全方位映像から音楽情報へのメディア変換に基づく視覚情報の伝達

池田徹志[†] 室田健吾[†,☆] 石黒浩[†]

特定のメディアでの通信や提示が利用できない状況などにおいて，信号を異なるメディア上の表現に変換して伝達するメディア変換の研究が行われている．従来のアプローチでのメディア変換は，パターン識別処理により識別された対象に応じ特定の信号を再生する方法や，形容詞などの感性語を介して対応づける感性情報処理の方法など，抽象度の高い表現を介した変換方法であった．本論文では，信号をより直接的に変換する新しいメディア変換手法を提案する．はじめに，環境内の様子を全方位カメラで撮影した映像を利用して，多数の画像特徴を音楽特徴に対応づけることにより音楽を生成する変換手法を提案する．次に複数の全方位カメラを用い，全方位画像の特性を利用して簡易に空間的な情報を伝達する手法を説明する．提案したメディア変換の有効性と可能性を確認するため被験者実験を行い，生成された音楽のみを聞くことにより環境内の様子を理解できることを確認した．

Media Conversion from Omnidirectional Vision to Auditory Signal

Tetsushi Ikeda,[†] Kengo Murota[†,☆] and Hiroshi Ishiguro[†]

A considerable number of studies have been performed on media conversion that converts a signal in a media into another media. In past studies, the signal in a media is converted into a symbolic representation such as adjectives before represented in another media. However, much information is lost during the abstraction process. In this paper, first we propose a direct conversion method based on mapping from visual features in omnidirectional images to musical features. Then we expand the proposed media conversion method to multiple omnidirectional camera network to represent spatial information in the environment. In experiments, we have confirmed that subjects can imagine the original visual scene only by listening to the music that is generated by the proposed method.

1. はじめに

異種のメディアで表現された信号を互いに変換するメディア変換の研究が行われている．たとえば特定のメディアによる通信や提示が利用できない状況下では，メディア変換を用いて他のメディアに変換し，相補的にメディアを用いることが有効と考えられる．また，メディア変換によって生成した信号を追加することにより，複数のメディアを同時に用いた通信や提示を行い印象を強めることができ，複数のメディアの相乗的な効果を得ることが期待できる．このようなメディア変換の可能性に注目し，感性情報処理の分野を中心に活発に研究が行われている[1)]．従来のメディア変換は，変換を行う対象の表現の抽象度に注目して，大きく3つに分類できる．

(1) パターン識別に基づく方法

あるメディア上のいくつかのパターンを識別し，別のメディアのあらかじめ用意した信号を再生する手法であり，認識モデルを用いて対象を分節化し記号のレベルで異なるメディアに対応づけているといえる．視覚情報を音信号に変換して提示する研究例として，Cronly-Dillon ら[2),3)] は，白黒で書かれた線画像を単純な幾何学的形状に分解し，あらかじめ用意した対応する音を鳴らすことにより形状を認識させる手法を提案した．また小林らは音を用いてシーンの理解を行い[4)]，頭部のカメラ画像からランドマークを検出し，3次元音響装置を利用して音を用いて歩行の誘導を行った[5)]．

(2) 形容詞などの印象を表す言葉を用いる方法

メディア上に表現された信号と人間が受ける印象との関係を，感性語などの形容詞を介して関係づける方法であり，言葉のレベルでメディア間を対応づけているといえる．上野山ら[6)] は，ドラムの音と人の受ける印象との間の関係を SD 法を用いて定量的に調べた．熊本ら[7)] は，音楽の検索における印象を表す語の選択方

† 大阪大学
Osaka University
☆ 現在，西日本電信電話株式会社
Presently with NTT WEST

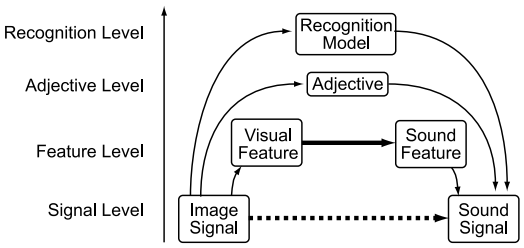


図 1 メディア変換手法の分類
Fig. 1 The approaches of media conversion.

法を提案した．山脇ら[8] は，形容詞を用いて音楽と色のイメージの類似性について議論した．

(3) 信号レベルで直接変換を行う方法

(1)，(2) の従来のメディア変換の問題点は，記号化や言語化できない情報が抽象化の過程で失われることや，事前に用意した対象の認識モデルが未知の対象には適用できないことである．異種メディア上で表現された信号は表層的にはまったく異なって見えるが，信号そのものの類似性や相関性などの関係が存在する．このような信号レベルの関係に注目することにより，新しいメディア変換が実現できる可能性がある．**図 1** は画像から音へのメディア変換の種々のアプローチを，変換対象の表現の抽象度に注目して分類したものである．

我々は図 1 の最下層の信号レベルでのメディア変換が実現できる可能性があると考え，はじめに全方位映像の情報から音信号を直接的に生成するメディア変換を提案した[9]（図中の点線）．文献 9) の手法では，全方位画像上の 1 つの円周上の輝度値の系列をそのまま音信号の時系列と見なして再生した．また環境内の対象の配置に関する情報を左右の音の位相差として表現した．異なる環境に対応する音を聴き分けて環境を識別できることを示したが，人間の聞きやすさを考慮しない変換手法であったため，生成した音を快適に聴くことができないという問題があった．そこで本論文では画像特徴を音楽特徴に対応づけることにより，映像を人間の聴きやすい音楽の形に変換する手法を提案する（図中の太線）．関連研究として，信号レベルの変換を利用して新しい楽器[10] やインタラクティブな空間[11] を作る試みが提案されている．また長田ら[12] は異種メディアの信号レベルの対応づけに注目し，音を聞くと色を知覚する色聴現象を保持する被験者に対し，音楽における調・音高・音色などのパラメータが，色の色相・明度・彩度などへどのように対応づけられるかを調査し，色聴現象を保持しない一般の人が同様の対応づけを保持する可能性を示唆した．しかし，メディア変換の情報伝達としての可能性を定量的に評価していない．本論文はこれらの研究の知見に基づき，メディア変換システムを構築し可能性を検討する試みである．

上記の色聴現象は共感覚[13] の 1 つとされる．共感覚は，あるモダリティの刺激が他のモダリティの刺激を不随意的に引き起こす現象であり，人間の知覚系において異種メディア間の相互作用が信号レベルで存在することを示唆する例である．共感覚は数千人に 1 人程度の人に現れるといわれ，普通の乳児でも生後 4 カ月までのあいだは共感覚を示すという仮説も提案されており[14]，共感覚は人間の無意識下で行われている知覚の過程が意識の上に現れた現象という解釈もある．共感覚よりも弱いモダリティ間の干渉は一般の人にも生じると考えられ，「黄色い声」「渋い色」といった言語における共感覚的表現は，それぞれ視覚と聴覚，味覚と視覚の間の干渉が反映されたものと解釈することもできる．「音色」という言葉もまた聴覚に対する視覚的な表現である．このような表現の存在は，普遍的なモダリティ間の干渉があることを示唆する．

また我々は，メディアが違うにもかかわらず信号に対して同じ印象を持つことがある．たとえば強い信号は音楽，画像，匂いなどメディアを問わず「激しい」という印象として感じられる．人間はこのように強さ，リズム，テクスチャ，方向といった刺激の物理量をメディアに依存しない印象としてとらえている可能性がある[15]．このように考えると，それぞれのメディアの様々な形式で表現された信号の中にも，メディアに依存せず同様な印象を作り出す成分があるのではないかと考えられる．

さらに，人間の知覚系における複数のメディアの信号の処理は，初期の段階から統合が行われている．たとえば，視覚が音声の認識に大きな影響を与える McGurk 効果[16]，視覚を用いることによる音声の認識精度の向上[17]，視覚，聴覚，触覚などの感覚が認識過程で互いに影響を与え合うこと[18] などが知られている．

以上のような人間の知覚系における異種メディア間の信号レベルの相互作用に注目することにより，工学的にも新しい異種メディア間の情報処理を実現できる可能性がある．工学的な信号統合の研究においては，各センサの信号に対し認識処理による抽象化が完了する以前の段階から統合を行い，より高度なセンサ統合処理を行う研究が進められている[19)~21]．

本研究ではメディア変換の実現においても，異なるメディア上で表現された信号の直接的な変換が有効であると考え，全方位カメラを用いて映像として表現さ

れた環境内の様子を，信号レベルで音楽情報に変換するメディア変換を提案する．多数の単純な画像特徴を利用することにより，対象を限定することなく，あらゆる入力映像に応じた音楽を生成することができる．また多数の全方位カメラを用いることにより，生成した音楽情報に対象の位置情報を合わせて提示し，空間的な情報を組み合わせた臨場感のある伝達を実現する．提案するメディア変換システムを構築し，生成した音楽のみを聞いた場合に環境内の様子を理解できるかを評価し，有効性を検証する．

以後，2 章では 1 台の全方位映像信号を音楽情報に変換する手法を提案する．3 章では多数の全方位カメラとスピーカを用い，空間的情報を伝える手法に拡張する．4 章で提案するメディア変換の有効性を示すために行った実験を示し，5 章で考察とまとめを行う．

2. 全方位映像から音楽へのメディア変換

2.1 提案するメディア変換の概要

全方位映像を用いたメディア変換

本研究では環境内の様子の撮影に全方位カメラを用いる．全方位カメラは下向きに設置された凸状鏡とその鏡を上向きに映すカメラで構成され，一度に全周囲の画像を撮影することができる（**図 2**）．全方位画像を用いることにより，環境内の全体の様子を 1 台のカメラで撮影することができる．

我々はすでに全方位映像の特性を利用した信号レベルのメディア変換を提案した[9] が，人間の聞きやすさを考慮していないという問題点があった．本論文では音楽の形式で音情報を表現することにより，この問題を解決する手法を提案する．全方位映像から多数の画像特徴を求め多数の音楽特徴に変換して音楽を生成することにより，多様な映像の性質を反映した人間の聞きやすい音楽情報に変換する手法を提案する（**図 3**）．

多数の特徴の組合せによる情報表現

提案手法は，複雑な画像処理や認識処理を行わず単純な画像特徴を組み合せて映像情報を表現することにより，認識誤りの影響を受けず，対象を限定せずに多様な映像情報の時々刻々の変化を反映した音楽を生成できる点が特徴である．従来のメディア変換では伝えられなかった情報を伝達することが期待できる．

対応づける画像特徴と音楽特徴の一覧を**表 1** に示す．映像情報を自然に表現する音楽を生成するため，人間が受ける印象に相関があることが実験的に検証または示唆されている対応づけを積極的に用い，人間が直感的に理解し学習しやすい自然な変換の実現を目指す．

ここで背景差分量は，環境内に移動する対象が存在しない状態で撮影した背景画像を用意し，現在の画像との画素ごとの輝度値の差分絶対値の平均である．前景とは背景との差分が一定値以上となった画像上の領域とし，人間などの対象が存在する領域に対応する．フレーム間差分量は，時間的に連続する 2 枚の画像に対して画素ごとの差分の絶対値を平均した量である．

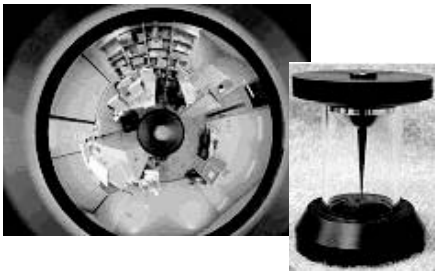


図 2 全方位画像と全方位カメラ
Fig. 2 An omnidirectional image and an omnidirectional camera.

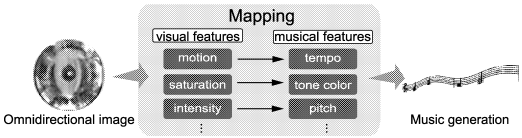


図 3 多数の画像特徴から音楽特徴への対応づけに基づくメディア変換
Fig. 3 Media conversion by mapping visual features to musical features.

表 1 画像特徴から音楽特徴への対応づけ
Table 1 Mapping from visual features to musical features.

	画像特徴	→	音楽特徴
(1)	平均輝度		調性の長短
(2)	平均色相		調性の主音
(3)	平均輝度		和音の音高
(4)	平均彩度		音色
(5)	フレーム間差分量		テンポ
(6)	背景差分量		音量
(7)	色相の差		和音
(8)	前景の平均輝度		旋律の音高

全体的な特徴と局所的な特徴に注目した変換

画像特徴と音楽特徴の中にも全体的な量と局所的な量があることに注目し，互いに同種の量に対応づけ，より直感的に理解しやすい変換を目指す．画像特徴の中でも平均輝度などは画像全体に関わる全体的な特徴量と考えられるため，音楽全体の特徴を表す調やテンポや和音などの特徴に対応づける（表 1 (1)〜(7)）．一方，前景領域の画像から求める画像特徴（表 1 (8)）は，動いている対象に対応する局所的な特徴量と考えられるため，音楽の中の旋律に関わる特徴に対応づける．

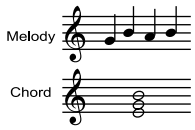


図 4 全方位画像 1 枚から生成される音楽の例
Fig. 4 An example musical sequence that is generated from a omnidirectional image.

表 2 平均輝度と調性の長短の対応づけ
Table 2 Mapping from brightness to major/minor key.

Brightness	Major/Minor
$I \geq I_0$	Major key
$I < I_0$	Minor key

表 3 平均輝度と調性の対応づけ
Table 3 Mapping from hue to key.

Hue	Key
White	C
Orange	D
Yellow	E
Green	F
Cyan	G
Red	A
Blue	B

表 4 平均輝度と和音の音高の対応づけ
Table 4 Mapping from brightness to octave.

Brightness	Octave
$I < 100$	-
$100 \leq I < 130$	+1 octave
$130 \leq I < 150$	+2 octave
$150 < I$	+3 octave

表 5 彩度から音色への対応づけ
Table 5 Mapping from saturation to tone.

Saturation	Tone
$0.16 \leq S$	～2 倍音
$0.24 \leq S$	～4 倍音
$0.31 \leq S$	～8 倍音
$0.39 \leq S$	～16 倍音
$0.47 \leq S$	～32 倍音

音楽生成の流れ

音楽の生成は 1 枚の全方位画像から 1 秒程度の長さの音楽を生成することの繰返しにより行う．1 枚の全方位画像から抽出された各種の画像特徴に基づいて音楽特徴を決定し，和音と旋律の部分から構成される音楽を生成する（**図 4**）．旋律は 4 つの音の系列とし，階名で表現したものを 50 種類用意する．画像特徴に基づいてその中から 1 つを選択し，調性の主音に応じて音名を決定し再生する☆．単調になることを避けるため，ときどきランダムな旋律を選ぶ処理を導入した．和音およびテンポなども画像特徴によって決定した．

再生が終わると次の全方位画像が取得され，同様に音楽を生成する．基本的にはその瞬間に取得された全方位画像に基づいて音楽を生成するため，長いフレーズや音楽的盛り上がりは意図的には作られない．

2.2 画像特徴と音楽特徴の対応づけ

図 1 の対応づけに基づく音楽生成の手法を具体的に述べる．提案するメディア変換の評価は 4 章で行う．

(1) 調性の長短

色の明度が高くなると，楽曲が明るい印象を受けることが報告されている[22),23)]．画像の明るさを表す平均輝度を用い，音楽の調の長短を決定する．画像全体を平均輝度値 I を経験的に定めた値 I_0 以上ならば長調，小さければ短調を対応づける（**表 2**）．

(2) 調性

音楽の調性と色の色相の対応が示唆されている[12)]ことに注目し，画像の平均色相に最も近い色に基づき調性を決定する（**表 3**）．

(3) 和音の音高

音の音高が上がると，イメージされる色の明度が高くなる実験結果が報告されている[12)]．この知見に基づき，画像の平均輝度値 I が大きいときは用いる和音の高さをオクターブ単位で高くする（**表 4**）．ここで平均輝度値 I は 0 以上 256 未満である．

(4) 音色

画像の彩度の上昇と音色の高調波成分の増加は相関することが示唆されている[12)]．この知見に基づき，画像の平均彩度を用いて音色を対応づけ，平均彩度 S が高いほど高次までの倍音成分を加える（**表 5**）．ここで彩度 S は 0 以上 1 以下である．

(5) テンポ

速い動きの映像には速い音楽が調和して感じられること[24)]，音楽と映像に現れるリズム間の引き込みが見られること[25)]などの知見に基づき，画像中の動きをフレーム間差分量 D で評価し，音楽のテンポに対応づける（**表 6**）．

(6) 音量

一般に，環境内に人間などの対象が多く存在する場合には，多くの音が聞こえる傾向があると考えられる．背景差分量がそれを反映すると考え，テンポの制御と同様に背景差分量を音量に対応づける．音量は背景差分量の一次式で決定する．

☆ 階名は調性の主音との相対的な関係を表し，音名は絶対的な音の高さを表す．

表 6 フレーム間差分量とテンポの対応づけ
Table 6 Mapping from frame difference to tempo.

Frame difference	Tempo (crotchet)
$D < 2$	40
$2 \leq D < 5$	48
$5 \leq D < 20$	60
$20 \leq D < 60$	80
$60 \leq D$	120

表 7 色相の差から和音への対応づけ
Table 7 Mapping from difference of hue to harmony.

Difference of hue	Code
$0 \leq H < 5$	Tonic
$5 \leq H < 20$	Subdominant
$20 \leq H$	Dominant

(7) 和音進行

形容詞を介したメディア変換の研究では，同じ配色とコード進行の色が類似の印象を与えることが示唆されている[26]．この知見に基づき，色相の調和によって和音進行を制御する．ここでは画像の背景部分と前景部分の平均色相の差の絶対値 H が色相の調和を表していると考え，H が小さい場合は落ち着いた印象を与える和音を，大きい場合は不安定な印象を与える和音を選択する（**表 7** ☆）．

画像の各フレームで独立に和音を決定した場合には，任意の和音の進行が現れる可能性がある．聴きやすい音楽を生成するため，次のような音楽制約を導入し，和音の進行を限定する．

(1) ドミナントの次には必ずトニックに進む．
(2) サブドミナントは任意のコードにも進行する可能性があるが，代理和音 II の次には必ずドミナント V に進む．

また，同じ和音が続く場合に単調になることを避けるため，代理和音を用いて和音を変化させる．

(8) 旋律の音高

音高が上がると，イメージされる色の明度が高くなるという知見に基づき，旋律の音の高さを前景領域の平均輝度によって定める．ここで音高と調性を独立に決定した場合には，調和した音にならない可能性があるため，決定した調性の音階中の音に制約する．

☆ トニック，ドミナント，サブドミナントはそれぞれ調の主音，調の主音の 5 度上の音（属音），属音の 1 度下の音（下属音）の上に組み立てられた和音である．たとえばハ長調（C-major）ではそれぞれ C-E-G（トニック），G-H-D（ドミナント），F-A-C（サブドミナント）の和音になる．

3. 多数の全方位映像からの空間情報を伝えるメディア変換

2 章で提案した手法は，環境内の対象の位置などの空間的情報が伝わらないという問題がある．多数のカメラやスピーカが利用できる場合には，全方位カメラネットワークによってとらえた空間的な位置を伝える変換を行うことにより，環境内の様子をより詳しく表現できると考えられる．3 章では，多数の全方位映像から環境内の対象の位置情報を抽出し，複数のスピーカの音量差を用いて位置情報を伝える手法を述べる．2 章で提案した手法に基づいて音楽を生成し，音楽の音量を各スピーカで変化させることにより，環境内の 1 つの対象の位置情報を加えて伝達する変換を提案する．複数の対象位置の伝達は今後の課題である．全体として 1 系統の音楽が生成され，全方位カメラの特性を利用することによって音量バランスがカメラごとに独立に決定される．最終的に各スピーカが出力する音量は，各スピーカでの音量の加算によって行われる．

3.1 同一配置の全方位カメラから無指向性スピーカへの変換

全方位カメラとスピーカを複数台用いた音源位置の提示を近似的に簡単に行う手法を提案する．カメラの位置を厳密に計測する必要はない．システムの構成としては以下の 2 種類が考えられる．

(1) カメラとスピーカを対にした装置を用い，記録された過去の映像情報をもとにして同じ場所で音によって提示する．
(2) カメラと同一配置になるように別の場所にスピーカを設置し，カメラの設置された環境の情報を，別の場所で音によって提示する．その際，撮影時のカメラの配置と提示時のスピーカの配置の位置関係は相似であるとする．

上記の (1) のシステムを実現するために作成したデバイスを**図 5** の左上に示す．上部に全方位カメラ，下部にスピーカを内蔵している．

3.2 全方位カメラから無指向性スピーカへの情報変換に基づく音源位置提示

全方位画像からの対象検出

背景画像との差分を計算することにより，全方位カメラに対する対象の方向を求めることができる．この様子を図 5 に示す．背景差分で得られた画像上に現れた領域に対し，最も中心から遠い点 (x, y) に基づき，カメラに対する対象の角度を求めることができる．

2 台のカメラとスピーカを用いた音源方向の提示

全方位カメラとスピーカを 2 台用い，2 台のスピー

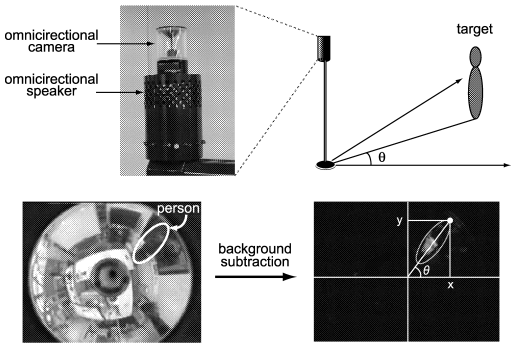


図 5 全方位画像からの対象の方向検出
Fig. 5 Position detection in a omnidirectional image.

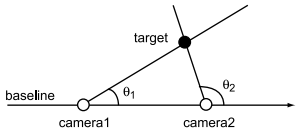


図 6 2 台の全方位カメラによる対象の観測
Fig. 6 Observing a target by two omnidirectional cameras.

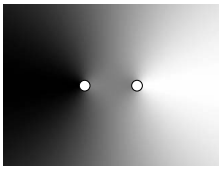


図 7 各位置での $\cos\theta_1 + \cos\theta_2$ の値
Fig. 7 Visual display of the equation 1.

カの音量の差によって音源方向を提示する手法を説明する．本論文では全方位画像の特性を用いて簡単に近似的な方向を提示する方法を説明する．

2 台の全方位カメラをある基線上に配置し，対象を観測する様子を**図 6** に示す．カメラ i を基準にした基線方向の対象の位置は，カメラからの角度を θ_i とすると $\cos\theta_i$ で表される．ここで 2 台のカメラと対象の相対的な位置は，近似的に $\cos\theta_i$ の和で表現することができる．すなわち，

$$C = \cos\theta_1 + \cos\theta_2 \quad (1)$$

は対象の基線方向の位置を近似的に表している．対象の位置による C の値の大きさを**図 7** に示す．濃度が C の値に対応し，薄い色ほど値が大きい．図より基線方向の位置がほぼ表現できることが分かる．同じ配置の 2 台のスピーカの音量 Vol_1，Vol_2 を次のように決定すると，基線方向の対象の位置を音量差で表現できる．

$$Vol_1 = V_0 - V * C \quad (2)$$
$$Vol_2 = V_0 + V * C$$

ここで V_0，V は音量の範囲を決定する適当な定数とする．

式 (2) で決まる音量は各カメラで検出された $\cos\theta_i$ の線形和であるため，複数のカメラの情報統合は各スピーカでの音量の和によって行われ，ミキシングによって実現できる．カメラ i による音量の増加分は，

$$\Delta Vol_1 = -V\cos\theta_i \quad (3)$$
$$\Delta Vol_2 = +V\cos\theta_i$$

となり，完全にカメラごとに独立に行うことができる．

さらに，図 5 上で検出する対象を人間に限定すると，図中の (x, y) は頭の位置を表し，人間がカメラに近すぎない場合には，中心から (x, y) までの距離が近似的に一定と仮定できる．このとき $\cos\theta_i \propto x$ となるので，

$$\Delta Vol_1 \propto -x \quad (4)$$
$$\Delta Vol_2 \propto +x$$

となる．すなわち，各全方位カメラは画像上での x 座標を検出し，式 (4) に従い独立にスピーカの音量を調整するだけで音源位置が提示される．ただし全方位カメラの映像は鏡像であるため座標変換が必要である．

2 台のカメラとスピーカを用いた音源位置の提示

基線と垂直な方向も含めた位置を提示するには，

$$S = \sin\theta_1 + \sin\theta_2 \quad (5)$$

を利用することにより実現できる．S は対象の基線と垂直方向の位置を表している．式 (2)，式 (3) を拡張し，次式に基づきスピーカの音量を決定する．

$$Vol_1 = V_0 + V * (-C - S) \quad (6)$$
$$Vol_2 = V_0 + V * (+C - S)$$
$$\Delta Vol_1 \propto (-\cos\theta_i - \sin\theta_i) \propto (-x - y) \quad (7)$$
$$\Delta Vol_2 \propto (+\cos\theta_i - \sin\theta_i) \propto (+x - y)$$

これは，提示された音源を聞く人が図 6 の基線の下側にいる場合に，奥行き方向を音量の大小で表すことに相当する．ただし 4 章の実験では，さらに拡張した次の手法を用いている．

多数のカメラとスピーカを用いた音源位置の提示

カメラおよびスピーカが多数ある場合には，隣接するスピーカの対ごとに独立に式 (3) に基づく音量調整を行い，音源位置を提示することができる．例として，カメラおよびスピーカが 4 台で正方形の頂点上に配置した場合（**図 8**）はカメラの対ごとに基線が異なるため，式 (3) 中の角度の基準が異なることに注意し，

$$\Delta Vol_1 \propto +\sin\theta_4 + \sin\theta_1 - \cos\theta_1 - \cos\theta_2 \quad (8)$$
$$\Delta Vol_2 \propto +\cos\theta_1 + \cos\theta_2 + \sin\theta_2 + \sin\theta_3$$
$$\Delta Vol_3 \propto -\sin\theta_2 - \sin\theta_3 + \cos\theta_3 + \cos\theta_4$$
$$\Delta Vol_4 \propto -\cos\theta_3 - \cos\theta_4 - \sin\theta_4 - \sin\theta_1$$

となる．

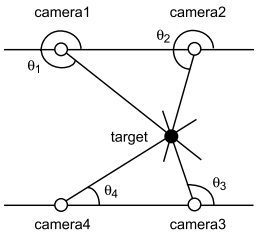


図 8 4 台の全方位カメラによる対象の観測
Fig. 8 Observing a target by four omnidirectional cameras.

4. 実　　験

提案したメディア変換手法の有効性を，構築したメディア変換システムを用いて実験的に検証した．生成された音楽情報のみを提示した被験者が元の環境の様子を推定できるかを，提案した 2 種類のメディア変換手法について実験し比較検討した．

4.1 実験に用いた全方位映像

全方位カメラ 4 台を用い，様々な場所の様子を撮影した．実験に使用した全方位画像の例と撮影環境の概要を**図 9** に示す．各環境で 2 回撮影し，実験時の試聴用およびテスト用にそれぞれ使用した．2 章で提案したカメラ 1 台を用いるシステムでは，特定の 1 台のカメラで撮影したデータを用いた．

4.2 実 験 設 定

カメラを 4 台を使う場合は，4 画面分を 1 系統の映像信号に統合して PC に入力した．4 台のスピーカを用いた再生は，RME 社製 Hammerfall DSP を用いて行った．

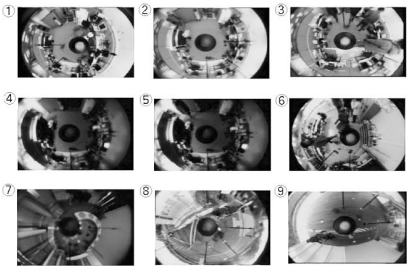
撮影に用いた全方位カメラを**図 10** に示す．撮影時の全方位カメラは 1.5 m 四方の正方形の頂点に，スピーカの位置は 2 m 四方の正方形の頂点に配置した．カメラ 4 台を使う場合は 3 章で提案した手法に基づいて各スピーカの音量を決定した．

全方位映像から音楽への変換の処理は，MAX/MSP および jitter（Cycling'74 社[27]）を用いて提案手法を実装し，PC 上でリアルタイムに変換処理を行った．同様のソフトウェア環境下では提案するメディア変換を動作させることができる．

4.3 実　　験

実験手順

被験者に対して映像情報を音楽情報に変換していることのみを説明し，各環境で撮影した試聴用の映像と生成された音楽の両方を提示した．次に，テスト用の映像から生成した音楽のみを提示し，被験者はディスプレイ上に提示された 9 つの全方位映像から対応すると思われるものを選択した．



	環境	移動対象の有無
1	明るい部屋	なし
2	明るい部屋	1 人が歩いている
3	明るい部屋	数人が歩いている
4	暗い部屋	なし
5	暗い部屋	1 人が歩いている
6	食堂	多数の人が歩いている
7	エレベーターホール	なし
8	道路沿い	車両が通行する
9	広場	なし

図 9 実験用データ中の全方位画像と撮影環境
Fig. 9 Example omnidirectional images in the experimental data and brief descriptions of the environments.

図 10 実験設定
Fig. 10 Experimental setup.

実験 1（提案手法 1，全方位カメラ 1 台使用）

2 章で提案した手法を用いた実験を 20 代から 30 代の被験者 15 人（男性 13 人，女性 2 人）に対して行った．被験者は音刺激を聴いてもその調名や音階が分からない者を対象にした．結果を**表 8** に示す．環境 6 の正答率は低かったが，環境 2，7，8 は 50%程度，そのほかは 60%以上の正答率を示した．平均は 58.5%であった．

実験 2（提案手法 2，全方位カメラ 4 台使用）

3 章で提案した手法を用いた実験を 20 代から 30 代の被験者 13 人（男性 11 人，女性 2 人）に対して行った．被験者は音刺激を聴いてもその調名や音階が分からない者を対象にした．同様に結果を表 8 に示す．環境 7 の正答率は 15%程度，環境 8 の正答率は 54%だったが，そのほかは 70%以上の正答率を示した．平均は 66.7%であった．

表 8 各データに対する平均正解率
Table 8 Correct answer rate in the experiment.

	環境	正答率 [%] 手法 1	手法 2
1	明るい部屋（無）	60	77
2	明るい部屋（1 人が歩いている）	47	70
3	明るい部屋（数人が歩いている）	60	77
4	暗い部屋の（無）	94	70
5	暗い部屋 （1 人が歩いている）	94	92
6	食堂	20	54
7	エレベーターホール	47	15
8	道路沿い	47	70
9	広場	60	77
平均		58.5	66.7

5. 考察とまとめ

5.1 考　　察

従来手法との比較

実験に用いた動画像を文献 9) のメディア変換手法で音信号に変換した場合には，環境の明るさや色相が変化した際の違いが音に現れにくいため，環境内の雰囲気などを伝えることが困難になることが考えられる．今回提案した手法は環境の多様な情報を多様な音楽特徴の形で伝達することができたと考えられる．

音楽らしさについての考察

文献 9) のメディア変換手法と比べ，提案手法では音楽の形式で情報を表現することにより聴きやすい音を生成することができた．これにより，BGM として長時間聴き続けることができる．

音楽らしさに関して，文献 9) の手法と比較して聴きやすい音になっていたが，音楽らしさという点では想像と違うという被験者の意見があった．これは提案手法では短い単位の音楽を独立に作成して再生しているため，明確な旋律がないことが被験者の期待とは違ったためと考えられる．また，提案手法の音色の変化は正弦波に高周波成分を加える割合によってのみ実現されていたため，一般的な楽器の音色とは違う点が違和感につながった可能性がある．これに対しては，ある程度の長さの旋律を維持するような特徴変換のしくみや，楽器音を導入することが有効である可能性がある．

空間的情報を伝達することの有効性

9 つの環境の正答率を実験 1，2 で比較すると，環境 5 はそれほど変化がなく，環境 4，7 で低下し，それ以外の環境では向上した．正答率が向上したデータの多くは，人の動きを含む環境であった．これは人や物の動きがある環境に対しては，空間的情報を伝えることにより環境の理解が容易になることを示していると考えられる．また正答率が低下したデータはいずれも環境内に動きのないデータであり，実験 2 ではカメラ 4 台分の映像を提示したことが分かりにくさにつながった可能性がある．

メディア変換システムの利用形態

本論文で提案するメディア変換手法の利用例として，仕事をしながら隣の部屋の雰囲気や人の動きなどの様子を何となく感じることができたり，過去に撮影された環境内の様子を音楽で聞くことなどがあげられる．いずれも視覚を用いた別の作業をしながら，視覚情報を知覚できることを利用している．

提案する画像特徴から音楽特徴への対応づけは，はじめてシステムを利用する際には分かりやすいと感じられるものではない可能性がある．この点には個人差もあると考えられる．これは提案するメディア変換のシステムをある程度の期間使い続け利用者が慣れるに従い，撮影された環境内の様子がよりはっきりと分かるようになることが期待される．提案手法では音楽の形式で情報を表現しているため，長時間聴き続けることが可能であり，利用者側がメディア変換システムに適応することができると考えられる．

たとえば画像の色相と音楽の調性の関係について，色聴などの共感覚保持者でない一般の人には調と色相の選択に再現性が見られないことが報告されている[12]．一方で，トレーニングによって再現性のある対応づけが可能なことを示唆する実験も報告されている．提案手法でも画像の色相を音楽の調性に対応づけており，はじめはこの変換が直感的ではない可能性がある．しかし利用者側が特定のメディア変換方式に利用者側が適応することにより，トレーニングの効果が期待できると考えている．どの程度適応による効果が期待できるかを検証する必要があり，これは今後の課題である．

メディア変換のパラメータの設定に関する考察

今回はメディア変換の一例を検証したが，メディア変換の変換方式や各種パラメータの値をより良いものに改良してゆくことは重要な課題である．特に調性の主音や長短など，不連続に変化する音楽特徴を決定する際のしきい値の決定や，なめらかに変化するような変換手法の検討などの課題があげられる．また音楽特徴の適切な変化の範囲も，適切なものを実験的に検証してゆくことが必要と考えている．

利用者の適応に関しても，パラメータによっては適応の容易さが異なると考えられる．どの部分を利用者の適応に任せられるのかを検討し，利用者を含めたメディア変換システムの設計指針を確立する必要がある．

利用者側がより積極的にメディア変換にフィードバックを行う方式も考えられる．たとえばはじめに映像と

生成された音楽の両方を利用者に提示し，利用者が違和感を感じた場合には変換のパラメータを変化させた例をいくつか提示し，利用者がより自然に感じられるパラメータを選択する方式が考えられる．フィードバックを利用することにより，利用者に対してシステム側が適応することが可能になる．このようなシステムの検討は今後の課題である．

5.2 まとめと今後の課題

全方位映像情報から音楽情報へのメディア変換において，多数の画像特徴を音楽特徴に変換することにより，映像信号の記号化や言語化を行わずに信号に近いレベルで変換を行う手法を提案した．また複数台の全方位カメラとスピーカを用いて，環境内の空間的な位置情報を伝達する手法に拡張した．提案手法に基づきメディア変換システムを構築し，生成された音楽のみを聴いたときに撮影された全方位映像を選択する実験を行い評価した結果，全方位カメラ 1 台の場合で正答率は 58.5%，複数の全方位カメラを用いた場合には 66.7%となり，環境の視覚的な様子を音楽によって伝達できることを示した．

今後の課題として，3 章で提案した空間位置を伝える手法は，対象が複数の場合に各位置を独立に伝えることができない点を拡張し，環境内を移動する複数の対象の位置を伝える変換に拡張する予定である．

今回提案した特徴間の対応づけは，画像特徴と音楽特徴の変化によって人間のイメージに相関があることが報告あるいは示唆されている特徴を選択して対応づけた．しかし先行研究の報告では異なる対応づけを示唆するものもある．今後は対応づけを変化させた際の影響も調べる必要がある．

また，今回の実験ではメディア変換の可能性を示すことを主眼に実験を行った．残る問題として，環境の様子の伝達という課題に対し各特徴が貢献する程度や特徴の組合せによる影響を評価する問題がある．ひきつづき詳細な検証を行い明らかにしてゆきたい．

参 考 文 献

1) 岩宮眞一郎：音楽と映像のマルチモーダル・コミュニケーション，九州大学出版会 (2000).
2) Cronly-Dillon, J., Persaud, K.C. and Blore, R.: Blind subjects construct conscious mental images of visual scenes encoded in musical form, *Proc. R. Soc. Lond. B*, Vol.267, pp.2231–2238 (2000).
3) Cronly-Dillon, J., Persaud, K. and Gregory, R.P.F.: The perception of visual images encoded in musical form: A study in cross-modality information transfer, *Proc. R. Soc. Lond. B*, Vol.266, pp.2427–2433 (1999).
4) 小林　真，太田道男：音像定位を利用した能動的情報取得が可能な視覚代行装置，バイオメカニズム学会誌，Vol.21, No.1, pp.39–42 (1997).
5) 小林　真，太田道男：全方位センサと 3 次元音響を利用した視覚障害者用歩行誘導システム，バイオメカニズム学会誌，Vol.24, No.2, pp.123–125 (2000).
6) 上野山努，樫村雅章，小沢慎治：ドラム音の音色における感性情報と工学的パラメータの対応づけ，音響誌，Vol.49, No.10, pp.671–681 (1993).
7) 熊本忠彦，太田公子：印象に基づく検索のための印象語選定法の提案，情報処理学会論文誌，Vol.44, No.7, pp.1808–1811 (2003).
8) 山脇一宏，椎塚久雄：音楽とカラーイメージの類似性について，情報学音楽情報科学研報，2002-MUS-047, pp.105–109 (2002).
9) 港　隆史，関戸智史，石黒　浩，河原英紀：全方位視覚の特性を利用した画像から音信号への変換，日本ロボット学会学術講演会予稿集 (2002).
10) 米澤朋子，間瀬健二：流体による楽器インタラクションの考察〜Tangible Sound #2 における展開，日本バーチャルリアリティ学会論文誌，Vol.5, No.1, pp. 755–762 (2000).
11) Eng, K., et al.: Ada — Intelligent Space: An artificial creature for the Swiss Expo.02, *Proc. IEEE Int. Conf. on Robotics and Automation* (*ICRA2003*), pp.4154–4159 (2003).
12) 長田典子，岩井大輔，津田　学，和氣早苗，井口征士：音と色のノンバーバルマッピング—色聴保持者のマッピングルール抽出とその応用，信学論，Vol.J86-A, No.11, pp.1219–1230 (2003).
13) Cytowic, R.E.: Synesthesia: Phenomenology And Neuropsychology, *PSYCHE*, Vol.2, No.10 (1995).
14) Baron-Cohen, S.: Is there a normal phase of synaesthesia in development?, *PSYCHE*, Vol.2, No.27 (1996).
15) 小嶋秀樹，矢野博之：発達ロボティクスからみたロボット聴覚研究，人工知能学会研究会報告，SIG-Challenge-0318-2, pp.7–12 (2003).
16) McGurk, H. and MacDonald, J.W.: Hearing lips and seeing voices, *Nature*, Vol.264, pp.746–748 (1976).
17) Grant, K.W. and Seitz, P.: The use of visible speech cues for improving auditory detection of spoken sentences, *J. Acoust. Soc. Am.*, Vol.108, No.3, pp.1197–1208 (2000).
18) Shimojo, S. and Shams, L.: Sensory modalities are not separate modalities: plasticity and interactions, *Current Opinion in Neurobiology*, Vol.11, No.4, pp.505–509 (2001).
19) Coen, M.H.: Multimodal Integration — A Bi-

ological View, *Int. Joint Conf. on Artificial Intelligence*, Vol.2, pp.1417–1424 (2001).
20) Hershey, J., Ishiguro, H. and Movellan, J.R.: Audio Vision: Using Audio-Visual Synchrony to Locate Sounds, *Proc. Neural Information Processing Systems* (*NIPS'99*) (1999).
21) Ikeda, T., Ishiguro, H. and Asada, A.: Sensor fusion as optimization: maximizing mutual information between sensory signals, *Proc. 17th Int. Conf. on Pattern Recognition* (*ICPR 2004*), Vol.2, pp.501–504 (2004).
22) 梅本堯夫：音楽心理学，誠信書房 (1966).
23) 安達太郎，岩宮眞一郎：色彩と音楽とが互いに及ぼす影響–ショパンのエチュードを手がかりに，日本音響学会九州支部学生のための研究会 (2003).
24) 菅野禎盛，岩宮眞一郎：音楽のリズムと映像の動きの同期が音楽と映像の調和に及ぼす効果，音楽知覚認知研究，Vol.5, No.1, pp.1–10 (1999).
25) 長嶋洋一：音楽的ビートが映像的ビートの知覚に及ぼす引き込み効果，芸術科学会論文誌，Vol.3, No.1, pp.108–148 (2004).
26) 北島吾郎，土居元紀：画像の構成を手掛かりとした音楽の検索，情報処理学会関西支部大会ビジュアルインフォメーション研究会，A-15 (2003).
27) Cycling'74. http://www.cycling74.com/

(平成 18 年 5 月 8 日受付)
(平成 18 年 10 月 3 日採録)

池田 徹志（正会員）

1972 年生．1994 年京都大学理学部卒業．1997 年同大学大学院工学研究科修士課程修了．同年三菱電機（株）入社．画像認識の研究開発に従事．2003 年大阪大学大学院工学研究科博士後期課程研究指導認定退学．2005 年大阪大学大学院工学研究科知能・機能創成工学専攻特任助手となり現在に至る．知覚情報基盤の研究に興味を持つ．

室田 健吾

1982 年生．2004 年大阪大学工学部応用理工学科卒業．2006 年同大学大学院工学研究科知能・機能創成工学専攻修士課程修了．同年西日本電信電話株式会社入社．

石黒 浩（正会員）

1963 年生．1991 年大阪大学大学院基礎工学研究科物理系専攻修了．工学博士．同年山梨大学工学部情報工学科助手．1992 年大阪大学基礎工学部システム工学科助手．1994 年京都大学大学院工学研究科情報工学専攻助教授，1998 年同大学大学院情報学研究科社会情報学専攻助教授．この間，1998 年より 1 年間，カリフォルニア大学サンディエゴ校客員研究員．2000 年和歌山大学システム工学部情報通信システム学科助教授．2001 年より同大学教授．1999 年より，ATR 知能映像研究所客員研究員．現在，大阪大学大学院工学研究科知能・機能創成工学専攻教授および ATR 知能ロボティクス研究所客員室長．知能ロボット，アンドロイドロボット，知覚情報基盤の研究に興味を持つ．