# 視覚障害者のための MIDI による画像認識方法の検討

筑波技術大学保健科学部情報システム学科

大西 淳児　小宮 厚一　小野　 束

**要旨**：視覚に障害がある人が物体の形状や画像の情報へアクセスする代表的な方法として、触覚に頼る方法が一般的に用いられる。しかしながら、触覚による認知の場合、認識に要する負担が大きく、また、複雑な画像になればなるほど認識が非常に困難になる。そこで、この研究では、画像を MIDI メッセージで作られた音信号に変換し、画像の内容を音色で感じ認識する方法の検討を行った。その原理は、聴覚に障害のある人が利用する手話通話とまったく正反対の変換、つまり、画像から音信号への変換を試みるものである。この報告では、画像の構造情報に依存した MIDI メッセージ符号を作り出し、MIDI メッセージから作られた音色によって、画像の内容を理解する方法を述べ、音信号で画像を感じ・認識へ発展させる可能性についての検討を行う。

**キーワード**：視覚障害、画像認識、画像処理

## 1. まえがき

視覚に障害のある人が下界の事物を観察・理解しようとする場合、触覚は極めて重要な役割を果たしている。たとえば、図形、ダイヤグラム、算数・数学におけるグラフ、職地図、点字の読みなどは、触覚的に理解することが多く求められている。触覚提示において、実用化されているものとしてはレーズライタやピンディスプレイなどがある。レーズライタは、引っ掻くとその部分が浮き出てくる特殊な紙である。ピンディスプレイは、複数のピンが上下することで立体図形を表現するものである。このように、従来、触覚提示に着目した研究が盛んに行われている [1]。この触覚による提示においては、図形があたかもそこにあるかのように受身的に提示することに主眼をおいていて、認識率や認識速度に関してはそれほど着眼していないことが多い。また、触覚感覚特性を考えると、空間解像度が低く、複雑な図形の認知にはあまり適していないため、複雑な事物を触覚で認識するには限界があると考えられる。

一方、文字、動作・表情語、点字、結縄文字、手旗信号、合図などに代表される視覚言語がある。手旗信号を例にとってみると、文字の音情報を手旗の振り規則で対応させていて、工学的な観点からすると、この法則は、文字音声の情報を手旗動作の映像情報にメディア変換していることと同等と言える。

そこで、この報告では、視覚に障害のある人が MIDI 音声によって、画像に描かれた図形を認識する方法について検討を行う。その原理は、離散コサイン変換（DCT）によって画像の構造情報を表す要素を取り出し、その要素から MIDI メッセージを作り出すものである。この方法によって作られる音声信号は、楽譜音であるためなじみやすく、また、図形の構造の違いを細かく表現することが可能である。そのため、触覚で困難であった、複雑な図形の認識にも適した方法として期待ができる。

## 2. MIDI とは

MIDI（Musical Instrument Digital Interface）とは、音楽信号を保存・再生する規格の一つである [2]。MIDI データは、音の高さ（音高）、ゆらぎ、大きさの３つの演奏情報で構成されており、ピアノやギターなどの音源の種類、キー、テンポなどは利用者が自由に選択できる。また、MIDI は各楽器の音色のデータをもたないため、容量が少ないという特徴もある。そのため、MIDI はカラオケや音楽制作などの分野で世界的に利用されている。

MIDI メッセージは複数のパート別に独立したコントロールをするためのチャンネルメッセージと、MIDI システム全体をコントロールするためのシステムメッセージに大きく分かれている。

チャンネルメッセージでは、最大16パートをコントロールするため、MIDI チャンネルという概念を用いている。

チャンネルメッセージはノート（音符）情報などのボイスメッセージとボイスメッセージの受信状態を設定するモードメッセージに区別される。ボイスメッセージには、ノートオン・オフ（音を出す・止める）、プログラムチェンジ（音色切替）、ピッチベンド等の設定を行う。

一方、システムメッセージは、シーケンサやリズムマシンなどにおける同期関係を扱うコモンメッセージとタイミングクロックなどのリアルタイム処理などを行うリアルタイムメッセージ、音色パラメータなど機器によって統一できないメッセージのためのエクスクルーシブメッセージがある。

提案する方式では、この MIDI メッセージ内の音を作り出すノートオンメッセージを画像の構造情報から作り出す。そこで、ノートオンとノートオフのメッセージ構成の説明をしておく。

MIDI による演奏情報は、「鍵盤を押す（ノートオン）」「鍵盤を離す（ノートオフ）」の動作をノートオンメッセージとノートオフメッセージで行う。ノートオンメッセージで発音された音はノートオフメッセージによって消音されなければいけない規則になっており、ノートオンメッセージとノートオフメッセージは必ず対になっていなければならない。つまり、ノートオンで発音された音に対してノートオフメッセージが送られてこないと音が鳴りっぱなしの状態になってしまう。

ノートオンメッセージは、図１に示すように、ステータスバイトにつづき、２バイト分のデータバイトで構成される。ステータスバイト9n の「9」はこのメッセージがノートオンメッセージであることを示し、「n」は任意のチャンネル番号が入る。２バイトで構成されるデータバイトの１バイト目は音階をあらわすノート番号、２バイト目は音の強さを示すベロシティである。ノート番号は音階を数字で表現したもので、ピアノの中央のドを番号60として、半音ごとに一つずつ増減していき、一番低い音が０、一番高い音が127となっている。２バイト目のデータバイトのベロシティは、鍵盤を押す速さ（楽器を鳴らす速さ）をその音の強さとみなして、その値をあらわしたパラメータである。ベロシティは最大で127、最小で０となるが、０のときは消音を意味しノートオフメッセージと同じ効果になる。

| メッセージ | ノートオン(note on)：チャンネルボイスメッセージ |
|---|---|
| フォーマット | 1:ステータス　2:データ　3:データ<br>3byte　9　(n)　ノートNo.　ベロシティ<br>n　：チャンネル１～16(0h～Fh)<br>ノートNo.　：０～127(0h～7Fh)<br>ベロシティ　：０～127(0h～7Fh) |
| 動作 | 発音(ベロシティ0のとき消音) |

図１　ノートオンメッセージ

| メッセージ | ノートオフ(note off)：チャンネルボイスメッセージ |
|---|---|
| フォーマット | 1:ステータス　2:データ　3:データ<br>3byte　8　(n)　ノートNo.　ベロシティ<br>n　：チャンネル１～16(0h～Fh)<br>ノートNo.　：０～127(0h～7Fh)<br>ベロシティ　：０～127(0h～7Fh) |
| 動作 | 発音停止(消音) |

図２　ノートオフメッセージ

一方、発音された音を止めるには、ノートオフメッセージを使う。ノートオフメッセージは、図２に示すような構成になる。ステータスバイト8n の「8」はノートオフメッセージであることを示し、「n」は任意のチャンネル番号を設定する。データバイトの１バイト目はすでにノートオンメッセージによって発音されているノート番号を指定する。ノートオフベロシティはノートオンベロシティとは反対で、音を消す速さを表す。

### 3. 画像構造情報の取得

画像を構成する信号と音声を構成する信号は、全く異なるものであり、互いに関連するものはない。この二つの信号を相互に変換することを考えたときに、特に、問題となるのは、その情報量の圧倒的な差である。画像の信号は２次元の信号であるが、音声の信号は１次元の信号で構成されている。そのため、画像の信号を単純に音声の信号に置き換えてしまうと、非常に大きな情報量を持った音声信号になってしまう。このことは、一つの画像を音声信号で伝える時に要する時間を増大させてしまうことを意味する。

一方、画像の情報量が膨大であることから、画像の圧縮に関する研究が盛んに行われてきた。これらの研究のほとんどは、限られた帯域で効率よく画像を伝送することを目的としたもので、現在、その成果は、ディジタルカメラによる画像の保存やディジタル映像伝送の分野など数多くの分野で実用化されている。

ここで、静止画像の圧縮に着目してみると、静止画像に対する画像圧縮の代表的なものには、JPEG や JPEG2000 といった符号化方式がある。この符号化方式は、画像の統計的な性質に基づいて画像符号化方式であり、画像の統計的な冗長を DCT や Wavelet 変換によって削減するものである。このような符号化方式によって、画像の情報は冗長さを削減することができ、より少ない情報に置き換えることができる。画像を音声信号へ置き換えることを考えたときに、この特徴は非常に有利に利用できる。ところで、統計的な性質を利用した画像圧縮符号化では、圧縮率を高くした場合に、定常性を満足しないエッジの周辺において視覚的に大きな歪みを生じることが知られている。最近の普及が進んでいる地上デジタルハイビジョンテレビで映像のエッジ付近をよく観察すると、ときどき、この歪みを認識できる。これは、画像にとってその特徴を表す重要な情報が冗長なものと判断され、画像から失われてしまっていることが原因となっている。そこで、齋藤らは、画像に対する DCT によって得られた DCT 列から、画像の幾何学的構造を表現するための特徴を抽出・保存した情報と雑音信号系列を用いて、符号化を目的とした画像表現方法を提案している [3]。この提案では、画像を DCT 変換した際に得られる DCT 係数の符号情報によって、画像におけるエッジ等の位置を表現可能であるという特徴を利用したもので、DCT 係数の正負符号は、画像の幾何学的構造におけ

る大局的特徴を表現していると報告している。

画像情報からMIDIメッセージを作る際のノート符号を作成することを考えると、この特徴は非常に都合がよい。なぜならば、画像の構造情報が正負符号で表現できるために、画像の情報を単純なビット表現で表すことができるからである。そこで、この報告では、DCT係数の正負符号の情報からMIDIの鍵盤音の情報に相当するノートNoを作成することによって、画像を音声表現化することを試みてみる。

## 4. 画像から音声への変換原理

画像の情報からMIDIメッセージを作成する原理について述べよう。

まず、画像$f(x,y)$に二次元DCTを実行して得られた係数を、$F(j,k)$とする。$F(j,k)$について、$j,\ k$における正負符号のみを取り出したものを、DCT符号係数

$$G(j,k)=sign(F(j,k)) \tag{1}$$

として定義する。ここで、$sign(x)$は、

$$sign(x)=\begin{cases} 1 & (x\geq 0) \\ -1 & (x<0) \end{cases} \tag{2}$$

である。このとき、$G(j,k)$に対して逆離散コサイン変換(IDCT)を適用したときに得られる画像$g(j,k)$は、原画像の振幅情報が欠落している状態であっても、画像のエッジ等の位置を表現できる[4]。その理由は、DCTの符号係数は、離散フーリエ変換における位相成分に相当する情報を保持しているためである。したがって、DCTの正負符号$G(j,k)$は、画像の幾何学的構造における大局的特徴を表していると考えることができる。

ここで、この符号の正負を1、0のビット列として取り扱うことで、MIDIメッセージのノートNoを構成することを試みる。

ノートNoは、第2節で解説したように、7ビットの情報から構成できるため、$G(j,k)$で得られた符号情報から直接MIDIメッセージを作成すると相当な量の鍵盤音が作られてしまう。この場合、画像の情報を表す鍵盤音の演奏が長くなってしまい、画像の判別に要する時間がかかることや鍵盤音を記憶する負担が大きくなりすぎる問題が生じてしまうなどの問題を生じてしまうと考えられる。この問題に対処するため、この報告では、$G(j,k)$で得られる符号列のうち、画像の基本的に重要な構造を示す低周波数から中間周波数領域に相当する部分の符号列を対象としてノートNoを構成することにした。

まず、ノートNoの構成に使うDCT符号列$G(j,k)$において、$0\leq j\leq s$及び$0\leq k\leq s$の範囲の符号列を取り出す。この符号列のマトリックスをラスタスキャン走査して得られる正負号列$M(n)$を用意する。$M(n)$の符号を式(3)により、1または0で構成される符号列$M'(n)$を得る。

$$M'(n)=\begin{cases} 1 & (M(n)>0) \\ 0 & (M(n)<0) \end{cases} \tag{3}$$

$M'(n)$にビット列から7ビットずつ取り出し、ノートNoのデータを作成する。このとき作成されるノートNoの数量はパラメータ$s$に依存する。つまり、$s$が大きくなるほど作られる鍵盤音の数が増えることになる。

図3に画像から音声への変換プロセスを示しておく。

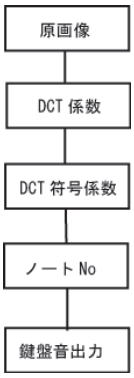



図3　画像から音声への変換プロセス

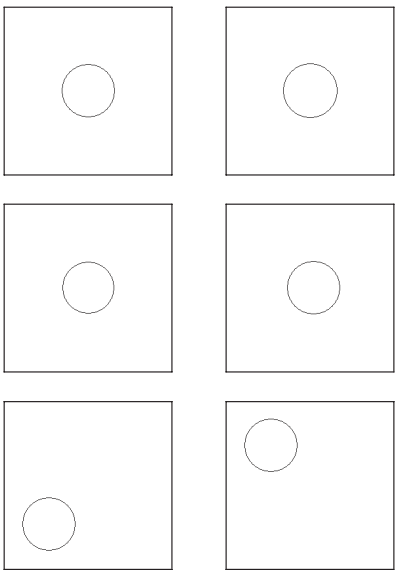


図4　テスト画像

## 5. 提案手法による鍵盤音評価

第4節の解説で示した方法によりディジタル静止画像からDCT符号係数を抽出することによって作られるMIDIメッセージを作成して、実際にどのような音になるか試験

してみた。

まず、入力画像として図４に示す複数の256×256画素の円形図形を描いた画像を用いた。また、MIDI メッセージの作成においては、パラメータ $s$ を14、また、一つの鍵盤音の鳴らす時間を100ms とした。

図４の(a)は、基本となる円形図形で半径が40画素で中心を画像の中央にしたものである。このときに、出力される鍵盤音は、図５(a)になる。このグラフは、縦軸にノート No を横軸に出力される鍵盤音の数を示している。この評価では、パラメータ $s$ を14としているので、出力される鍵盤音は、28個となる。また、ノート No は、ドレミファソラシドといった音符を表す番号となるので、音の階名としての意味を持つ。ここでは、一つの音を100ms 鳴らすことになる。

図４の(b)は、円の中心位置は同じで、半径を41画素にしたものである。見た目上は、ほとんど違いを見分けることができないが、微妙に円の大きさが大きくなった状態である。このときに、提案する手法で出力される鍵盤音は、図５の(b)となる。このグラフを図５の(a)と比較すると、鍵盤音の後半でノート No が異なって現れていることが分かる。実際に発音される音は、前半はほとんど同一に聞こえるが、終わり付近になると異なった音がでてくるのが確認できる。そのため、画像信号は、ほとんど似ていて、微少な違いがあるととらえることも可能と考えられる。

図４の(c)は、円の中心位置は同じで、半径を39画素にしたものである。これは、図１(a)の円形の半径を１画素小さくしたものである。このときに、出力されるノート No は、図５(c)になる。この図を観察すると、ほとんどの鍵盤は、同じ音になっているが、中盤から後半の手前の音が異なって出力されることがわかる。出力される鍵盤音を聴取した場合、上記の１画素半径を大きくした円と同様に全体的には、似ている音に聞こえるが、局所的に違う音が混じっていることが確認できる。そのため、聞いた場合の印象は、図４(b)と同様になると考えられる。

一方、図４(d)の図形は、図４(a)の図形を５画素右へ平行移動したものである。このとき、出力されるノート No は、図５(d)となる。この図と図５(a)を比較すると、出力される鍵盤音がほぼ同一であるとみることができる。実際に、音を聴取してみたところ、半径を１画素分、大小変化させた場合と異なり、違いを聞き分けるのは、ほとんど困難であった。この結果を踏まえて、さらに10画素の右へ平行移動させてみたが、同様の結果が現れ、20画素右へ平行移動させると、極端に違う鍵盤音が出力された。その結果を、図６に示しておく。このような結果になる原因の詳細については、現在、調査中であるが、DCT 変換の低周波に近い成分を利用して、鍵盤音を作成していることから、平行移動した場合の違いの差が DCT の正負符号の状態に表れないことが原因と思われる。

図４の(e)は、円形図形の中心を左へ60画素、下へ60画素移動させた画像である。このときの、鍵盤音の出力は、図５(e)となる。このケースでは、図形の表示位置が大きくずれることが顕著に鍵盤音の出力にも反映されており、グラフからもその差が明確に分かる結果となった。これは、図４の(f)のケースでも同様の結果を得ている。図４の(f)は、中心を左へ60画素、上へ60画素移動させたもので、その出力鍵盤音（図５(f)）でも、大きく異なる音が出力される。この二つのケースについては、実際に聴取した場合に、極めて容易に違いが判別される音となった。

## 6. むすび

この報告では、視覚障害者のための MIDI による画像認識方法の検討を行った。画像の情報が膨大であることから、従来からある画像圧縮の技術を利用して、短い MIDI メッセージを作り、短い時間で画像を認識する方法を試みた。現状での実験では、まだまだ有効性の実証には及ばないが、実際に出力される鍵盤音について調査したところ、画像の判別に利用できる可能性があると分かった。

一方、短時間で画像識別を可能にするために、画像の構成情報のより細かい情報を切り捨てており、微少な平行移動などのケースにおいて、違いの認識が困難となる結果がでた。また、逆に細かい情報は必要とせず、大局的な形だけを判別したい場合、たとえば、大きさはどうであれ、形が四角であれば、四角の図形が描かれた画像と認識したいといったとき、提案する方法をそのまま適用しても認識は容易でないなどの問題もある。

音で画像を認識するといった方法を考えた場合、もともと、性質のまるで異なる信号であるため、触覚認識と異なり、提案する方法では、物理的に図形があたかも存在するような形での提示はできない。しかしながら、画像を信号としてとらえたときに、画像の信号状態がどうなっているのかというのを知り得る一つの手段として利用できるのではないかと考えている。したがって、この方法を利用することで、たとえば、文字や漢字の形状を認識や、印刷された文字のフォントの違いを音で認識させ、かつ、フォントを使い分ける意味となる感性的情報を融合させることで画像に含まれる言語音声で表現困難な情報を伝える手段としての応用などさまざまなアプリケーションの開発への可能性が開けると考える。

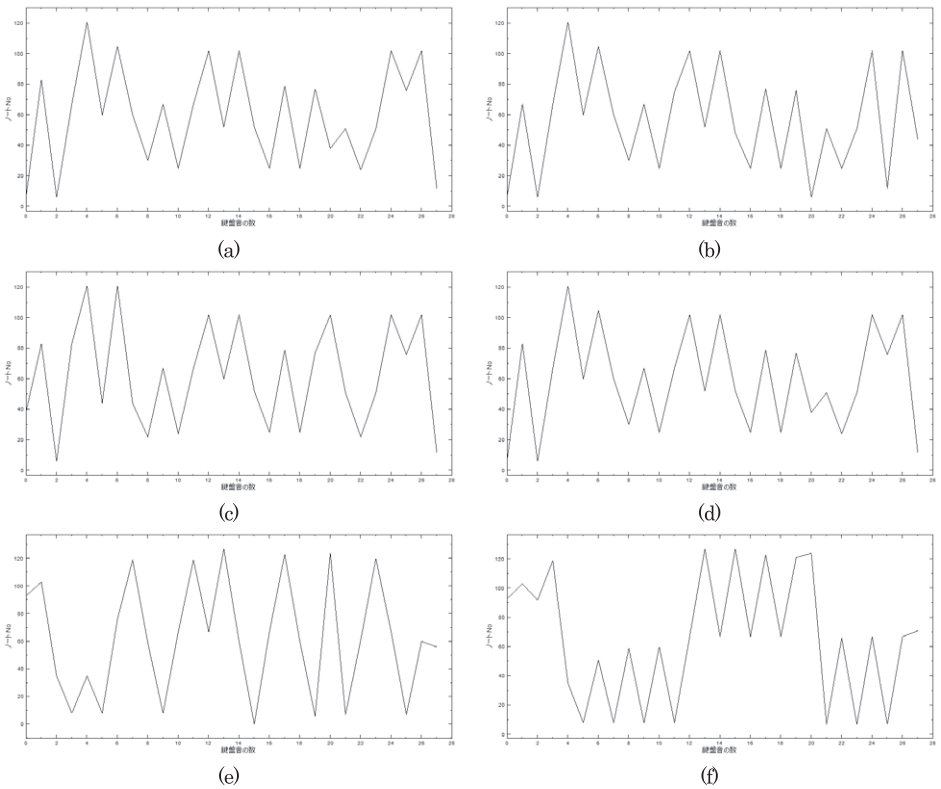



図 5　類似 画像におけるノート No 構成

今後は、この報告の成果を生かして、さまざまな応用分野への適用と図形を音声によって認識することによる新たな娯楽分野への適用などを考えていく予定である。

**謝辞**

本研究は科学研究費（21531009）及び、筑波技術大学保健科学部長裁量経費の支援のもとで実施した。ここに感謝を申し上げる

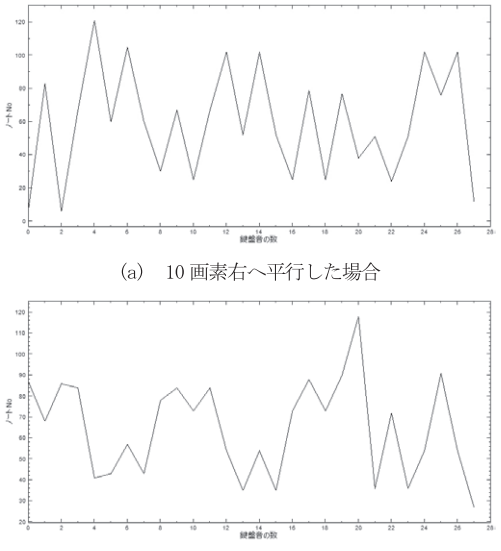
**参考文献**

[1] 渡辺哲也，久米祐一郎，伊福部達：「触覚マウスによる図形情報の識別」映像情報メディア学会誌，Vol.54，No.6，pp.840-847，June 2000.

[2] http://en.wikipedia.org/wiki/Musical_Instrument_Digital_Interface

[3] 齋藤太郎，亀田昌志：「DCT 係数における構造的特徴の保存に基づいた画像表現手法の提案」電子情報通信学会論文誌 D，Vol.J91-D，No.8，pp.1967-1970，Aug. 2008

[4] 伊藤泉，藤吉正明，貴家仁志：「DCT 係数の正負符号と位相限定相関との関係について」電子情報通信学会論文誌 A，Vol.J90-A，no.7，pp.567-577，July 2007



(a)　10 画素右へ平行した場合

(b)　20 画素右へ平行移動した場合

図 6　図 1 (a)を右へ平行移動した場合の鍵盤音

National University Corporation Tsukuba University of Technology Techno Report Vol.17 (2), 2010

# A Study on Image Recognition by Musical Instrument Digital Interface for People with Visual Impairment

ONISHI Junji, KOMIYA Koichi, ONO Tsukasa

Faculty of Health Science, Tsukuba University of Technology

Abstract: In most cases, people with visual impairment use tactile devices to recognize objects. However, it is difficult to recognize the detailed shape of objects by using tactile devices. To solve these issues, image recognition using musical instrument digital interface is proposed. Our method is based on the principle behind the use of sign language by hearing-impaired people. In sign language, voice sounds are converted to visual images. Our goal is to develop a method to convert digital images to sounds by using MIDI in order to enable visually impaired people to recognize images.

Key words: Visually impaired, Image recognition, Image processing