2005－CVIM－148（14）
2005／3／4

動作と物体の統合的認識とそのモデル化

北橋忠宏† 樋口未来†† 小島篤博‡ 福永邦雄††

† 関西学院大学理工学部 E-mail:kt@ksc.kwansei.ac.
†† 大阪府立大学大学院工学研究科 E-mail:{higuchi@com., fukunaga@}cs.osakafu-u.ac.jp
‡ 大阪府立大学総合情報センター E-mail:ark@center.osakafu-u.ac.jp

人間は主に視覚を通じて外界の事物を認識していると言われているが、自らの行為・行動によって実体を伴った認識ができることは日常的な体験に照らしても確かめられる。本稿では、この認識に基づき、人間の行為・行動を事物への働きかけとして認識し、行為・行動がもたらす物体の変化から、行為の内容とともに物体の機能属性を求めることができることを示し、コンピュータビジョンに応用しようとするものである。

Cooperative Recognition of Human Movements and Objects and Its Modeling

Tadahiro Kitahashi †　Mirai Higuchi ††　Atsuhiro Kojima ‡　Kunio Fukunaga ††

† School of Science and Technology, Kwansei Gakuin University
†† Graduate School of Engineering, Osaka Prefecture University
‡ Library and Science Information Center, Osaka Prefecture University

In current computer vision systems, a person, if acting in a scene, is often treated as one of unsteady and complex shaped objects. However, humans' movements always put some effect onto the objects. Accordingly, it must be more reasonable to recognize humans and their movements together with the objects that they are manipulating. In this paper, we try to formulate a scheme along the notion.

１．はじめに

コンピュータの計算性能・記憶容量の急速な向上に支えられて、因子分解法や固有空間法、部分空間法など計算量の大きな特徴袖出法が実用に供されるようになり、風景画像からの３次元情報の抽出や顔画像による個人識別や表情認識に有効な手段として盛んに用いられ、認識能力や認識対象領域が広がりを見せている。また分散視覚や全方位カメラなど画像取得手段の多様化も別の角度から CV の対象領域の拡大を支援し、その領域は移動する人物を含む環境にまで及んでいる。

このように拡大を続ける CV に求められる認識結果は、いわゆるパターン認識の目的である物体名による対象物へのラベル付けの他に、ロボットビジョンとして利用される際には、対象物の３次元空間内での大きさ・位置・姿勢といった単なる名称以外の情報の抽出が求められる。近年の画像・映像の各種の映像メディアへの応用あるいは表情の認識を含む感性情報処理においては、パターン認識やロボットビジョンとは異なり、「形容詞」で表現されるような対象物の内部状態というべきものを認識することがしばしば求められる状況にある。

本稿での考察内容は、人間の行為・行動を事物への働きかけとして視覚情報から認識し、行為・行動が物体にもたらす変化から、行為の内容とともに物体の機能属性を求めるものであり、単なる対象物への呼称の対応付けではない点で、上記の分野と共通性をもつが、人間と事物という二者間の相互関係を捉えているという点でこれまでにない側面をもつ認識と考えられる。

筆者の一部は、室内環境の認識として、視覚情報に基づく物体認識の際、従来からの認識手法である物体の幾何属性として大きさ・高さ・面の向きなどの抽出に加え、人間の行動（例えばものを置くという動作）の認識結果に基づき、「ものが置かれた空間にはものを支えるといった機能・効用が備わっている」という知識を参考に、事物の機能属性を知ることにより、環境に関する認識（歩行による移動可能性を見出すことにより、経路に当たる空間の無障害物性、壁面部では出入り口の認識など）の領域の拡大や認識確度の向上を目指す新たな一手法を報告している[1, 2]。

これは外観情報から直ちに対象に概念ラベルを対応付けという従来の認識方式の枠組みとは異なり、一旦、機能属性の認識という迂回あるいは並行処理を経て対象物の認識に至るという方式の提案であると解釈できる。同時に、対象への人間の行為・行動による働きかけを暗黙のうちに前提としており、従来のCV、PRよりも一層実際的な環境を想定した画像処理であるとも解釈でき、一定の評価が与えられようとしている[3]。

また筆者の別の一部は、これと類似しながら、機能属性を主体にして、着座状態にある人間の行う動作における、物体の移動と状態の変化を抽出することにより、食べ物と摂食行動とを相互補完的に認識できる枠組みを提案している[4,5]。本稿では共通性と隔たりのある２つの研究をやや詳細に検討し、両者に共通する一つのモデルを形作ろうと試みた。

２．手法の位置づけおよび前提条件

上記の主張を一層具体的に検討し問題点を洗い出すとともに、提案の定式化に基づき人間の行為・行動とそれに関わる物体との同時的な抽出により、行為・行動と物体の機能属性を認識するシステムの作成に道筋を与えることを試みる。

２．１ 外観的認識から実体的認識へ

本稿は、最終的には行為・行動とそれに関わる物体を認識する新たな枠組みとして、対象の外観情報に基づく認識に加え、一種の実体認識を導入することを提案するものである。少数の例外を除いて[6]、これまでの CV では人間の行為・行動とそれに関与する対象物とが共存する世界を認識・理解する場合であっても、人間が物体と相違するのは複雑な形状と動きをする点のみであると捉えてきた。これに対し、本報告では行為・行動はほとんどすべて物を操作する過程であるとして、両者は相互関係をもつ協調的対象として捉えようとするものである。歩行のように物体と関わりをもたないように見える場合でも、人体を取り巻く空間に対し、歩行という形態で通過するという操作を加えていると考えられる。いわばこれまでの認識とは異なり、多くのものを人間の行為・行動の対象となって位置・形状などが変化するものとして取り扱うという、動的な画像認識・理解であり、さらに積極的にいえば、事物はこのように行動によってどのように変化するかということを捉えてこそ、はじめてその実体を認識できるとまで意識するものである。

前章に述べたように、形状による認識とは、認識しようとする属性を異にするが、実体認識もまた、画像処理に基づく対象認識の一つの必然的な発展方向である。実際、視覚・映像メディアと関連する場合には、物体の光学的特性を抽出しようとする処理になり、周知の通り現在、研究活動が活発な分野である。

もう一つロボット工学の見地からは、材質・重量・重心などの認識が必要不可欠であり、物体の物理特性を抽出する研究はこの分野におけるパターン認識に続く発展の道筋に沿うものであり、触覚センサ等に基づく対象認識は正にその方向に則った技術である[7]。

先にも触れたように、近時の画像に基づく認識では、人間が行動する空間としての環境を認識の対象としながら、環境を構成する人間と物体とをそれぞれ相互関係をもたない別個のものと仮定しているかのような認識が支

配的であると考えられる。これに対し前記のように、われわれの観点は、極端に言えば人間が対象に働きかけることこそが、実際的環境での主要な現象であり、環境認識はその働きかけの観察に基づいて、対象とその働きかけの行為・行動を関連付けて認識すべきであり、そのことが両者の認識に有効であると主張・提案している。

２．２ 対象の限定

このような認識の頭初の段階として、考察の範囲を限定するために行為・行動に制約を加える。動作に関わる対象が、位置変化、生成・消滅（場合によっては、量的変化も含める）を受ける行為・行動に焦点を当てた。このような限定は画像処理技術に係わる有効な制約になる。というのは、少し限定を強めて、行動の過程での量的な変化は無視し、その結果としての位置移動・対象の生成・消滅のみに着目するならば、時間的経過での対象物領域の変化として現れるため、基本的には差画像処理という単純な処理が基本となるからである。

２．３ 動作と物体操作の基本構造

筆者らは当初、手法も目的も見かけ上はそれぞれの特徴を備えた独自の研究を進めていた。先ず動作の種類に大きな違いがあった。一方は部屋全体を視野に収め、歩行に伴う人体の位置移動と物体の搬入・設置を中心とする状況が認識対象世界であった（図１）。従って頭部の左右移動を主体として、上下動を従にした人物の室内での位置移動を捉え、これに手の動きと対象物の位置移動を付加した認識を手掛けていた。他方は着座した人物の動作が対象であり、人物は定位置にほぼ不動で、手とそれに伴う対象物の動きが認識の鍵となった（図２）。

両者は一見すると異なるが、単に視野が異なるだけの一連のアプローチであると捉えることもできる。例えば、後者は、前者の環境で人物が大きな移動を終えて着座状態に入っ

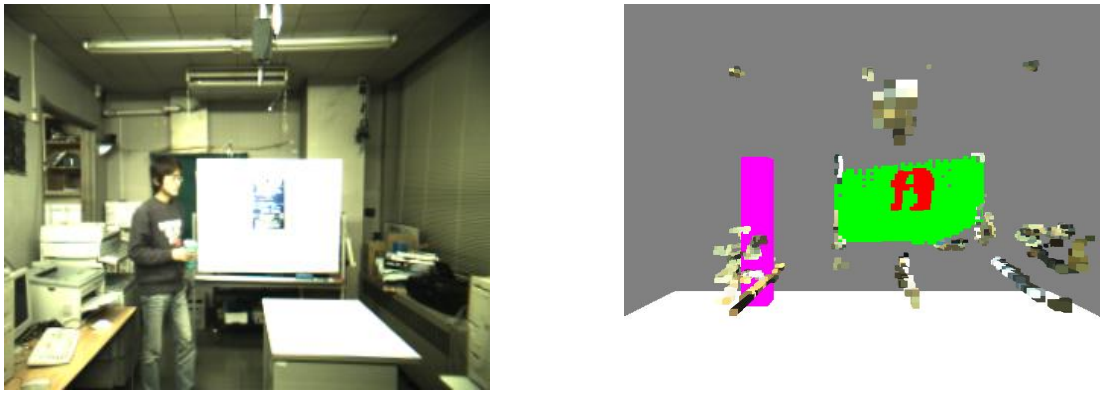


図１．室内情景と人物行動の抽出による物体の機能属性の認識

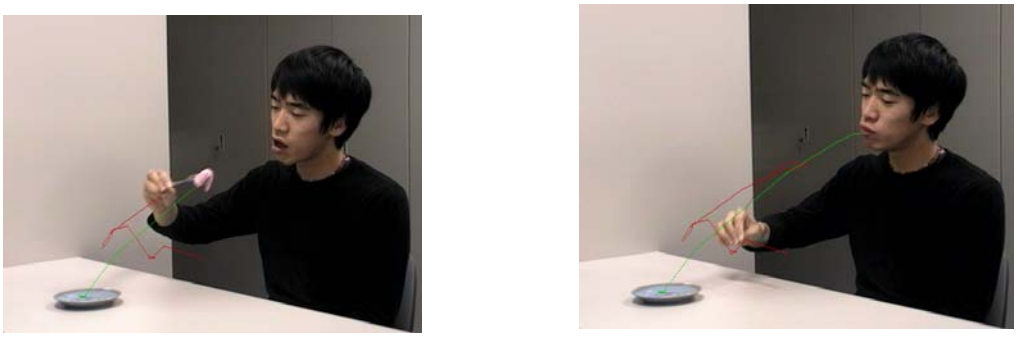


図２．着座人物の摂食行動における手と食べ物の移動軌跡

たときに、人物をクローズアップし腕や手さらには指に観察領域を絞り込んだと捉えることができる。

実際、人物の行動認識において、必要に応じて注視点を制御できるようにカメラパラメータを切り替えられる観測機能をもつシステムを作成し、画像獲得を目指す構想を現在も捨て切れずにいる。しかし観測機器の性能向上により室内全景データから手の操作を認識するに必要な精度の観測が可能になることは、つぎの発表[7]の解析で明らかにする。

２．４ 動作の形態と身体構造

もう一つ別の統一解釈の枠組みも考えられる。人体の構造に基づく動作形態の制約という考え方の導入である。人体の可動部分として下肢、上肢、手・指はそれぞれ独自の機能をもち、人間は、状況に応じてそれらを組み合わせ、必要な行動を実現していると考えられ、このような認識に基づく定式化である。すなわち、下肢が全身を行為・行動の目標が存在する位置の近くに移動し、次いで上肢が手・指をさらに行為・行動対象に近付け、最終的に正確な操作位置に導かれた手・指が対象に意図された作業を実行する。このような観点から、筆者らのこれまでの２種類の実験は、それぞれ下肢による移動と、上肢・手による移動・操作とを別々に実行したものと解釈することができる。

当初の研究には、その他に認識の枠組みにも差異が見て取れる。前者は空間および対象物の３次元配置や３次元形状を計測した上で、人間の動作から導かれる機能属性を付加することによって、対象物の同定の精度向上を図っていた（図１）。これに対し、後者の研究は手の動きとそれに伴って移動する対象物が消滅することや消滅の場所からその機能を推論し、この結果から得られる機能属性から対象物を認識し、同時に動作自体も認識できることを示し、知識偏重の機構になっている（図２）。現在、後者の画像処理性能を前者のシステムの性能向上により補完しようとしている。

３．機能属性の認識

本稿で認識しようといている「機能」は、物と物との間における関係として現れるものであり、そのものが何らかの形をもつ単体として視覚によって捉えられるものではないことは自明である。このため、その認識には関連性をもつ２つ以上の事物が係わる状況を観察し、その結果について知識に基づく推論によって決定することが不可欠になる。したがって、従来の視覚属性に基づく特徴による識別を基礎とする物体認識以上に、知識情報処理の分野との強い連携が必要とされるものと考えられる。この点でも従来の認識とは異なる要素を含んでいると言うこともできよう。

３．１ 仮定の効用

機能が現れる状況を「２つ以上の事物が係わる状況」と規定するだけでは、ほとんど無規定に等しい。そこで２．２に述べたように仮定を設ける。行為・行動によって対象物に位置移動、出現・消滅（場合によっては、見掛けの面積の増加・減少）という変化が生じる場合のみを注目する。

この単純な仮定が有効に働く局面が存在する。先ず、これらの状況は系列画像の差画像という簡単な処理により判定可能であると予測できる。次いで、それにも増して有効と考えられるのは、このように限定された状況に係わる機能のみを取り扱うため、必要とされる常識を大幅に限定できると考えられる。現在のところ、行為の結果が離れた場所に現れるようなものは除外され、

身近に結果・効果が現れる行為・行動に対象が限定される。引き続く発表[7]では、手持ち荷物の設置、ボードへの書類の貼り付け、板書を順次行う過程を机上面、ホワイトボード垂直面、板書過程との相互関係によって認識する機構を紹介する。

３．２ 具体的事例と認識機構の問題点

上記のままではなお抽象的である。具体的な事例を紹介しよう。すべての事物は重力場中に存在している。したがって、水平で凹凸の少ない丈夫な面は事物を重力に抗して支えることができる。これに対し、垂直に近い面をもつ面には通常はものを支える機能はない。

しかし面の素材を適当に選ぶと、例えば磁石やピンなどである程度の制約を満たすものを止めることができる。

前者では上向きの面をもつ一定の大きさを持つ空間に人間が近づき手に付随していた事物を手から離なし、その位置で面上に事物が移り、静止したとすれば、その状況は人がテーブルや机とか、棚とかに近づき、手にしたものをそこに置いたことを表していることは明らかである。ただし、このような手法で認識できるのは、水平面がものを支えたという事実、水平面がその機能属性をもつというのみである。その面がテーブルや机なのか、棚なのかは従来の物体認識において用いられていた高さ、形状などの幾何学的・外観的情報が必要になる。

これらの行為の結果は、人間による物体の搬送という動的な状況から物体の据え置きという静的な状況への変化をもたらすものであるため、行為が遂行された前後の変化は画像の差を求めるという単純な画像処理操作によって認識できる。

画像中の行為・対象の認識と解釈を導くキーは、先にも述べたように差画像により抽出はできるが、このようなシステムでは、知識利用が不可欠となり、認識対象となるシーンがシステムに付与される辞書的知識および推論知識の内容に大きく依存するという一般的な問題が避けられない。これまでの報告はいずれもこの点を克服できていない。

引き続き発表する報告では、画像および３次元センサー情報を基本とし、行動と対象との依存関係を重視したシステムを紹介する。

４．認識機構

これまでに観察・認識の対象としたのは数事例であり、動作の内容はいずれも実験的な単純なものである。その実験的な解析結果では、共通する特徴が明らかになった。しかし歩きながら電話をするなどの行為は、身体各部位の独立性の現れであり、逐次性の仮定には留保が必要になるが、ここではこれらを除外した行為・行動を対象として、その認識のとの関係について以下で考察する。

４．１ 日常的動作と身体構造

数例の実験的な行為・行動の映像では、その構成は階層性とそれに従う逐次性を呈した。ある面上に物を置く場合、まず歩行による全身的な位置移動の結果、目的近辺に到着し歩行を止める。その後、腕を動かして手にしたものを置く所定の位置まで面の上部空間を移動して次第に腕を下げ、物体が面に接したところで手の動きを止め、指を離して手にしたものを面上に置いている。

よって下肢による大きな移動が滞留状態に入ると、ときには下肢の屈伸による身体の上下動に移り、これが停留状態に入ると、上肢による手の移動が生じ、その滞留状態で最終的に手・指が対象に直接動作系列が見受けられる。

これは日常的に頻繁に生じる行為・行動に共通する身体の動きであり、その特徴は、行為・行動を構成する動作が移動単位の大きさによる階層性をもち、その順位に従う動作が逐次的に組み立てられているものと見なし得る行為の一つであり、実験的事例における特例的なものではないと考える。

以上の動作と身体構造との関連を容認すると、行為の解析に当たっては、一般的にも動作解析で用いられてきた特徴抽出の合理性が裏付けられる。すなわち、全身的な位置変化を頭部の移動によって代表し、手の位置変化によって上肢・手の動きを認識することである。このとき、当然ながら上肢・手の動きは正確には肩を支点とする身体の内部運動である。このため、近似的に頭部との相対的位置によって記述することが望ましい。

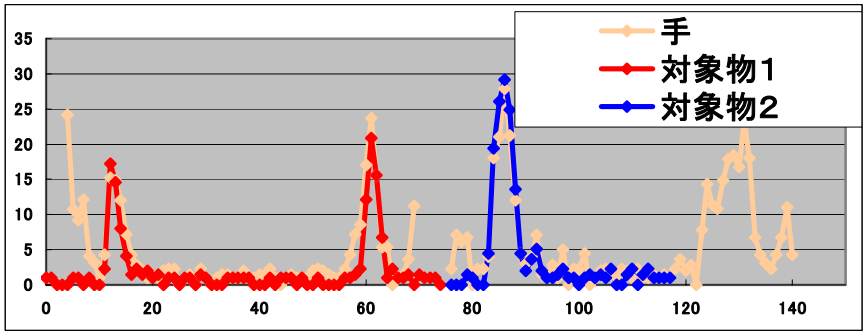
４.２ 身体部位動作と動き認識

上記のように人体の対物的な行為・行動、引いては環境への働きかけに関して、その過程に係わる身体各部位の階層性と機能分担、逐次的な動作が一般的に用いられている。したがって逐次性を前提とすることは、処理の対象となる行動にさほど大きな制約を加えるものではないと考えられる。認識に当たっては、階層性に基づいて粗から密あるいは大局から局所への注視点の移動が必要になり、これを制御する手掛かりを画像情報あるいはそれに基づく情報から導く必要が生じる。

(a) その映像断片

(b) 手とジュース領域の移動軌跡

図３．ジュースの摂取



(a)　手とジュース領域の移動量の変化
(b)　ボトルとコップ内のジュース量の変化

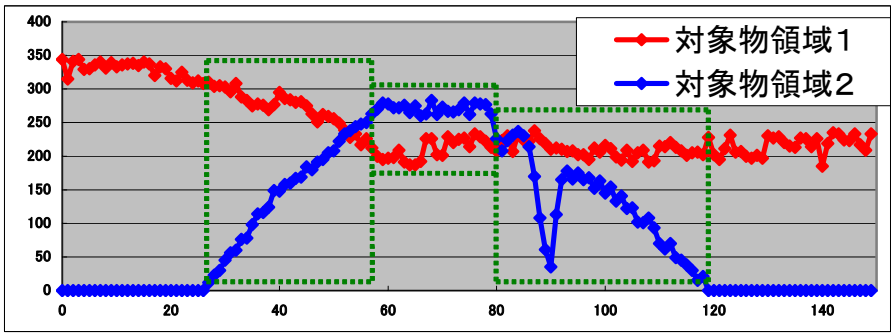


図４．ジュース摂取映像における各種領域の位置と面積の変化

動作の階層的逐次性を仮定あるいは是認するならば、実時間性を保証するには課題があるが、検出容易な単純な手立てが存在する。すなわち、上記のように上位から下位の動作への切り替えは上位の動作の滞留によって示されるため、頭部に次いで手（頭部中心座標系での）の位置変化が切り替えの時点で急速に低下する。したがって、これを検出すればカメラパラメータの切り替え、注視点の切り替えの制御が可能になると考えられる。切り替えのタイミングの検出が重要な鍵になるが、これを実験から求めた。

図３(a) に示す行動は、図２の行動と同様の状況を設定しながら、対象物をジュースに変更し、その摂取過程を撮した映像の一部であり、図３(b)は、手とジュース領域の重心の軌跡を示している。ボトルからコップにジュースを移して飲むことにしたため前回の動作よりも操作が増加している。図 4(a)では肌色の線が手の移動量を表し、赤い線がボトルからジュースをコップに移す過程を、青い線がコップのジュースを口元に移す過程を表している。このブラフから切り替えのタイミングを検出するため、グラフには示されていないが、手の移動と二つのジュースの移動とを比較検討し、次の規則を得た。

規則：移動量の顕著な下降の切り返し点が動作内容の変換点を表す。

図３に示した実験では､下肢の動きはなく、手の動きのみであるため、手と対象物の移動量の大きな期間は単純な移動過程にあることを示し、移動量がやや緩慢な期間は、手が行為・行動を特徴付ける操作を実行中であることを示していると考えられる。

一連の手の動き、一般的には身体あるいはその部位の動きの変化量が頻繁にしかも著しく変化する場合には、その行為・行動が複雑であることを示唆していると考えられる。

４．３ その他の行為・行動パターン

前項では、階層的・逐次的な行為・行動パターンを前提として認識手法を提案した。しかし理論上では必ずしも一般的なものとは言えない。それは下肢・上肢・手指はそれぞれ独立した身体部位であり、それぞれ独立に動作可能だからである。したがって、逐次的に作動する必要はないからである。階層構造の並列的な動きが支配的な行為・行動を今後に残された課題である。

４．４ モデルの構造

図４の a. と b.は横軸が処理対象の画像のフレーム番号を表し、ほぼ対応付けられているため、これらを比較しながら解釈すると、これまでの議論を検証できる。

しかし３．２の末尾に述べたように､図２，図３に示した状況を上述のような規則に従って認識するシステムには、摂食行為に関する知識が付与されていることが必要になり､２．２で述べた制約を満足する行為・行動の一般的動作モデルとともに、視覚あるいは距離センサのデータからこれらの知識の記述要素となる概念を導くための特徴量の決定とそれに基づく認識辞書が必要になる。

現在のところ、それぞれの要素としては次のようなものを考えている。

行為・行動：通常はまず、頭部および手領域の判定と各領域の移動か滞留かの判定が求められる。
頭部の移動については、水平移動と垂直方向移動の分類が必要である。
手領域については、行為者自身に近づける場合と、他の対象物に近づける場合とに２大別すべきと考えている。
同時に、帯同物の有無も類別を要する。

事物：基本的には、行為・行動に併せて、生成・消滅する領域、あるいは、見掛けの面積・形状の変化する領域であって、その属性は当初は不問である。

対象物の属性こそを行為・行動から決定しようとするところに、本稿提案の手法の特徴がある。また、主たる画像処理は、肌色領域の抽出とフレーム間差分とに限定されることも特徴である。

しかし、他の事物に「近づく」という関係の判定が必要になる。このとき、人物の３次元位置と環境に固定された事物の認識および３次元位置が求められ、同時にその幾何学的特性(水平面をもつなど)が必要となる。この処理は通常の物体認識である。

５．おわりに

以上の議論は、単純でかつ有用な解析手法を提供しているようであるが、提案手法が基本的には非実時間解析に適した手法であるという一つ大きな制約をもっている。

引き続き発表する報告[8]では、認識のレベルを画像データと３次元位置データというレベルまで引き下げ、これらの情報を統合し、それらから導かれる３次元的幾何情報および手・物体領域の変化情報を抽出し、対象と動作とを記述する特徴として用い、その記述の信頼性を確率の形を取って表している。したがって、動作（行為・行動）と対象との判定に関する評価の信頼性を、相互補完的に向上できることを示したものであり、本稿で論じた相互補完性とは異なるものである。それにもかかわらず、相互補完性が有効に機能しているのは、ベイジアンネットワークに備わる性質によるものと考えられる。

この手法が実時間性と整合性をもつと考えられる点も長所であり、板書行為も判定できる性能も評価できる。本稿に述べた手法と結合できれば、画像処理との親和性を期待できると考えられる。

参考文献

[1] 樋口未来，小島篤博，福永邦雄: “人間の動作と動作対象の関連性に基づくシーンの統合的認識”，画像の認識・理解シンポジウム(MIRU2004), pp.I-469-I-474, July 2004.

[2] Mirai Higuchi，Shigeki Aoki，Atsuhiro Kojima，Kunio Fukunaga: “Scene Recognition based on Relationship between Human Actions and Objects”，Proc. of 17th ICPR，Vol.3, pp.73-78, Aug. 2004.

[3] A. Kojima，T. Tamura，K. Fukunaga: “Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions,” International Journal of Computer Vision，Vol.50，No.2，pp.171-184, 2002

[4]吉田成希，北橋忠宏，”人間動作を通じての物体の機能的認識,” 信学会技報，PRMU2002-213, pp.13-18，Feb. 2003.

[5] 山平貴督，北橋忠宏：人間の動作に基づく物体の認識，情報処理学会,第66回全国大会,第２分冊4Z-7, Mar. 2004

[6] 小川原光一、高松 淳、木村 浩、池内克史，"観察に基づく手作業の獲得における視覚の利用," 情処学会論文誌:コンピュータビジョンとイメージメディア, Vol.44, pp.No.SIG17(CVIM8), Dec. 2003

[7] 山岡勝、山崎佳代子、田中弘美 :”仮想空間シミュレータ自動構築のためのハプティックビジョンに基づく物体間の水平支持接触拘束抽出,” 信学会論文誌, Vol.J84-D-II, No.7, pp.1439-1447, July 2001

[8] 樋口未来，小島篤博，北橋忠宏，福永邦雄：”協調型ベイジアンネットワークを用いた動作と動作対象の統合的認識,” 情処学会，CVIM 研究会資料，Mar. 2005