■8 群（情報入出力・記憶装置と電源）－ 1 編（センシングとインタラクション）

4 章 入出力のための知的情報処理

【本章の構成】

本章では以下について解説する.

4-1 入出力のための音声・言語処理

4-2 実世界を理解するための画像認識

4-3 マルチモーダルインタラクション

4-4 ヒューマンロボットインタラクション

■8 群 － 1 編 － 4 章

4-1 入出力のための音声・言語処理

（執筆者：甲斐充彦）［2018 年 9 月 受領］

コンピュータの入出力装置としては本編 1 章で述べられているようにキーボードやマウス，ディスプレイなどが代表的である．本節では人とコンピュータとの間の入出力のユーザインタフェースとして古くから注目されている音声や言語による入出力に関わる技術に注目する．音声認識や音声合成などの音声言語処理技術や自然言語処理技術が基盤となるため [1]，それらとの関わりについて述べる．

4-1-1 音声言語メディアとインタフェースの種類

人とコンピュータとの間の入出力インタフェースは，サポートするメディアや伝達手段の違いによって**表 4・1** のように分類することができる [1]．人とコンピュータの間のユーザインタフェースとして最も一般的に用いられるキーボードやマウスなどの入力装置は，この分類において接触型に分類される．これらの装置は人がコンピュータにデータ入力する効率や正確性を向上させるために発展してきた．一方で人との対面コミュニケーションで用いられるバーバルインタフェースは，コンピュータとの間での意図伝達をより自然で円滑にするものとして期待されている．本節ではそのような入出力のユーザインタフェースの実現に関係する音声・言語処理について述べる．

表 4・1 メディアと伝達手段の違いによる人と機械のインタフェースの分類

サポートするメディアによる分類基準		インタフェースの種類
人の対面コミュニケーションに用いられるものと同じかまたは類似のもの	言語	バーバルインタフェース
	非言語	ノンバーバルインタフェース
人の対面コミュニケーションに用いられるものと異なるもの	接触型	ハプティックインタフェース
	非接触型	ノンハプティックインタフェース

4-1-2 入出力のための音声・言語処理技術の進展

コンピュータにおける入出力のための言語処理技術の進展としては，古くはキーボードを入力装置とするテキスト入力インタフェースが挙げられる．特にワードプロセッサ（ワープロ）専用機やオペレーティングシステム（OS）の開発の黎明期では，自然言語文の入力のためのユーザインタフェースの開発が進んだ．日本語の入力を例にすると，1980 年代にワードプロセッサの実用化に入ってから仮名漢字変換方式が主流となり，キーボードによる仮名入力方式と併せた日本語入力方式が活発に検討されてきた．また検索・質問応答システム，対話システム，機械翻訳システムなどの自然言語処理の応用技術が考えられるようになり，それらの要素技術となる形態素解析，構文解析，意味解析，文脈解析のモデルやそれらに関わる言語知識体系の構築が進められた [2]．

[1] 詳細は 2 群 7 編「音声認識と合成」や 2 群 10 編「自然言語処理」を参照
[2] 詳細は 2 群 10 編（自然言語処理）を参照（2018 年 8 月時点では未公開）

一方で，コンピュータの記憶容量や計算能力の向上などに伴い，1980 年代頃には大量の音声データから音声言語知識を獲得する統計的方式の音声認識技術の進展があった．1990 年代に入ってインターネットや Web の利用の広がりによって自然言語テキストを含む大規模な文書群をデータとして扱えるようになってからは，自然言語処理の分野においても音声認識技術と同様に統計的なアプローチが広く研究されるようになった．それに伴い，人手によって言語知識を構築するのではなく，辞書や文法的な知識を大規模なテキストデータから構築する統計的なアプローチによる自然言語処理のモデルが多く採用されるようになってきた．また，Web コンテンツなどを対象として自然言語文のクエリ（検索要求）入力による情報検索技術や，コンピュータから人に提示する情報を自然言語文として生成するなど，自然言語文による入出力技術の開発が進められた．

自然言語処理技術の進展に伴い，音声をテキスト情報に変換する大語彙音声認識（Large Vocabulary Continuous Speech Recognition；LVCSR）システムや，自然言語文から音声へ変換するテキスト音声合成（Text-to-speech; TTS）システムの要素技術として応用されるようになった．初期の実用化への試みとしては業務関連でのニーズに関するものがあり，医療所見の記録においてその記録を音声で入力するために大語彙音声認識（ディクテーション）システムを利用する例や，新聞などのテキスト編集時の校正の目的のために音声合成システムを利用する例などが挙げられる．個人利用への初期の応用例としては，パソコンで音声からテキストに変換して入力する用途のディクテーションシステムの商品化や，カーナビゲーションシステムの操作を音声コマンドで可能にする組み込みシステムの事例などがある．2000 年代に入ってから，パソコン用のおもなオペレーティングシステム（OS）に音声認識機能やテキスト音声合成機能が標準的に搭載され，テキスト入力だけでなくコンピュータの操作も一部可能となったほか，文書や操作内容の音声読み上げが利用可能となった．そしてこれらの仕組みは主要な OS が標準的に提供するようになったアクセシビリティ機能の重要な要素となっている．その後，スマートフォンにおいても同様な機能が提供されるようになったが，ウェブ検索をはじめとしたクラウドサービスの普及と大きく関わる．ウェブ検索ではユーザの検索クエリが大量にサービス提供側に蓄積されており，クエリを音声認識するための言語情報を容易に入手できるからである．

一方，スマートフォンが普及してきた 2010 年代には音声入出力の仕組みはウェブ検索だけに留まらず，スマートフォンが備えるメールやリマインダなど多くの基本機能との連携も図られた．ユーザが“仮想的なアシスタント”に対して自然な文体の音声で問いかけることによって，その要求に応えて音声やテキスト表示で応答するものである．同様な仕組みはその後，スマートフォンのアプリや後述のスマートスピーカ（AI スピーカ）などに搭載する機能として製品発表が相次いでおり，音声 AI や音声アシスタント機能などとも呼ばれる．

以下の節では，このように進展してきた音声・言語処理技術の中から入出力のための音声・言語処理としての応用の側面に分けて概説する．

4-1-3 入出力のための自然言語処理

人とコンピュータとの間での自然言語テキストによる入出力の用途は大きく 2 つに分けられる．一つはワープロによる文書作成のように自然言語テキストそのものをデータとして扱うもので，もう一つは自然言語をコンピュータとの意図伝達の手段として用いるものである．ここ

では，それぞれの側面での代表的な言語処理の概要について触れる [3].

(1) テキスト入力のための言語処理

前述のように，自然言語処理の応用として日本語テキスト入力への応用と実用化は早い段階になされている．そこで日本語の自然言語入力として主流となっている仮名漢字変換方式に関わる典型的な言語処理について述べる．

仮名漢字変換は，単語から文章くらいまでの単位の平仮名文字列を入力として，最も適切な仮名漢字文字列に変換する問題として定義される．そのために，あらかじめ形態素単位の仮名文字列と仮名漢字文字列との対応のリストからなる辞書を持つ．そして，与えられた平仮名文字列の部分文字列について辞書を検索し，一致する部分文字列を仮名漢字文字列に変換する．このとき，平仮名文字列から変換可能な形態素列が複数存在することがあるため，形態素の並びに対する接続可能性を考慮する．具体的な方法は，仮名漢字文字列を形態素の並びに分割して品詞や読みを同定する形態素解析と同様であり，接続可能性を表すものとして接続コストを定義する方法や，大量のテキストデータの形態素解析結果から得られる連接の統計的確率を用いる方法が代表的である．いずれの場合も，変換された文字列全体でのコスト合計や連接確率によって最適な変換を効率よく求めるため，動的計画（Dynamic Programming）法の原理に基づく最適解探索法のビタビ（Viterbi）アルゴリズムが用いられる．また，適切な変換は話題や文脈にも依存するため，コストや連接確率に文脈の影響を反映する仕組みや，直前までの仮名漢字変換の選択の履歴をもとに適応的にコストを更新する工夫が用いられる [2].

(2) 意図伝達のための言語処理

応用例としてコンピュータとの対話を目的とする対話システムや機械翻訳システム，情報検索システムなどがある．実用化の面では自然言語テキストを対象とした情報検索への応用が進んでいる．一方で，用途は限定的ながらインターネットやクラウドサービスの普及に伴い，対話システムとしての実用化も徐々に進んでいる．そこでこれらに関わる典型的な言語処理について述べる．

情報検索の問題において，大規模な自然言語テキストがファイル単位など適当な単位で分割されているものを文書と呼び，その単位でユーザが要求する内容と一致する文書を見つける問題を文書検索（document retrieval）と呼ぶ．この場合，ユーザの意図を伝える検索要求（クエリ）の入力として文書中に現れる文字列そのものが与えられるとは限らないため，検索者が情報要求内容を自然言語として与え，内容が一致する文書を見つけ出すのが一般的な問題設定となる．文書間またはクエリと文書間での内容の類似性を考えるとき，最も基本的な方法としてそれらの文書やクエリに共通して出現する単語を手がかりとする方法が用いられる．その代表的な方法としてベクトル空間法や文書中の単語の出現頻度などに基づく確率モデルによる方法 [3]などがある [4].

対話システムの大まかな分類としては，一つの自然言語文の入力に対して一つの応答を出力する一問一答型の入出力を対象とするものと，より複雑な問題解決を図るために自然言語文の

[3] 詳細は 2 群 10 編（自然言語処理）も参照のこと（2018 年 8 月時点では未公開）.
[4] 詳細は 2 群 11 編（マルチメディア），2 群 12 編（文書処理）も参照のこと.

入力を繰り返して，文脈を含めてユーザが意図していることをシステム側が理解して適切な応用を出力するものがある．狭義での対話システムは一般に後者のことを指し，テキスト入力でのチャットシステムなどが例として挙げられる．一方，一問一答型としては顧客対応サービスの負荷を軽減するための特定のアプリケーションや Web 上のサービスとしての QA（質問応答）やヘルプシステムが代表的である．

対話システムでは，自然言語文から意味表現へと変換する言語理解処理や，意味表現の結果を踏まえて意図を理解し，適切な応答を生成する対話制御処理が要素技術となる．言語理解では通常，形態素解析，構文解析，意味解析を経て何らかの意味表現を得る．文脈を考慮する対話システムでは，さらに文脈解析を経て対話履歴を考慮し意味表現としてあいまいな部分を補完する．最終的な意味表現としては動詞を中心として文中の語と語の意味関係を表す考え方に基づき，フレームという表現形式が用いられる．特に，日本語では動詞の用法の別に格助詞「が」や「を」などが取り得る名詞の意味的制約を記述した格フレーム（case frame）を用意し，文との整合性に基づいて意味解析を行うのが一例である．一問一答型の対話システムの場合，このようにして得られた意味表現をもとにデータベースの検索式（SQL クエリ）や検索コマンドへ変換して応答結果を得る．文脈を考慮した対話システムの場合は，過去の文脈情報と最新の意味表現から不足する情報を補ってユーザの意図を理解し，得られた意味表現からデータベースを検索して応答結果を得る [3]．

一方，対話制御処理では，対象とする話題や意図の範囲が限定的な場合，有限状態オートマトンによる状態表現を用いて可能な対話状態や状態遷移，それぞれの対話状態でどのような応答を返すかを記述できる．しかし，ユーザが主導権をもつ対話（ユーザ主導対話 ; user initiative dialog）やユーザとシステムの両者に主導権をもつ混合主導対話（mixed initiative dialog）などでは，対象とする話題が広くなるにつれて対話の状態遷移をすべて書き尽くしておくことは難しくなる．そのような場合，前述のように格フレームを用いることでより汎用的な対話制御が可能となる．例えば，各フレームによる言語理解の意味表現から WH 疑問文であるか，Yes/No 疑問文であるかなどがわかり，それによって次に起こす行動（応答文）を決定することで，柔軟な対話制御が可能となる．

最近では，大量の学習用データを用いることで意味解析を行うアプローチも用いられる [4]．一つの発話に対してドメイン，意図，スロットの 3 種類を埋める処理を言語理解と考える．例えば，天気ドメインの意味フレームの意図スロットとして気温，日付，場所を持つとすると，「今日の浜松の気温は」という文から，「ドメイン＝天気」「意図＝温度」「場所＝浜松」という内容を同定する．意図の推定の問題は，分類問題と考えることができるため，発話に含まれる単語列から何らかの特徴量を抽出し，機械学習の方法を適用することができる．自然言語文から抽出する特徴量としては形態素解析で得られる単語列そのもののほか，語彙の各単語の出現数（bag-of-words），単語 2-gram の出現数などが利用される．また，機械学習の方法としてはサポートベクターマシン（SVM），最大エントロピー法，ニューラルネットワークなどが用いられる．一方，スロットの充足の問題については，単語列を句の単位にまとめあげる chunking や固有表現抽出（named entity recognition）の問題として扱われる．これらの解決法としては，系列ラベル付けの問題に用いられる最大エントロピー法や CRF（Conditional Random Field）などの方法が用いられる．

4-1-4 入出力のための音声言語処理

前項と同様に，音声に含まれる言語情報（自然言語テキスト）そのものをデータとして入出力する用途と，音声を意図伝達の手段として入出力する用途に分けて述べる．

(1) 自然言語文と音声間の変換としての音声言語処理

4-1-2 項で述べたように，音声入力によってテキストへ自動変換する用途として大語彙連続音声認識システムが，その反対にテキストから音声へ変換する用途として音声合成または音声テキスト変換（TTS）システムが用いられる．後述のように，音声言語を利用した対話システムや機械翻訳などのアプリケーションの要素技術としても用いられることが多い．その基本的な仕組みを以下に述べる [5]．

コンピュータへのテキスト入力を目的とした大語彙音声認識システムでは，人間を相手としての発話と比べて予期される言語表現や発話スタイルの範囲は限定的となる．そのため，大規模な書き言葉のテキスト資料から単語 n-gram に代表される統計的な言語モデルを獲得し，書き言葉を多くの人が読み上げた音声資料から隠れマルコフモデル（Hidden Markov Model ; HMM）に代表される統計的音響モデルを構築し，それらを併用して照合する方法が基本となる．語彙の追加に柔軟に対応するために，音素単位での音響モデルを構築し，単語の音素単位での発音辞書の情報を併用して照合するのが一般的である．しかし，自然言語文を連続して発声する場合には，音素単位の区切りが明確ではなく時間的に隣接する音素の影響を受けて音響的な特徴も変動する（この現象を調音結合と呼ぶ）ため，前後の音素のコンテキストの違いを考慮した単位（典型的には前後の一音素の違いを区別する triphone 単位）で音響モデルを構築する．

同様に，テキストから音声に変換する方法としても，語彙の追加に柔軟に対応するために音素や音節単位など単語より細かな単位で音声のテンプレートや音響特徴の統計的モデルを用意し，それらを滑らかに繋ぎ合わせることで音声合成を実現する方法が代表的である．音声合成の場合はピッチや音の大きさの時間変化として表されるアクセントやイントネーションが重要な情報をもつため，形態素解析の段階で単語固有のアクセント情報を得るとともに，構文解析や係り受け解析によって適切なアクセントの区切り（アクセントフレーズ）を付与し，ピッチやパワーの時間変化を制御可能な音響モデルによって音声合成を実現する．

なお，2010 年代に入ってからは人工ニューラルネットワークモデルを中心とした数理モデルの深層学習（Deep learning ; ディープラーニング）技術の発展に伴って，音響モデルや言語モデルの一部や全体を多層の DNN（Deep Neural Network ; 人工ニューラルネットワーク）モデルに置き換えるアプローチがより高い認識精度や合成品質を与えるようになっている．音声認識システムでそのような音響モデルを実現する方法としては，隠れマルコフモデルによるサブワード単位の時系列生成確率を与える生成モデルとしての HMM を基盤として，識別モデルとしての DNN によって求まるサブワード・状態単位での音声特徴の事後確率推定値と組み合わせる方式（DNN-HMM 方式と呼ばれる）が挙げられる．従来の生成モデルに基づく方法では，音素状態のクラス別の特徴量分布を確率モデルとして表現するために事前の特徴抽出方法を工夫したり，音声特徴に影響を与える要因をモデルに組み込む工夫をしたりすることが高い認識精度を得るために不可欠であった．しかし，ニューラルネットワークでは全クラスを一つの

[5] 詳細は 2 群 7 編（音声認識と合成）を参照のこと．

識別モデルで表現し，特徴抽出の段階をニューラルネットワークの一部に暗黙的に含めることができるため，限られた量の学習用データを有効利用して，より識別性の高い音声特徴表現を事例から獲得できる利点がある．

同様に言語モデルについても，単語 n-gram 確率モデルとの併用や置き換えとしてニューラルネットワークに基づくモデルの有効性が示されている．この場合，離散的なシンボルとしての単語の情報をどのようにニューラルネットワークへの入力として与えるかが問題となる．その方法としては，語彙単語数の次元数を持つベクトルにおいて各要素を単語の種類に対応させ，入力とする単語に対応する次元のみ要素が 1，他は 0 の値をとるベクトル表現（one-hot ベクトルと呼ぶ）を元に，それをより低次元のベクトルに変換する方法が用いられる．そのような表現を単語の分散表現（distributed representation）または埋め込み（embedding）と呼ぶ．この単語の分散表現を得る方法としてはいくつか提案されており，word2vec として知られる方法がある [5]．このように求めた単語の分散表現は，意味の加減算表現ができることが事例として示されており，言語モデルだけでなく依存構造解析や文の分類などニューラルネットワークを用いた言語処理の要素技術となっている [3]．n-gram と同様に直前までの単語を入力として次単語の単語確率を予測する言語モデルをニューラルネットワークとしてモデル化する場合には，系列モデル化においてより高いモデル化能力を発揮する RNN（Recurrent Neural Network；リカレントニューラルネットワーク）やその改良型である LSTM (Long-Short Term Memory；長・短期記憶) - RNN などが用いられ，従来の n-gram を超える言語モデルの性能が示され，音声認識精度の改善に寄与することが明らかになっている．

さらに近年では，テキスト機械翻訳の分野では原言語から目標言語への変換過程を一つのモデルとして包含する End-to-end 型の人工ニューラルネットワークモデルに基づくアプローチが成功を収めており，音声からテキストまたはテキストから音声への変換モデルとしての応用例でも同等以上の性能を得る研究事例が報告されている [6, 7]．従来の音素や単語，文などの単位を明示的に仮定する階層的な知識モデル化のアプローチと比べて，単語発音辞書や形態素解析器などの中間段階での知識表現にあわせた明示的な中間処理が不要となり，学習データの準備コストを大幅に削減できる．しかし，学習データに含まれていない単語や発音を新たに追加する場合は工夫が必要であり，従来の統計モデルベースではよく用いられる適応化技術についてはまだ重要な課題となっている．

(2) 意図伝達のための音声言語処理

音声言語を入力とする音声対話システムでは，音声認識システムと前項で述べた言語理解のシステムを組み合わせるのが基本的な方法となる．しかし，音声認識システムの出力誤りに影響される問題があるため，完全に切り離して考えることはできない．そこで最近では，言語理解から対話制御までを確率的モデルとして扱う例や，機械学習の方法と組み合わせて解決する例なども用いられる．前者の例としては，不確実な環境情報（確率的に与えられる状態）から最適な制御行動（政策 ; policy）を確率モデルとして与えるものとして用いられる部分観測マルコフ決定過程（Partially observable Markov decision process: POMDP）を，音声対話システムに応用する例がある [8]．POMDP では，各種行動の結果として得る報酬（reward）を決めておくことで，最適な政策はたくさんの事例データからの強化学習（Reinforcement learning）によって求められる．音声対話システムへの応用の場合，不確実な環境情報が音声認識結果や言語理解結果

にあたり，最適な制御行動は応答の決定の問題に置き換えて適用される．

一方で，カーナビゲーションシステムのように比較的扱う対象ドメインが限られている場合には，前項で述べたように有限状態オートマトンによって対話の状態を記述し，対話制御を行うことができる．また，対話システムの記述について多用なアプリケーションへの汎用性を持たせる例として，W3Cにおいて標準化が進められた対話システム記述のためのマークアップ記述言語 VoiceXML の例がある．VoiceXML の仕様は**図 4･1** のような構成を想定しており，音声認識，DTMF（電話のトーン信号）キー入力，録音，音声合成，オーディオファイルの再生，電話転送機能などを用いて音声対話コンテンツを記述できる．**図 4･2** に簡単な対話コンテンツの記述例を示す．form 要素はユーザが音声や DTMF キー入力する複数の項目を個々に定義する field 要素を含む．各 field はフォーム解釈のアルゴリズムに従って処理されその順番は自由度があるため，有限状態オートマトンによる記述と比べてシンプルに柔軟な対話制御の流れを記述できる．

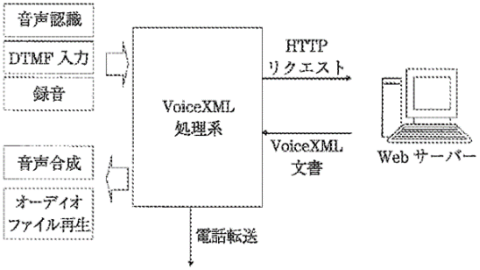


図 4・1 VoiceXML **のシステム構成** [10]

(a)　記述例（form 部分のみ）

```
<form>
  <block>
    切符購入システムです.
  </block>
  <field name="fromstation">
    <prompt>
      出発の駅名を言ってください.
    </prompt>
    <grammar src="station.grxml"
      type="application/srgs+xml"/>
  </field>
  <field name="tostation">
    <prompt>
    行き先の駅名を言ってください.
    </prompt>
    <grammar src="station.grxml"
      type="application/srgs+xml"/>
  </field>
  <field name="cnt" type="number">
      <prompt>
        切符の枚数を言ってください.
      </prompt>
    </field>
    <filled>
      <submit next="/mysite/buy"/>
    </filled>
</form>
```

(b)　対話例（S : システム，U:ユーザ）

S：切符購入システムです.
　出発の駅名を言ってください.
U：浜松駅
S：行き先の駅名を言ってください.
U：東京駅
S：切符の枚数を言ってください.
U：1 枚

図 4・2 VoiceXML **による音声インタフェースの記述例**

4-1-5 実環境利用のための音声処理

最近ではスマートフォンのように雑音の影響を受けやすい屋外での利用や，スマートスピーカ（AI スピーカ）のように生活空間内で雑音・残響の影響をうける遠隔からの音声入出力を想定するシステムが増えている．このような場合，多チャンネルのマイクロフォンアレイを利用し，音声を時間同期加算して雑音・残響を抑圧するビームフォーミングが用いることが多い．また，雑音除去オートエンコーダの応用として深層ニューラルネットワークモデルを用いて遠隔音声から接話マイクの音声の特徴に写像するように学習して利用することで，雑音・残響下の音声認識精度を改善できる [9]．

■参考文献

1) 黒川隆夫：“ノンバーバルインタフェース,”オーム社, 1994.
2) 増井俊之：“予測/例示インタフェースの研究動向,”コンピュータソフトウェア, vol.14, no.3, pp.4-19, 1997.
3) 中川聖一, 小林　聡, 峯松信明, 宇津呂武仁, 秋葉友良, 北岡教英, 山本幹雄, 甲斐充彦, 山本一公, 土屋雅稔：“音声言語処理と自然言語処理(増補),”コロナ社, 2018.
4) 颯々野学：“音声発話からの意味理解,” 電子情報通信学会誌, vol.101, no.9, pp.891-895, 2018.
5) T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean：“Distributed representation of words and phrases and their compositionality,”Proc. 26th NIPS, pp.3111-3119, 2013.
6) 河原達也:“音声認識技術の変遷と最先端－深層学習による End-to-End モデル－,”日本音響学会誌, vol.74, no.7, pp.381-386, 2018.
7) Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R.J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R.A. Saurous：“Tacotron: Towards End-to-End Speech Synthesis,”Proc. INTERSPEECH, pp.4006-4010, 2017.
8) J.D. Williams and S. Young：“Partially observable Markov decision processes for spoken dialog systems," Computer Speech & Language, pp.393-422, 2007.
9) 三村正人：“深層学習に基づくフロントエンド特徴量強調と頑健な音声認識,”日本音響学会誌, vol.73, no.1, pp.47-54, 2017.
10) 荒木雅弘：“ボイスウェブの可能性 —VoiceXML 概説—,”情報処理学会誌, vol.44, no.10, pp.1044-1051, 2003.

■8 群－1 編－4 章

4-2 実世界を理解するための画像認識

（執筆者：佐藤　敦）［2018 年 5 月 受領］

実世界にあるモノや起こっているコトを把握し，そこから有益な情報を抽出し活用することは，効率的で安全・安心かつ快適な社会を実現するうえで重要である．視覚情報は，人間が外部から受ける知覚の 8 割以上を占めると言われるように，重要かつ豊富な情報を含んでいる．この視覚情報を工学的に取り扱う画像認識技術は，人間が取り扱えない大量な画像を高速に処理することができ，目視作業の代替として有用なだけでなく，人による作業では得られなかった新たな価値の提供も可能にする．

4-2-1 画像認識の処理過程

生物は長年に亘る進化の過程を経て，強力なパターン認識能力を獲得してきた．生きるうえで必要な食糧を得て，捕食される危険を避けるには，外界を感知し即座に対応しなくてはならない．五感と呼ばれる視覚，聴覚，触覚，味覚，嗅覚のなかで，より遠方の情報が得られる視覚は特に有用である．この視覚を工学的に実現する画像認識技術は，**図 2･1** に示す処理過程から成り立っている．

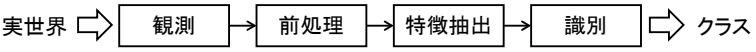


図 2・1　画像認識の処理過程

「観測」とは，実世界の認識対象を処理可能な電気信号に変換する処理で，可視光カメラや赤外線カメラ，イメージスキャナなど，認識したい対象の特性に応じた画像入力装置が用いられる．一般に，信号の空間的変化を画像と呼ぶが，更に時間的変化を含む場合は動画像，含まない場合は静止画像と呼んで区別する．「前処理」とは，後段の処理負荷を軽減するための処理で，認識対象の画像中の位置や大きさ，あるいは信号の強さを揃える正規化や，ノイズ除去などが行われる．「特徴抽出」とは，識別に有効な特徴を画像から取り出す処理で，複数の特徴値を並べた，特徴ベクトルと呼ぶ多次元ベクトルで表現することが多い．「識別」とは，特徴ベクトルがどのクラスに属するかを判定する処理で，判定したクラスを出力する．最も簡単な例では，クラスごとに代表となる特徴ベクトルを登録しておき，入力された未知の特徴ベクトルとの類似度が最大となるクラスに判定する．クラスとは認識対象が属する集合概念であり，目的に応じて人間が定める．例えば，画像中の人と車を識別したい場合は，それぞれを別のクラスとして定義し，個人を識別したい場合は，人ごとに別のクラスとして定義する．

画像認識技術の処理過程は，後段になるほど抽象度が高くなり，数理的手法との整合性が良い．例えば，識別処理の最適性はベイズ決定理論によって保証され，機械学習による識別器の自動設計が一般的となっている．1980 年代後半のニューラルネットブーム [1] を経て，サポートベクトルマシン [2] をはじめとするマージン最大化に基づく手法や，カーネル関数を用いた非線形手法の有効性が示されている．一方，処理過程の前段になるほど対象依存性が高くなり，数理的手法との整合性が悪くなる．例えば，特徴抽出は認識対象ごとにその分野の専門家が試行錯誤的にアルゴリズムを考案してきた経緯がある．しかし，深層学習 [3] の進展によって，専門

家の設計を上回る特徴抽出の自動設計が可能になっている．2010 年代にブームとなった深層学習のモデルは，基本的には 1980 年代に提案された畳み込み型ニューラルネット[4]をもとにしているが，より深い層構造の学習を可能にしたことで，識別だけでなく特徴抽出も含めた自動設計を実現した．

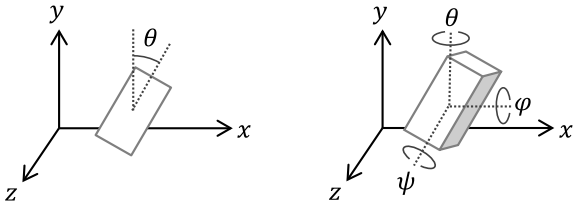
4-2-2 画像認識の難しさ

画像は，3 次元の実世界を 2 次元平面に投影しているため，その見え（アピアランス）は，認識対象そのものの変化のみならず，撮影条件によって大きく変わる．このような画像変動は認識精度の低下を招くため，実用にあたっては画像変動を如何に除去するかが重要な課題となる．認識精度が低下する要因とその対策例を**表 2･1** に示す．姿勢変化，照明変化，対象自体の形状変化のほかに，対象が置かれた背景の複雑さや，対象の一部が隠れることによる精度低下，更には撮像に用いたカメラの影響も考えられる．もちろん，姿勢や照明がそろうように管理された条件下で撮影したり，遮蔽が生じにくい位置にカメラを設置したりするなどの「観測」の工夫は重要であるが，それが難しい場合は，表に示した対策例が「前処理」として行われる．

表 2･1　認識精度が低下する要因と対策例

要　因	対策例
姿勢変化	位置正規化，回転正規化
照明変化	輝度値正規化
形状変化	対象に特化した形状正規化
複雑背景	背景差分，多重解像度画像に対する総当り探索
遮　蔽	部分認識，複数カメラの利用
カメラ	歪み補正，ぼけ除去，高解像度化，ノイズ除去

姿勢変化及び照明変化の対策は，認識する対象が平面か立体かによって難易度が異なる．姿勢変化は，位置変化と回転変化に分けることができる（**図 2･2**）．位置変化は対象によらず 3 自由度であり，画像内の位置 (x,y) と z 方向の違いによる大きさ（スケール）をそろえることが位置の正規化になる．回転変化も 3 自由度であるが，重力を利用して対象を鉛直方向にそろえることができれば，回転の自由度を減らすことができる．対象が平面の場合は，撮像系に正対させることが比較的容易であり，その場合は画像内回転のみの 1 自由度となるため，慣性モーメントなどに基づき回転をそろえることができる．



(a) 平　面　　(b) 立　体

図 2･2　姿勢（位置・回転）の自由度

照明変化については，理想的な完全拡散反射（ランバート反射）を仮定すると，観測される輝度は光源位置のみで決まるので，3 自由度となる．実際には影（キャストシャドー）や鏡面反射の影響で自由度は更に多くなり，立体形状や反射特性の情報がないと，照明をそろえることは難しい．対象が平面で完全拡散反射が仮定できる場合，光源が十分離れているときの照明変化は，平面上で一様に変化する 1 自由度とみなせるので，輝度値をそろえることは容易である．

対象自体の形状変化は，対象ごとに個別に対処する必要がある．人物の動作や手話における手の動きなどは，人体モデルや手のモデルをあてはめることで，形状の変化に追随することが行われる．手書き文字はバリエーションの多さで知られ，文字の傾きをそろえたり，ストローク間隔を均等にそろえたりする形状の正規化が有効である．

認識対象が複雑な背景を伴うと，対象領域を画像から正確に切り出すことが難しくなる．正確に切り出すためには，それが何であるかを既に知っている必要があり，「ニワトリと卵の問題」と同様のジレンマが生じる．そこで，例えば文字認識では，紙面に枠を設定しその枠内に 1 文字ずつ記入してもらったり，あるいは動画像の場合はあらかじめ背景を覚えておき，それと差分によって前景の認識対象を切り出したりする処理が行われる．しかし，コンピュータの処理能力の向上に伴い，このような処理を必要としない総当り探索が可能になってきている．すなわち，画像中のあらゆる場所において，様々な大きさで画像を切り出しながら認識する処理を繰り返すことで，対象の認識と切り出しを同時に行う．

遮蔽とは，認識したい対象が，前景にある何らかのオブジェクトに隠れてしまう状態である．対象全体でなく部分領域でも認識できるような工夫は考えられるが，認識に重要な部分が遮蔽されては，認識精度の低下は免れない．そこで，遮蔽が生じないような位置から撮影できるように複数のカメラを利用する，あるいは動画像であれば，前後のフレームから情報を得て補完するなどの工夫が行われる．

カメラに起因する認識精度の低下についても様々なものがあり，レンズによる歪み，焦点が合わないことによるぼけ，対象の動きに比べてシャッタスピードが遅いことによる動きぼけ，解像度の不足，あるいは照度不足によるノイズの重畳などが挙げられる．前処理の負担をなるべく軽減するような機種の選定や使い方が求められる．

以上述べたように，様々な要因によって認識精度が低下するため，これらの画像変動を除去する前処理は極めて重要である．そこで，このような前処理や，あるいは画像変動にロバストな特徴を，従来から専門家が対象ごとに設計してきた．これが画像認識研究の泥臭さと呼ばれる所以であるが，良い前処理や特徴が設計できれば，逆にそれが強みになると言える．

4-2-3 深層学習

一方で，これらの変動を含む画像を大量に収集し，得られた特徴ベクトルをクラスごとに識別器に登録することで，変動に対して頑健にするアプローチも考えられる．機械学習は，基本的にはこの考え方に基づいており，大量な画像を学習することで，画像変動に対して頑健な認識を可能にする．例として，2010 年から始まった大規模画像認識コンペティション ILSVRC（ImageNet Large-Scale Visual Recognition Challenge）[5] の結果を**図 2･3** に示す．画像中に含まれる一般物体 1000 クラスを識別するタスクで，縦軸は上位 5 候補までに正解が含まれなかった誤り率，横軸はコンペティションの開催年を表す．2012 年に深層学習（Deep Learning）が登場

してから大幅に識別誤りが低減され，2015 年には人間のレベルを超えたと言われている．

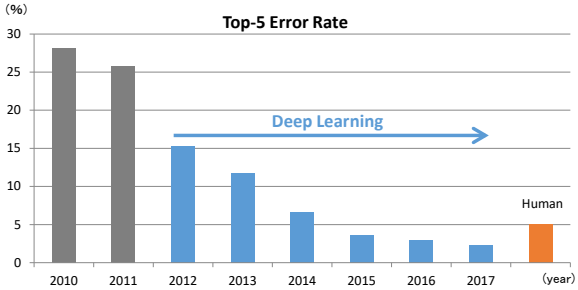


図 2・3　ILSVRC における一般物体認識の誤り率

深層学習の登場により，学習用の画像を大量に集めることができれば，特別な前処理や特徴抽出を行わずに，End-to-End で非常に高い認識精度を実現できるようになってきた．画像変動の自由度を d とし，1 自由度当たり N 枚の画像を撮影すると，合計で N^d 枚の大量な画像を学習する必要があるが，経験的には，クラス当たり 5000 枚の画像を学習すれば受け入れ可能なレベル，クラス当たり 1000 万枚の画像を学習すれば人間と同等以上の精度を達成できると言われる [6]．深層学習が生まれた背景には，深い層構造のネットワークを学習できるテクニックが開発されたほか，インターネットの普及やクラウドソーシングにより，正解クラスが付いた大量な画像を準備できるようになったこと，GPU などの高性能コンピューティングにより，膨大な試行錯誤によるネットワークのチューニングが可能になったことが挙げられる．しかし，実問題への適用を考えると，認識対象の画像が必ずしもインターネット上にあるとは限らず，画像データの収集や正解付けが高コストなことから，学習用の画像を大量に集めることが難しい場合も少なくない．このような場合は，深層学習で必ずしも高い精度を出すことができず，画像変動を抑える前処理を組み合わせるなどの工夫が必要になる．

4-2-4 実用化例

画像認識技術の初期段階では，実世界に存在する 2 次元のモノを認識対象とした．文字は紙という 2 次元平面に書かれたものであり，指紋は 3 次元物体である指を紙面に押し付けることで 2 次元化したものである．文字認識は，文書を電子化する OCR（光学式文字認識）ソフトや，郵便物を仕分ける郵便区分機などで実用化されている．指紋認証は，犯罪捜査向け自動指紋照合システムとして実用化されている．顔については，カメラに正対してもらうことで 2 次元の正面顔画像とみなし，更には影ができないように照明条件を工夫することによって 3 次元的な画像変動を除去し，実用化を進めてきた．2001 年の米国同時多発テロ事件以降，世界的に入出国管理が強化されたのを契機に，空港などで顔認証が実用化された．米国国立標準技術研究所（NIST）は，1993 年から顔認証のベンチマークテストを継続的に実施してきた（**図 2・4**）．他人を本人に間違える誤り率を 0.1 %に設定したときの，本人を他人に間違える誤り率は，1993 年に行われた FERET 1993 では 79 %だったのが，2010 年に行われた MBE 2010 では 0.3 %にまで大幅に低減していることが分かる [7]．指紋認証や顔認証は，身体の特徴を用いて本人を確認

する生体認証（バイオメトリクス認証）であるが，単なる本人確認にとどまらず，大量なデータベースとの高速照合によって，人手では難しかったブラックリストの即時照合を可能にしている．

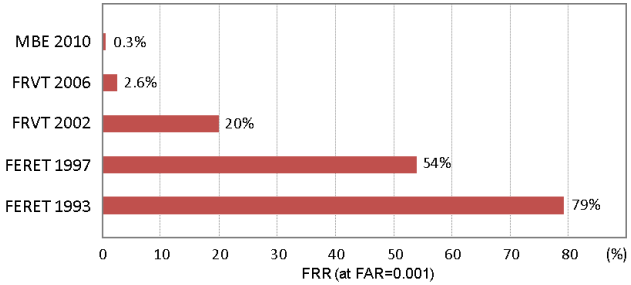


図 2・4　NIST ベンチマークにおける顔認証の誤り率

これまで目視によって欠陥や異常を見つけていた作業を，画像認識で置き換える例は多い．ファクトリーオートメーション（FA）向けではウエハやフォトマスクなどの欠陥検査，医用向けではマンモグラフィなどの X 線写真のスクリーニング検査などが実用化されている．欠陥や異常の発生頻度は低いため，大量データを必要とする機械学習アプローチでなく，正常からの外れ値として欠陥や異常を検知する手法が多い．

3 次元物体の認識に関しては，ディジタルカメラ向け顔検出が実用化されている．インターネットの普及により顔画像は大量に収集することが容易であり，AdaBoost などの機械学習が採用されている．一般物体の認識は，深層学習が標準的な手法となっている．これを用いた人物や車両の検出は高い精度を実現しており，自動運転として実用化される日も近い．

4-2-5 今後の展望

深層学習の登場により，画像認識の研究は大きく様変わりしている．オープンソース化によって，世界中の誰もが最新手法を入手でき，インターネット上の画像データを用いて評価も容易に行える．このような状況は研究の活性化に貢献しているものの，実用化にあたっては認識対象となる画像が必ずしもインターネットで得られるわけではないため，その場合は自前で画像データベースを構築する必要がある．目的に応じた独自の画像データベースを構築することは大きな強みになる反面，深層学習では大量の正解付き画像が必要なため，高コストとなる問題がある．更に，欠陥や異常など発生頻度が低い事象については，そもそも大量な画像を得ること自体が難しい．このようなデータ不足を補うために，転移学習の研究が進められている．転移学習とは，認識対象ではない別のデータを流用して学習することを意味し，別のデータを変換して流用する方法や，別のデータを学習したモデルを流用する方法などがある．このような技術開発により，不十分なデータ量であっても高い認識精度を実現することができれば，実世界を理解するための画像認識技術の実用化は，更に広がるものと期待される．

■参考文献

1)　D.E. Rumelhart, G.E. Hinton, and R.J. Williams：“Learning Internal Representations by Error Propagation,” D.E. Rumelhart, J.L. McClelland, and the PDP research group (editors), Parallel distributed processing: Explorations

in the microstructure of cognition, vol.1: Foundations, MIT Press, 1986.
2) C. Cortes and V. Vapnik：“Support-vector Networks,” Machine Learning, vol.20, no.3, pp.273-297, 1995.
3) A. Krizhevsky, I. Sutskever, and G.E. Hinton：“ImageNet Classification with Deep Convolutional Neural Networks,” Advances in Neural Information Processing Systems, vol.25, pp.1097-1105, 2012.
4) Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel：“Backpropagation Applied to Handwritten Zip Code Recognition,” Neural Computation, vol.1, no.4, pp.541-551, 1989.
5) O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei：“ImageNet Large Scale Visual Recognition Challenge,” Int'l J. of Computer Vision, vol.115, no.3, pp.211-252, 2015.
6) I. Goodfellow, Y. Benjio, and A. Courville：“Deep Learning,” MIT Press, 2016.
7) P.J. Grother, G.W. Quinn, and P.J. Phillips：“Multiple-Biometric Evaluation (MBE) ―Report on the Evaluation of 2D Still-Image Face Recognition Algorithms,” NIST Interagency Report 7709, National Institute of Standards and Technology, Aug. 24, 2011.

■8 群－1 編－4 章

4-3 マルチモーダルインタラクション

■8 群－1 編－4 章

4-4 ヒューマンロボットインタラクション

（執筆者：加藤由花）［2017 年 4 月 受領］

ロボットという言葉の定義は明確には定まっておらず，その用途，形態，機能は様々である．当然，ヒューマンロボットインタラクション（HRI：Human Robot Interaction）にも様々な形態が存在する．近年では，人にサービスを提供する知的なエージェントとしてロボットを捉え，このようなロボットと人とのインタラクションとして HRI を捉えることが多い．しかし，本来，HRI はより広い概念を含む言葉である．例えば，パワーアシストロボットと人とのインタラクションなども対象になる．本節では，まず，インタラクションの特性に従って HRI を分類し，形態ごとに具体例をまとめる．その後，知的情報処理という観点から，バーチャルリアリティ（VR：Virtual Reality），ユビキタスコンピューティング，エージェンシーと HRI の関係を説明する．

4-4-1 インタラクションの特性による HRI の分類

文献 1）では，HRI の形態を，時間的特性，空間的特性，主体的特性の 3 つの軸により分類・整理している．

- **時間的特性**：通信のリアルタイム性による分類．人とロボットがリアルタイムに通信を行う同期型と，タスクの依頼を受けたロボットが非同期処理を行う非同期型がある．
- **空間的特性**：人とロボットの物理的距離による分類．遠隔地にあるロボットを操作する遠隔型と，眼前にあるロボットを操作する臨場型がある．
- **主体的特性**：ロボットの身体の捉え方による分類．人がマスタとなりスレーブであるロボットを操作する一体型と，人と人とのインタラクションと同様，ロボットを自身と異なる個体として捉える対面型がある．

同期型/臨場型/一体型に近づくほど通信が密になり，広帯域高信頼の通信が必要になる．一方，非同期型/遠隔型/対面型に近づくほど，ロボットには自律性と信頼性が要求される．

以上の分類による HRI の具体例を**表 4･1** に示す．

表 4・1　インタラクションのタイプによるロボットの分類（文献 2）表 7.1 を一部修正）

空間的特性	時間的特性	主体的特性	
		一体型	対面型
臨場型	同期型	外骨格型ロボット パワーアシストロボット	産業用ロボット（教示型） サービスロボット（インタラクティブ型） 介護ロボット ペットロボット ユビキタスロボット
	非同期型	なし	産業用ロボット（プレイバック型） サービスロボット（バッチ型） 搬送ロボット
遠隔型	同期型	遠隔サービスロボット 遠隔医療ロボット	なし
	非同期型	宇宙ロボット	なし

(1) 一体型/臨場型/同期型

人が，自分の手足の代わりとしてロボットを操作するタイプのインタフェースである．人の随意運動指令を筋電位などにより取得し，外骨格型ロボットを自分の身体の一部のように操作するシステムや，パワーアシストロボット[3]などがある．物理的な HRI と考えられ，人の特性を考慮し，機械が人に合わせて振る舞いを適応的に変える機能が必要となる．

(2) 一体型/遠隔型/同期型

遠隔地のロボットを，ネットワークを介して操作するタイプのインタフェースである．遠隔医療ロボットのように具体的なタスクを遂行するシステムのほか，ロボットそのものをコミュニケーションインタフェースとすることで，人と人の遠隔地コミュニケーションを実現するロボット（テレプレゼンスロボット）などがある．

(3) 一体型/遠隔型/非同期型

通信遅延を回避できない環境で，遠隔に存在するロボットにタスクを遂行させるタイプのインタフェースである．惑星探査ロボットなどがある．予測ディスプレイを用いた作業教示機能，VR モデル介在型遠隔操作システムによる時間遅れ吸収機能などが用いられる．

(4) 対面型/臨場型/同期型

人とロボットが共存し，対面してインタラクションを行うタイプのインタフェースである．知的エージェントとして人にサービスを提供するロボットのほか，環境に埋め込まれたセンサ，アクチュエータ群を統合し，人の活動を支援する形態のシステムもこのタイプに分類される．後者はユビキタスロボティクスとも呼ばれ，知能化された環境全体が環境型ロボットシステムと捉えられる．

(5) 対面型/臨場型/非同期型

人が，対面したロボットに対してタスクを依頼するが，その遂行に時間を要するタイプのインタフェースである．倉庫内の荷物を運搬する搬送ロボットや，家庭用掃除ロボットなどがある．

4-4-2 バーチャルリアリティと HRI

一体型/遠隔型インタフェースでは，遠隔地の実世界とのインタラクションをスムーズに行うために，VR 技術を用いることが多い．同期型インタフェースにおいては，空間共有システムとして，非同期型インタフェースにおいては，時間遅れ吸収機能として VR 技術が利用される．

(1) 空間共有システム

遠隔地に存在する人と人の間でロボットを介したコミュニケーションを行う場合，VR 技術を利用した空間共有システムにより，高臨場感通信が可能になる．これには，3 次元空間の共有を目的とした没入型ディスプレイを用いるシステムや，高い臨場感を持ってロボットに乗り移った感覚でロボットを遠隔操作するテレイグジスタンス[4]などがある．

(2) VR モデル介在型遠隔操作システム

時間遅れのある環境で遠隔地に存在するロボットを操作する場合，人とロボットの間に VR モデル（シミュレーション環境）を介在させることで，時間遅れへの対応が可能になる．ここでは，あらかじめロボットの存在する遠隔地の作業環境を VR モデルとしてモデル化しておき，人は，実際の作業環境から遅れて伝わるセンシング結果の代わりに，VR モデルをとおして得

られる時間遅れのないセンシング結果を用いてロボットを操作する．時間遅れは，実ロボットと VR モデルの間の非同期通信により吸収する．

このような仮想予測環境モデリングに関する研究は，宇宙遠隔操作ロボットにおいて古くから行われており [5] 大きな遅延を持った遠隔操作ロボットへの仮想現実技術利用の代表的な例として，火星探索ロボット [6] がある．

4-4-3 ユビキタスコンピューティングと HRI

臨場型/対面型/同期型インタフェースの一形態として，ユビキタスコンピューティングによる空間のロボット化がある．ユビキタスコンピューティングは，1991 年に Mark Weiser により提唱された概念 [7] であるが，クラウド環境やモバイルネットワークの普及，センシング技術の高度化により，近年では，Internet of Things（IoT）と呼ばれる形態に進化している．ここでは，知能化された空間や分散したロボット全体の複合体と人がインタラクションするため，この複合体と人とのコミュニケーション機構が重要になる．ビジョンやマルチモーダル情報を用いたインタフェースや，人の意図を汲み取るためのモデリング手法，近年では，機械学習アルゴリズムや人工知能技術なども活用されている．

部屋の中にセンサやアクチュエータを分散配置し，人の行動を監視・支援するという考えに基づき空間を知能化する試みはこれまでも多く行われており，代表的なものに，MIT メディアラボの SmartRoom[8]，ジョージア工科大学の Aware Home[9]， Micorosoft 社の EasyLiving[10]，東京大学のロボティクルーム [11] などがある．近年では，スマートシティを構成する社会インフラとネットワークロボットの連携として，街まるごとロボット化 [12] なども提案されている．

4-4-4 エージェンシーと HRI

臨場型/対面型/同期型インタフェースのもう一つの形態として，ロボットを人という外界とのインタラクションを持つ自律システム（エージェント）とみなし，そのエージェントとのインタラクションを考えるものがある．これは，人間が認知するシステムのインタフェースとしてロボットを捉えた，ヒューマンロボットコンピュータインタラクションという概念 [13] に基づくものである．実際，人工知能技術の進展により，人に能動的に働きかけ，自律的に動作する存在としてのロボットが増えている．特に，人と知的に関わる対話型ロボットが多く登場しており，人型の Pepper（http://www.softbank.jp/robot/）や RoBoHoN（https://robohon.com/），据え置き型の Jibo（https://www.jibo.com/），家庭用小型ロボットの Kuri（https://www.heykuri.com/）など様々なタイプのロボットが存在する．

これらのロボットは，人とロボットの間で情報のやり取りを行う新たなメディアと考えられる．ここでは，人の状況を認識し，人が操作しなくても端末の方から働きかけるべきタイミングと内容を判断し，働きかけが可能なインタフェースが要求される．クラウド環境との連携を前提に，人に働きかけるための様々な情報をネットワーク経由で取得することで，ロボット単体ではなし得ない高度なサービスを実現している．

4-4-5 これからの HRI

今後，超高齢化社会の到来，独居人口の急増などが予想される．そのような社会では，人の

役に立つことはもちろんのこと，人を心地よくさせ，人を楽しませる存在としてのロボットに対する期待が高まる．HRI 技術においては，エージェントとしてのロボットとのインタラクションを高度化することが，ますます重要になってくる．

このように，ロボットは人々の日常生活にますます深く入り込んでいく．そこでは，倫理的・法的・社会的問題（ELSI：Ethical, Legal and Social Issues）への配慮が不可欠である [14]．社会実装を通じて，ユーザの受容性，ロボットの信頼性・安全性などに関する HRI 技術が進展し，ロボットサービスが普及していくことを期待したい．

■参考文献

1) 中内　靖, 安西祐一郎：“ヒューマン・群知能ロボット・インタフェースシステム―人間とロボットの協調について―,”計測と制御, vol.31, no.11, pp.1167-1172, 1992.
2) 日本ロボット学会編：“新版ロボット工学ハンドブック,”コロナ社, p.733, 2005.
3) Y. Sankai：“HAL: Hybrid Assistive Limb Based on Cybernics,”Robotics Research, The 13th International Symposium ISRR, pp.25-34, 2010.
4) S. Tachi：“Telexistence 2nd Edition,”World Scientific, 2015.
5) W.S. Kim and A.K. Bejczy :“Graphics Displays for Operator Aid in Telemanipulation,”Proc. of IEEE International Conference on Robotics and Automation, pp.1059-1067, 1991.
6) L. Edwards, M. Sims, C. Kunz, et al.：“Photo-Realistic Terrain Modeling and Visualization for Mars Exploration Rover Science Operations,”Proc. of IEEE International Conference on Systems, Man and Cybernetics, pp.1389-1395, 2005.
7) M. Weiser：“The Computer for the 21st Century,”Scientific American, pp.94-104, 1991.
8) A. Pentland：“Machine Understanding of Human Action,”Technical Report 350, MIT Media Lab., 1995.
9) G. D. Abowd, et al.：“Living Laboratories: The Future Computing Environments Group at the Georgia Institute of Technology,”Extended Abstracts of the ACM Conf. on Human Factors in Computing Systems, pp.215-216, 2000.
10) J. Krumm et al.：“Multi-Camera Multi-Person Tracking for EasyLiving,”Proc. of 3rd IEEE Int. Workshop on Visual Surveillance, pp.3-10, 2000.
11) 佐藤和正：“ロボティックルームの知能―ユービキタス知能,”日本ロボット学会誌, vol.20, no.5, pp.482-486, 2002.
12) K. Kamei, S. Nishio, N. Hagita, and M. Sato：“Cloud Networked Robotics,”IEEE Network, vol.26, no.3, pp.28-34, 2012.
13) Y. Anzai :“Human-Robot-Computer Interaction: A New Paradigm of Research in Robotics,”Advanced Robotics, vol.8, no.4, pp.357-369, 1993.
14) 土井美和子, 小林正啓, 萩田紀博：“ユビキタス技術ネットワークロボット―技術と法的問題,”オーム社, 2007.