　市村　哲† 　松下　温†
†東京工科大学

1. はじめに

近年、インターネットの急速な普及、特にブロードバンドによる高速アクセス回線常時接続環境が整備され、テキスト情報中心だったやり取りも、音声や映像を含むマルチメディア情報のやり取りへと、利用形態が変化しつつある。これらの、利用形態の中でも、企業中心として広く普及しつつあるのが、インターネットテレビテレビ会議システムである。この傾向は、日本だけにとどまらず、世界規模で広まっている。2002 年度の中国テレビ会議システムの市場規模は 9,492 万ドルだった。今後も毎年平均 29%増の成長率で拡大し、2003 年市場規模は 12 億元（約 1.4 億ドル）、 2007 年まで 3.36 億ドルになると予測している[1]。

また、出張費を削減できるテレビ会議システムを利用する企業は少なくない。さらに、2001 年 9 月 11 日の同時多発テロ以降、日本でも海外出張の自粛や削減をする企業が増えている。このような危機管理の一環としても企業のテレビ会議の導入が検討されている。

本研究では、画像処理とマイクロホンアレーを用いることで、ランニングコストを抑えつつ、多人数参加ができるよう映像表示を工夫したテレビ会議システムを提案する。

2. 背景

現在のインターネットテレビ会議システムは、一地点あたりの利用人数が 1 人または多くても 3 人でしか使えないものが多く、また、画質音質ともに低いという問題がある。実際に、画面の表示サイズが限られており、相手の表情を捉えにくく誰が発言者かわかりにくい、誰が参加しているかわかりにくい、的確な意思が伝わりにくいといった問題がある。

そこで、本研究では、画像処理を用いることで、ランニングコストを抑え、通常の PC と同等の設備で、多人数が参加できるよう映像表示を工夫したものを提案する。

Method of displaying speaker for TV conference
†Takeshi Tomino†; Takahiro Yamada†; Akihumi Inoue†; Satoshi Ichimura†; Yutaka Matsushita†
†Tokyo University of Technology

3. システム概要

本システムでは図 1 のような 10 人程度の会議環境を想定している。これを 1 地点とし、発言者拡大部と会議室全体部を 1 つの出力として、他の地点に送信し、同時に、他の地点の画像を受け取ることにより会議システムとする。

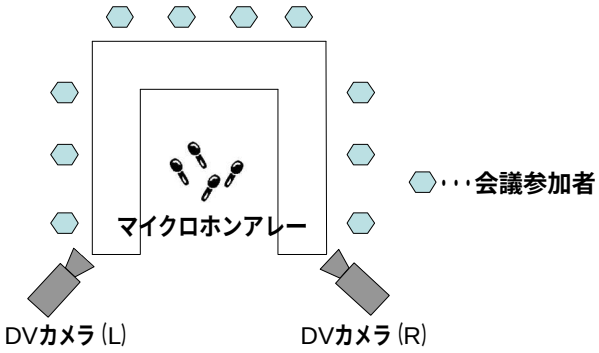

図.1　会議室想定環境

システムの流れは、2 台の DV カメラの入力映像と、マイクロホンアレーから得られた話者位置推定データを統合して、会議参加者および話者の検出、拡大をし、上部に、発言者を 3 人まで拡大したもの、下部に会議室全体を写したものを 1 地点の会議用映像として出力する。

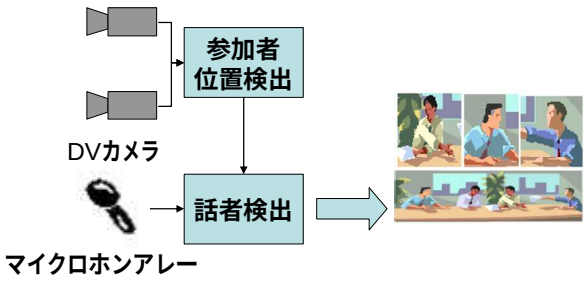

図.2　システムの流れ

3.1　会議参加者検出

会議参加者の検出にあたり、いくつかの方式、アルゴリズムを試し、会議時の人の動きの特徴

を調べた。
その結果、試した手法の一つである二値化によるテンプレートマッチング法が適していないことがわかった。テンプレートマッチング法とは顔の画像のテンプレートをデータベースとして用意しておき、それに近似したものを探し出す方法である。会議において、顔の向きが人や時間によって様々であることや、明るさによっても二値化の閾値が変わる等、同じ顔でも、多くのテンプレートを参照せねばならず、リアルタイム会議システムには不向きであることがわかった。
そこで、差分認識に着目した。
調査の結果、会議中やその他の場合においても通常時の他、特に話しているときに、人がまったく動かないこということがなく、それを元に会議参加者を検出できることが。
また、差分認識だけでは人の検出は不完全なため、肌色認識も組み合わせた。
肌色認識は、RGB 値のばらつきにより、RGB 表色系による肌色認識が困難であることが判明した。しかし、デジタルカメラで撮った画像の肌の色の RGB 値と HSV 値を比べてみたところ。表.1 のようになり、色相(H)と明度(V)の値が独立していため、部屋の明るさに関係なく肌色が認識でき、色相(H)の値が 0～35 となるところで肌色となることがわかった[2]。

表.1　デジタルカメラで撮った肌色の RGB 値と HSV 値

肌サンプル	R 赤	G 緑	B 青	H 色相	S 色彩	V 明度
A	229	196	181	19	53	229
B	196	205	191	20	46	233
C	203	163	128	28	94	203
D	141	96	77	17	115	141
E	234	173	110	31	134	231
F	284	125	121	4	87	184
G	159	101	94	10	112	159
H	95	60	30	28	174	95

まとめると、はじめに、背景差分をとり、動物体以外を削除し、次に、HSV 肌色認識を行い、肌色以外をなくし、その上で、フレーム間差分を数回とることにより、一瞬だけ動いたものや、カーテンの開閉による明るさの変化に対応し、より動きが大きいものを会議参加者、その中でも特に発言者の候補と特定するようにした(図.3)。

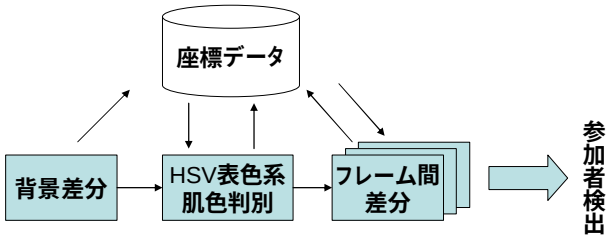

図.3　会議参加者検出

3.2　話者検出
話者検出に関しては、図.2 の通り、会議参加者認識に加えて、マイクロホンアレーによる話者位置推定もシステムに統合した。
方法としては、マイクロホンアレーの話者位置推定により、話者の推定角度と音量のデータを受け取り、それを元に会議参加者検出で得られた、発言者の候補を参照し、話者の位置を検出する。そこで検出された会議参加者を有力な発言者とし、3 人までを拡大して、発言者拡大部として出力した。

4.　おわりに
本システムにより、大掛かりなハードウェアを用意しなくても、通常の PC 環境に機材を少し足すだけで、多人数参加型会議を行うことのできるインターネットテレビ会議システムを構築した。
今後は、多地点でも利用できるシステムを目指し、より、実際の会議に適したものにするとともに、処理軽量化や、アルゴリズムの再検討などにより話者検出率、話者検出効率の向上を図りたい。

参考文献
[1]月間シティ
http://www.cityshanghai.com/

[2]Science and Engineering
at The University of Edinburgh
http://www.inf.ed.ac.uk/

[3]松尾、北川、長田、棚橋、“視聴覚情報の統合による話者位置検出システム”情報処理学会第 59 回全国大会、特 1-57-63（1999）