視聴覚情報の統合による会議映像の自動撮影

大西　正輝†（学生員）　　影林　岳彦†
福永　邦雄†（正員）

Production of Videoconferencing Images by Computer Controlled Camera Based on Integration of Audiovisual Information

Masaki ONISHI†, *Student Member*,
Takehiko KAGEBAYASHI†, *Nonmember*, and
Kunio FUKUNAGA†, *Member*

† 大阪府立大学大学院工学研究科，堺市
Graduate School of Engineering, Osaka Prefecture University
1-1, Gakuen-cho, Sakai-shi, 599-8531 Japan

あらまし　ネットワーク環境の整備に伴い，遠隔会議システムの利用が増えてきている．遠隔会議を行うためには会議を撮影する必要があり，固定カメラを通して得られる映像を用いることが多いが，固定カメラから得られる映像は変化に乏しく単調である．臨場感のある遠隔会議を行うためには，相手サイトにいる参加者の注目する領域が効率良く撮影されている映像を用いることが望ましい．本論文では，マイクとカメラを用いて抽出した視聴覚情報を統合することによって決定した注目者をもとにしてカメラ制御を行うことで自動的に会議映像を生成する手法を提案する．

キーワード　会議映像，視聴覚情報，情報の統合，自動撮影

1. まえがき

近年，ネットワーク環境の整備が進むにつれて，大容量のマルチメディア情報の通信が可能となり，遠隔会議システムを活用できる環境が整ってきている．遠隔会議において，各サイトの参加者は相手サイトの映像と音声をもとに会議を進める．このとき，一般には固定カメラを用いて会議を撮影することが多い．しかし，固定カメラから得られる映像は変化に乏しく，また相手サイトの参加者が注目する領域を効率良く映像化しているとは限らない．これを回避する方法の一つとして，カメラマンが撮影した映像を用いることも考えられるが，人手がかかるなどの問題があり，自動的にカメラを制御して撮影する方法が検討されている．

自動撮影に関する代表的な方法では，カメラワークを決定するための特徴量を抽出するために，マイクなどの聴覚センサを用いるかカメラなどの視覚センサを用いることが多い．聴覚情報をもとに発言者を調べ，テレビ討論番組のカメラワークの知識に基づいて画面を切り換える多人数テレビ会議システム [1] では，固定カメラによる会議システムとの比較を行い高い評価を得ている．これに対して，視覚情報を映像上の時空間情報発生量の分布から求めて，カメラワークを設計する手法 [2] では，遠隔講義の撮影自動化を実現している．

本論文では，効果的な会議映像の生成を行う方法の一つとして，聴覚情報と視覚情報の両者を用いて注目領域を調べ，カメラワークを設定する新たな手法を提案し，固定カメラによる会議システムやカメラマンが撮影する会議システムとの比較を行い，その有効性を明らかにする [3]．

2. 映像生成におけるカメラ制御

いくつかのセンサから得られるデータを統合する方法をセンサフュージョンといい，ロバストな認識システムへの応用が期待されている [4], [5]．本研究では五感の中でも外界の認識に重要な役割を果たす聴覚情報と視覚情報を統合することで必要な情報を抽出し，注目領域を決める．まず，聴覚情報からは空間的に離れた複数のマイクの入力をもとに音源定位を行い，人物の発声位置を推定する．一方，視覚情報については時空間情報発生量をもとに，活発に動作などを行っている映像上の人物領域を抽出する．そしてこれら両者の情報を統合することで注目領域を選択し，カメラ制御アルゴリズムを用いて可動カメラを制御する．

3. 音源定位による聴覚情報の抽出

本研究では聴覚情報を参加者の発声とみなし，参加者の発声位置を求めるために音声の音源定位を行う．まず，音源定位について説明する．

3.1　3次元音源定位

音源定位とは，音源が発する音波を観測・解析することによって，音源の位置を推定することである．音源定位の代表的な手法として，ビームフォーマによる手法 [6] や時空間勾配法による手法 [7] などが挙げられる．一般的に，音源定位においては方向定位に比べて距離定位の精度が低い．これは，距離情報に対する分解能が低いことに起因しており，マイクロホンアレー中のマイク間の距離に比べて音源までの距離が遠い場合には，距離定位の困難さが顕著に現れる．

ここでは，定位の対象となる音源は参加者の音声となる．このため音源数は常に一つとは限らず，複数音源の定位が可能であることが望ましい．しかし，会議という場面の性質から，複数の参加者が同時に長時間発声することは考えにくいため，必ずしも複数の音源を同時に定位する必要はない．そこで，4 本のマイク

を空間内に配置し，リアルタイム処理に適している音の到達時間差を用いる手法によって，単一音源を 3 次元空間で定位する．

3.2 マイクロホンアレー

音源定位が可能な範囲とその分解能はマイク間距離に依存する．マイク間距離が大きいほど定位可能な範囲は広がるが，到達時間差を一意に決定できる周波数が低くなるという問題が生じる．ここでは，装置のコンパクトさや配置の簡単さを優先して，図 1 のようにマイクを配置するマイクロホンアレーを用意した．$M_1 \sim M_4$ がマイクの設置位置を表しており，マイク間距離 M_d は 45cm とした．

サンプリング処理は，4 本のマイクからの入力に対して同時に行い，マイク M_i $(i = 1 \sim 4)$ からの入力信号を $g_i(t)$ $(t = 1 \sim N_t)$ で表す．また，サンプリング周波数 F_s は 25 kHz，サンプリング周期 T_s は 40 μ 秒とする．

3.3 音声信号の検出

まず参加者が発声しているか否かを判断する．マイク入力に音声信号が含まれる場合には，音声によって低周波成分が大きくなる．一方，音声信号が含まれない場合，低周波成分が小さくなり，白色雑音による零交差数が増加する．したがって，零交差数が少なくなれば，音声信号が含まれている可能性が高いと考えられる．また，参加者の他に周囲で大きな音を発するものが存在しない場合には，参加者の音声によってのみ入力信号の平均エネルギーは増加すると考えられる．そこで，入力信号の零交差数が小さくなり，かつ平均エネルギーが大きくなった場合に，参加者が発声しているものとする．

3.4 到達時間差による音源定位

会議の参加者が発声している場合にのみ，音声の音源定位を行う．音源定位の方法として，本研究では到達時間差に基づく手法を用いる．

各マイクで音声を受信したときには，それぞれの入力信号間には強い相関がある．そこで，各入力信号の相互相関関数を求めることで，到達時間差を求める．二つの入力信号 $g_i(t)$ と $g_j(t)$ 間の相互相関関数 $\phi(t_m)$ は次式で求める．

$$\phi(t_m) = \sum_{t=1}^{N_\phi} g_i(t) \cdot g_j(t + t_m) \qquad (1)$$

ここでは，$\phi(t_m)$ が大きいところに強い相関があることになり，そのときに得られた t_m が到達時間差となる．式 (1) 中の N_ϕ は，マイク間の距離を音波が進むのにかかる時間 M_d/V_s をサンプリング周期 T_s で割った次式の値を用いる．

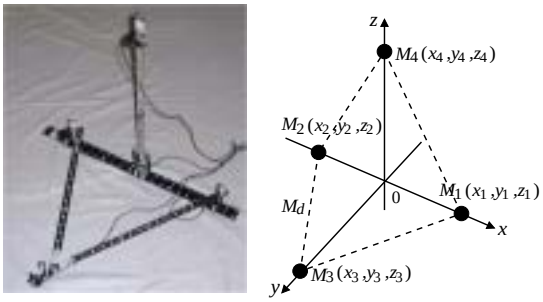

図 1 マイクロホンアレー
Fig. 1 Microphone array.

$$N_\phi = \frac{M_d}{V_s \times T_s} \qquad (2)$$

ただし，V_s は音速を表す．

ここでは 4 本のマイクを用いてマイクロホンアレーを構成しているため，三つの独立した到達時間差が得られる．三つの独立した到達時間差から音源の 3 次元位置を数学的に求めることができるが，音源位置によって三つの到達時間差は固有の値をもつことを利用して，リアルタイム処理の観点から，あらかじめ対象空間のボクセルごとに計算で求めておいた理論的な到達時間差と，得られた到達時間差を比較して音源位置を導き出す方法を用いた．

以上の処理を 20 m 秒ごとに繰り返し行い，過去数秒の定位結果を累積して音源位置の分布を求めることで，その間に発声した参加者が複数人である場合にも，複数の音源位置を求めることが可能になる．

4. 情報発生量による視覚情報の抽出

本研究では，参加者の動きに伴う映像中の画素の色変化を視覚情報として用いる．文献 [2] では，映像中における空間的に存在する情報と時間的に存在する情報に着目して情報発生量を定義しており，本研究でも情報発生量の分布を求めることで視覚情報の抽出を行う．情報発生量の詳細については文献 [2] に解説を譲る．

5. 注目者の遷移に基づくカメラ制御

5.1 視聴覚情報の統合による注目者の選択

会議における注目領域を調べるため，聴覚情報と視覚情報の二つの情報の抽出を行った．しかし，聴覚情報と視覚情報は本質的に異なる情報であるため，統一的に扱うための情報量（注目量）が必要になる．

聴覚情報については，3 次元の音源位置が 2 次元のモニタカメラ画像上に投影される画素位置を求め，その x 座標における注目量へ変換する．その際に過去の一定期間に発生した聴覚情報は累積する．これにより，この間に複数の参加者が発言した場合には，それらを複数の音源として抽出することができる．この期間の長さは t_a 秒とする（本論文では 4 秒）．一方，視覚情報については，モニタカメラ画像から計算した情報発生量に対する画像平面上の水平方向への射影ヒストグラムを求めることで，x 座標における注目量へ変換する．また，視覚情報は発生した直後に情報を失うわけではないと考えられ，この際にも過去の一定期間の情報発生量を累積する．この期間の長さは t_v 秒とする（本論文では 3 秒）．

次に，聴覚情報による注目量と視覚情報による注目量を統合する．ここでは実験での経験に基づいて聴覚情報を中心に考え，聴覚情報による注目量が一定のしきい値を超えた領域内においてのみ，聴覚情報による注目量に視覚情報による注目量を加算した値を最終的な注目量とする．聴覚情報による注目量がしきい値以下の領域では注目量は 0 としている．視聴覚情報を統合して求めた注目量をカメラ制御アルゴリズムへ適用するために，注目すべき参加者（注目者）を選択する．注目者を選択するために，各参加者ごとに参加者領域内の注目量を合計して，参加者全員の平均を求める．そして，平均値と合計値との差を計算して，その値がしきい値以上となる参加者を注目者とする．

5.2 カメラ制御アルゴリズム

テレビ討論番組のカメラワークの知識を利用して会議を撮影する手法 [1] では，参加者を話者とそれ以外の人（第三者）の 2 種類に役割付けし，オペレータがリアルタイムで参加者の役割を与えることで，カメラを制御している．

本研究では，参加者の役割付けを，注目者とそれ以外の人（第三者）に置き換え，この提案手法に基づいたカメラ制御を行う．参加者の役割に基づいて撮影領域を決定するために，ショットの種類を表 1 のように分類する．この分類では，映像が単調にならないように，画面に入る人数を変化させるように配慮している．また，これらのショットを切り換えるタイミングは，次の 2 種類を考える．

- 注目者が交替する
- 同一ショットが一定の持続時間を超過する

前者は，注目者の交替によって，交替前の注目者を撮影する重要性が低くなることから切換を行うものである．後者は，同一ショットが長時間続くと，単調な映像となってしまうことから，参加者の関心を持続させるためのに切換を行うものである．ここでは，持続時間を 10 秒とした．また，ショット切換が頻発すると見づらい映像になるため，直前のショット切換後の一定期間は，これらの条件が発生しても切換を行わず，一定期間が経過するまで待機する．このショット切換を行わない期間を 2 秒とした．

次に，ショットを切り換える際にどのショットへ遷移させるかを，遷移確率行列で定義する．遷移確率行列は，テレビ討論番組において，どのショットからどのショットへ遷移したかを，すべての遷移について統計をとり，確率で表したものである [1]．遷移確率行列は，ショット切換の発生条件ごとに定義されており，表 2 及び表 3 に，2 種類の遷移確率行列を示す．

表 1　ショットの分類
Table 1　Classification of shot.

ショット	説明
s_1	注目者 1 人だけを映す
s_2	注目者とその周囲の参加者を映す
s_3	注目者が複数いる場合に注目者複数人を映す
s_4	注目者でない人物 1 人だけを映す
s_5	注目者でない人物を複数人映す
s_6	参加者全員を映す

表 2　注目者交替時の遷移確率行列
Table 2　Transition probability matrix on attended human switch.

		遷移先					
		s_1	s_2	s_3	s_4	s_5	s_6
遷移元	s_1	45.1%	18.5%	3.4%	4.1%	13.9%	15.0%
	s_2	40.1%	30.4%	11.7%	1.2%	2.4%	13.4%
	s_3	30.8%	25.9%	34.7%	1.3%	2.3%	5.1%
	s_4	44.3%	31.5%	11.9%	2.8%	8.5%	1.0%
	s_5	36.7%	31.5%	8.3%	8.9%	4.1%	11.5%
	s_6	32.2%	23.3%	30.1%	0.0%	4.4%	10.0%

表 3　持続時間超過時の遷移確率行列
Table 3　Transition probability matrix on overpass duration.

		遷移先					
		s_1	s_2	s_3	s_4	s_5	s_6
遷移元	s_1	14.8%	21.7%	25.6%	21.2%	6.7%	11.0%
	s_2	30.4%	17.2%	22.5%	10.7%	7.2%	12.0%
	s_3	34.4%	18.0%	25.4%	10.7%	2.1%	9.4%
	s_4	33.2%	22.9%	26.0%	4.0%	2.9%	12.0%
	s_5	42.2%	19.0%	24.3%	7.2%	2.3%	5.0%
	s_6	43.1%	24.1%	31.4%	1.4%	0.0%	0.0%

6. 実験及び考察

6.1 会議映像生成実験

音源定位の精度を評価するために，男性の発声を音源として発声位置を変えて音源定位を行ったところ，

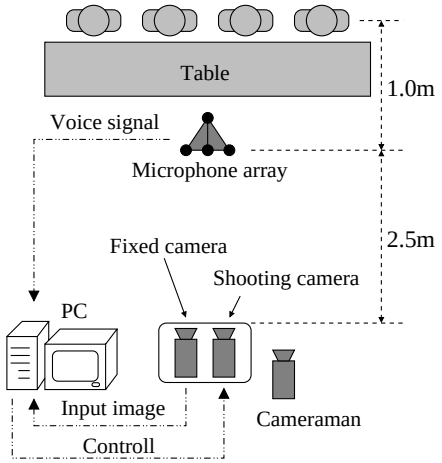

図 2 マイクロホンアレーとカメラの配置
Fig. 2 Microphone array and camera layout.

図 3 撮影映像例
Fig. 3 Example of created video images.

マイクロホンアレーからの距離が大きくなるほど定位の誤差は大きくなるが，1.5m の範囲内では比較的精度良く音源位置を定位できることがわかった [3]．そこで，マイクロホンアレーは会議の参加者から 1.5m 以上離れないように設置することにした．本実験で用いた自動会議撮影システムにおけるマイクロホンアレー及びカメラの配置を図 2 に示す．

本手法を用いて撮影用の可動カメラを制御して，リアルタイムで実際の会議を撮影した．マイクからの音声と固定カメラからの画像を並列して 1 台の計算機に取り込み，音源定位処理と画像処理による情報発生量の計算を行いながら，撮影カメラの制御を行った．認識用の固定カメラから得られる入力画像の解像度は，320 × 240 画素である．会議の参加者は 4 人として，参加者全員が映像中に収まるように認識用の固定カメラを配置した．また，音源定位処理に要する割合は，1 秒間のうちで 0.6 秒程度であり，残りの 0.4 秒で画像処理を行った．この 0.4 秒で処理できるフレーム数は 2.5 フレーム程度であった．なお，固定カメラ，可動カメラ，マイクロホンアレーの位置関係，及び参加者の移動範囲はあらかじめパラメータとして与えておいた．特に，参加者の移動範囲については，注目者の交代を正しく抽出するために，各参加者が重ならないように考慮した．

本手法による撮影映像例を図 3 に示す．図中の左段は全体ショットから注目者単独ショットへ切り替わったためにズーミングが行われた例を示し，右段が注目者の交替によりパンニングが行われた例を示している．

6.2 映像評価のためのアンケート

本手法の有効性を確認するために，**6.1** の実験において撮影した映像を，大学生の被験者 14 人にアンケート評価してもらった．評価の対象の映像は次の 3 種類である

- 固定カメラを用いて撮影した映像（固定映像）
- 本手法を用いて自動で撮影した映像（自動映像）
- 人手により撮影した映像（人手映像）

それぞれ，同時刻における 5 分程度の映像である．また，人手による撮影はカメラマン 1 人で行った．この 3 種類の映像を，

- 固定映像と自動映像
- 固定映像と人手映像

の 2 度に分けて，横に並べた 2 台のモニタで 2 種類の映像を同時に呈示した．被験者に映像の撮影方法は知らせていない．

表 4　アンケート結果
Table 4　Result of questionnaires.

	質 問 項 目	人手	本手法
Q.1	参加者の表情はよく見えた	3.57	4.36
Q.2	会場の雰囲気が伝わりやすかった	2.71	3.43
Q.3	話し手は誰かがよくわかった	3.79	3.14
Q.4	身振り・しぐさはよく伝わった	3.79	3.71
Q.5	画面の切換は効果があった	3.50	3.79
Q.6	第三者への切換は効果があった	–	3.57
Q.7	映像は退屈だった	2.79	1.57
Q.8	映像は単調に見えた	3.00	1.57
Q.9	映像は見やすかった	3.43	2.64
Q.10	ストレスを感じた	2.93	3.29

（固定カメラ映像の評価値 3.00 に対する評価）

アンケートの評価項目は，文献 [1] 及び文献 [2] を参考にして本手法に関係すると思われる 10 項目を選んだ．その集計結果を表 4 に示す．評価は，質問項目に当てはまるかどうかを，固定カメラの映像の評価を 3 として，1 から 5 までの 5 段階評価で行った．表中の人手・本手法の各項目の値は評価値の平均値を表し，Q.6 については人手による撮影には第三者への切換がなかったため，評価ができていない．

6.3 考　　察

実験・アンケートにより得られた本手法の特徴について考察する．

アンケートの結果，一部の項目で人手により撮影した映像を上回る評価を得ることができた．特に Q.7 及び Q.8 において高い評価を得ることができ，映像が退屈・単調なものになることを防ぐためのカメラ制御の有効性が確認できた．また，Q.1 及び Q.2 については，注目者を適切に特定できたことや，時折第三者のショットを撮影することで会場全体の雰囲気が伝わりやすかったことが評価されたものと考えられる．しかし反対に，Q.9 及び Q.10 においては，人手により撮影した映像の評価には及ばなかった．また，人手によって撮影された映像の評価は固定カメラの映像の評価を上回ったが，本手法を用いて撮影した映像の評価は固定カメラの映像の評価を下回る結果となった．これは，ショット切換の発生条件である「注目者が交替する」が頻発し，ショットの切換が頻繁に起こったため，映像が見にくくなったことが原因であった．Q.3 において，人手により撮影した映像の評価が高かったのは，撮影者がほぼ話し手を中心に撮影したことが原因であると考えられる．本手法では，第三者のショットや全体ショットへの切換を行ったため，Q.2 の評価は高かったが，Q.3 では固定カメラの映像の評価とほぼ同程度となったと考えられる．

これらの結果から，テレビ討論番組のカメラワークの知識を利用したカメラ制御を行うことで，映像が単調・退屈になることを防ぐことが可能となり，参加者の注意が持続するような映像を生成することができた．つまり，テレビ討論番組の分析結果を利用した遷移確率行列を用いて，遷移先のショットを確率に合わせてランダムに選択することで映像が単調になることを防ぐことができた．また，同一ショットが一定の持続時間を超過した場合にショット切換を行うことで，映像が単調なものになることを防ぐことができた．

これまでの関連研究 [1], [2] では，聴覚情報または視覚情報のどちらか一方を用いて映像の生成を行っている．本研究では，これら両方を用いることで，より効果的な会議映像生成が可能になった．具体的には，比較的誤差の大きい音源定位結果を視覚情報との統合により排除することができた．また，複数人で対話しているときには身振りが大きい参加者を注目者として選択することで，話者だけではなく身振りも効率良く映像化することができた．

一方，参加者 1 人ひとりの前にマイクを用意することや，参加者の位置を完全に固定してしまうことで，注目者の推定精度を上げることが可能になると考えられる．しかし，参加人数が多くなった場合や参加者が自由に動き回る場合に対応した自動撮影を考えると，ある程度の汎用的なシステム作りが重要になる．このため，情報発生量による視覚情報と音源定位による聴覚情報の統合による撮影手法を提案したが，注目者の交代を正しく抽出するために，参加者同士が重ならないように参加者の移動範囲に制約を設けることでショットを切り換える問題などを簡単化した．この制約をなくした場合にでも注目者の交代を正確に抽出するためには，視覚情報に対しては情報発生量だけではなく画像を用いた個人識別やトラッキング，聴覚情報に対しては音源定位だけではなく音声情報を用いた個人識別などの処理を組み込んで，重なった参加者を分離して抽出する作業が必要になるが，今後の課題とする．

7. む す び

本研究では，マイク入力を用いた音源定位によって聴覚情報の抽出を行い，カメラの入力画像から情報発生量を計算することで視覚情報の抽出を行った．そして，聴覚情報と視覚情報を統合して注目量を求め，注目量の分布から注目者の選択を行った．また，注目者に基づいて，テレビ討論番組のカメラワークの知識を

利用したカメラ制御を行うことで，より効果的な会議映像をリアルタイムで自動的に撮影する手法を提案した．最後に，本手法により会議映像を生成する実験を行い，視認アンケートで評価することで本手法の有効性を確認した．

文　献

[1] 井上智雄，岡田謙一，松下　温，“テレビ番組のカメラワークに基づいた TV 会議システム，” 情処学論，vol.37, no.11, pp.2095–2104, Nov. 1996．

[2] 大西正輝，泉　正夫，福永邦雄，“情報発生量の分布に基づく遠隔講義撮影の自動化，” 信学論 (D-II)， vol.J82-D-II, no.10, pp.1590-1597, Oct. 1999．

[3] M. Onishi, T. Kagebayashi, and K. Fukunaga, “Production of video images by computer controlled cameras and its application to TV conference system,” Computer Vision and Pattern Recognition, vol. 2, pp. II-131–II-137, Dec. 2001.

[4] 高橋弘太，“視覚と聴覚を統合するシステム―センサフュージョンの具体例，” 信学誌, vol.79, no.2, pp.155–161, Feb. 1996．

[5] D. Zotkin, R. Duraiswami,and L. S. Davis, “Multimodal 3D tracking and event detection via the particle filter,” IEEE Workshop on Detection and Recognition of Events in Video, pp.20–27, July 2001.

[6] 浅野　太，速水　悟，松井俊浩，“話者方向同定と雑音抑制による音声認識性能の改善，” 音響誌，vol.53, no.11, pp.889–894, Nov. 1997．

[7] 安藤　繁，篠田裕之，小川勝也，光山 訓，“時空間勾配法に基づく 3 次元音源定位センサシステム，” 計測自動制御学会論文集，vol.29, no.5, pp.520–528, May 1993．

（平成 13 年 6 月 25 日受付，11 月 5 日再受付）