2005.01.14

ユビキタス領域周辺の技術　特集

音声認識と画像認識技術

Voice and Image Recognition Technologies

望月 博文　Hirofumi Mochizuki
●研究創発センター　システム技術研究室

Abstract

What do most people envision when they hear the terms voice recognition and image recognition technologies? Is it the visual functions or aural functions of robots? It can be said that voice recognition and image recognition technologies are technologies involving the reproduction of some portions of human recognition function in mechanical devices. However, these are areas of technology about which many people might think they understand the general concepts but in fact there are many mistaken conceptions. These misunderstandings come from the fact that the process of recognizing things at a single glance is very simple for human beings and the fact that we are aware that computers can perform complex and difficult calculations. In this paper we present a simple discussion of how much can be done in the areas of voice and image recognition technologies, how these functions are used in the world around us and what uses will be made of them in industry in the future.

1 はじめに

音声認識・画像認識技術というと、どんなことを想像するだろうか？ ロボットの視覚、聴覚だろうか？ 音声認識・画像認識技術は、人間が持っている認知能力の一部を機械で実現しているとも言えるのだが、分かっているようでも意外と誤解の多い技術領域である。それは、人が一目瞭然に分かることは簡単なことで、コンピューターは複雑で難しい計算ができるという認識があるからである。ここでは、音声認識・画像認識とはどの程度のことができるのか、私たちの身近なところでどのように使われているのか、今後の産業応用はどうなっていくのかについて簡単に述べる。

2 音声認識とは

音声認識の究極的な目標としては、人間が普段使っている言葉（音声）を機械が理解することだと言える。しかし、多くの研究者が膨大な努力をしているにもかかわらず、未だ実用的なシステムにはなっていない。そこで音声認識の中のコマンド認識に焦点を絞ることによって、近年、実用化につなげている例が見られるようになってきた。音声認識はなぜ難しいのだろうか。以下に自然な会話例を示して説明する。

A「それついていないのですか？」
B「これはまだ、これは古い形のなので、まだもっていないのですけど」
A「私、持ってる」
B「どんなんですか、あれ」
A「まだねぇ」

B「まだ使わないんですか、あれ」

これは、Aさん、Bさんが、実際に携帯電話について会話しているところを正確に抜き出したものである。前後関係も分からず会話を途中からみると、何を話しているかよく分からない。実際の会話は、文法的には正しくないところがあったり、省略されている部分があったりする。また、Aさん、Bさんには共有できている「これ」「あれ」が第三者には何を指しているのか分からない。コンピューターでは、なおのことである。また、同音異義語がどの単語を指しているか(ex. はし＝橋、端、箸、嘴???)や、同じ単語でも複数の意味がある場合(ex. 持つ＝①手に入れて保つ　②身につける　③所有する　④身にそなえる・・・)は、人の場合、前後関係やその言葉が使われている状況を元に判断している。このように連続的に自然に発話している音声を正しく認識することは、極めて難しい課題である。現実的には、ある使用目的に絞ってコマンド認識で音声認識の実用化を試みている。

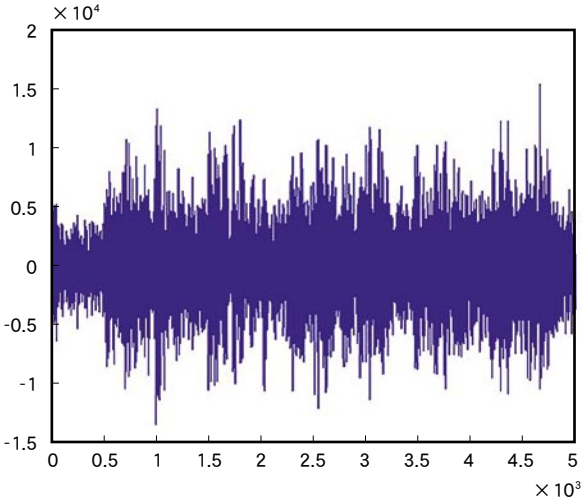

図１　自動車内の音声波形

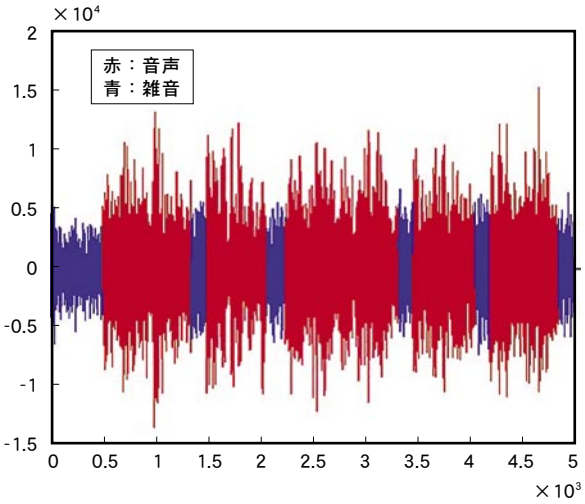

図２　自動車内の音声波形

しかし、コマンド認識に絞っても、発音、イントネーション、発話速度などの個人差、男女差、年齢差がある。さらに同一個人の中でもその時々によって変わっている。同じ騒音環境で音声認識率90％以上をマークしていたにもかかわらず、別の被験者になると20～30％の音声認識率に落ちてしまうことも珍しくはない。また、周りの雑音にも大きく影響を受ける。**図1**は、自動車の中で「今日の天気は」と発話している時の音声波形である。どこが音声で、どこがその他の雑音だろうか？ それを示したのが**図2**である。

このように雑音が混じった音声を、いかに認識するかが大きな課題となっている。音声の部分の波形も音声に雑音が重畳され、静かな環境の音声とは異なった波形になっている。一般に雑音環境下の音声認識を行うためには、使用が想定される環境の雑音を含む音響モデルを作ることになる(ex.自動車内での音声認識には、自動車用音響モデル)。また、雑音と音声が区別できなければ、雑音を音声として処理して認識ミスの原因になる。そのため、雑音成分の除去や低減をしたり、音声の部分だけ抜き出したりする(発話区間検出)技術が一般に使われている。一定の回転数で回るエンジン音など、時間的に変動の少ない定常雑音であれば、Spectral Subtraction法などで比較的効率よく除去できるが、非定常雑音に対応する技術の実用化は今後の課題となっている。

3 音声認識の応用分野

音声認識技術は、コマンド認識に絞ることによって、カーナビ、パソコンの操作、コールセンターの自動受付などで使われるようになってきた。また、最近では携帯簡易翻訳機への搭載も試みられてもいる。自動車の情報化分野では運転を妨げないようにするため、ハンズフリーのインターフェースが要求されている。その実現手段のひとつとして、音声認識は、ますます注目を集めるだろう。現在の技術では、個人差に対して十分に対応できていないので、人によっては使いにくいと感じることもある。その中でも、コールセンターでの応用例は高い認識率を実現しており、成功例と言える。

4 画像認識とは

機械、すなわち、コンピューターで画像を認識する（以下、画像認識と略す）ということは、どのようなことだろうか？ **図3**の写真を見ると、どのようなことが分かるだろうか？「ある建物の通路を2人が歩いている。」「この2人は白人女性のようだ」「時間は昼で寒くはない」等々。人であれば、この1枚の写真から様々なことが分かる。しかし、以上の人間ならなんでもないことも、画像認識では極めて難しい課題になる。この写真に写っている様々なものを一般的に認識するためには、膨大な知識と計算量が要求され、現在の技術では実用的なものにはならない。このため、通常、画像認識では認識対象を特定している。認識対象以外の背景はノイズなので、写っていない方がよいぐらいである。

図３　人間には簡単でもコンピューターに不得意な画像例
（Q：何が写っている?）（コンピューターA：???????）

図４　人間には難しくコンピューターが得意な画像例
（Q：色の付いた部分は全体の何%?）（コンピューターA：81.59%!）

一方、**図4**の色の付いた部分の面積を知りたいといった課題は、画像認識の得意な分野である。こういった画像を大量・高速・高精度に処理するのは、いたって容易である。

最近、注目されている画像処理の応用として、人の認識がある。生体認証としての人認識を考えると、指紋認識、掌紋認識、虹彩認識、静脈認識など精度の高い方法が開発され、実用化が進んできている。しかし、これらの方法は、いかにも検査されていると感じることも多く、指紋認証は犯罪捜査を連想される場合もあり、必ずしもユーザーフレンドリーなインターフェースとは言えない。そこで用途によっては、ごく自然にユーザーの認識ができる顔画像認識が適していると考えられる。しかし、離れたところから自然に検出でき、どのようなところでも使えるという画像認識の特徴を活かすためには、例えば以下のような課題をクリアしなければならない。

図5の写真は、同一日時に撮影の方向だけを変えた著者の写真である。日光のあたり方で画像が大きく変化してしまっている。この変化に画像処理は弱い。この問題を解決するために、顔モデルを用いて照明による変化を補正したり、ステレオカメラを用いて３次元認識を試みたりされているが、実用化に向けた決定的な解決方法にまでは至っていない。

図５　太陽光との位置関係による顔画像変化例
（室内でも照明条件で顔画像は大きく変化する）

5 画像認識の応用分野

ヤマハ発動機株式会社（以下、当社）の製品の中では、表面実装機（サーフェースマウンター）で画像認識技術が使われている。画像認識は、当社の実用化例をはじめ工場向けの応用が多かったが、最近、世の中の身近なところで画像認識が使われているのを見かけるようになった。例えば、高速道路や主要幹線道路に設置されている自動ナンバー読み取りシステムにも画像認識技術が応用され、高い信頼性と高速処理を両立している。また、最新の自動車のクルーズコントロールでは、画像認識とレーダー技術を組み合わせて、前方を走行する車両が車線のどこを走っているかを検出して、一定の車間距離を保つものが実用化されている。当社における研究事例としては、**図6**の後方車線検出があげられる。これは、オフライン処理で車線（白線→図中の緑線）認識を試みたものである。この技術を二輪車に応用する場合、カメラを搭載するスペースや位置の制約や、コーナリング時や車線変更時の二輪車特有のバンク角に対する対応が、四輪車の場合と比べ技術的に困難な課題となっている。

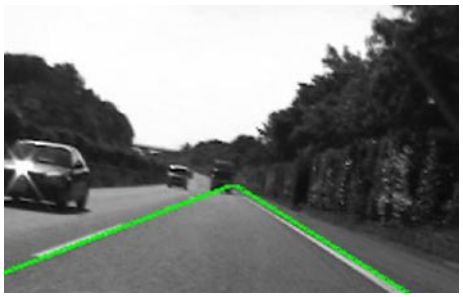

図６　後方の車線検出例

その他、一般的な研究事例として顔画像認識技術を使った入国管理システムや入出門管理システムへの応用などがみられ、より人間に近いところで、複雑な画像を対象としたものが増えてくると予想される。

6 音声と画像の組み合せ

音声と画像認識技術は、パターン認識という枠組みで似通った技術である。実際、隠れマルコフモデルや独立成分分析などの技術が音声認識、画像認識の両者で使われて研究がなされている。また、私たち人間も外界を認識する際、あるいはコミュニケーションをする場合に、一般的には音声だけでなく視覚（画像認識）も使ったマルチモーダルな情報処理をしている。例えば人と話をする際には、相手の表情、唇の動き、仕草等々、全体をみて相手の言おうとしていることを理解している。人同士が話をしていている際には、視覚情報をはじめとするその他のモダリティーの方が、音声から得られる情報よりも多いとも言われ、このことは今後の音声認識技術開発を考える上でのヒントとなっている。

7 おわりに

音声認識と画像認識の概要について説明してきたが、最後に今後の方向性について述べる。音声認識では、コールセンターの自動受付やカーナビのハンズフリーインターフェース、画像認識では、文字認識、部品認識、表面検査等で応用が広がりつつある。効率追求をねらった自動化から、さらに人間よりの技術への進化が期待されている。例えば、ドライバーの識別だけでなく、認知レベル、感情状態などを検出して、それに応じたナビゲーション、制御特性の変更などが今後実用化へ向かう可能性がある。そうなれば、音声認識と画像認識は世の中で広く使われるようになり、マシンは身近なパートナーとなるだろう。

■著者

望月 博文