装着型能動視覚センサを用いた注目対象映像の獲得と理解

戸田　真人† 　鷲見　和彦† 　松山　隆司†

† 京都大学情報学研究科　〒 606-8501 京都市左京区吉田本町
E-mail: †{masatoda,sumi,tm}@vision.kuee.kyoto-u.ac.jp

あらまし　ウエアラブルビジョンセンサを用いて人の行動を記録したり認識することは，近未来の知能機械が人に質問したり指示を待ったりせずに自然なインタラクションを行うために重要な機能である．これまでカメラを装着して人と視覚情報を共有使用という試みは多く見られたが，人が持っている広い視野から人が注目しているものに焦点を当て，注目対象と人の意図とに即した的確な情報を提供できるような気の効いたシステムは実現できていない．そこで我々は人の視線情報を活用し，次のような手法を考案した．まず，人の視覚行動を分析して，ウエアラブルカメラとの情報共有に適した，注目対象に対する興味の有無を表す二つの状態に視線状態を分類した．次に，この視線状態と視線方向とを用いて，注視対象の三次元位置の計測を行い，注視対象を適切な視野と解像度で撮影した．さらに，注視視野情景の三次元情報，手の動き，注視点の位置と視線状態を用いて，人が対象物を手にとって観察したり移動させたりするという把持行動を，人をサポートするシステムにとって意味のある動作記述に分類できることを示した．この動作分類によって，例えば同じように手に物を持っていても，それを観察しているのか，操作しようとしているのかを判別し，人の意図に応じた適切な支援を選択することができるようになる．

キーワード　視線, ウエアラブル, ステレオ, カメラ制御, 行動理解, 把持

Acquisition and Recognition of Gazing Object Images with an Active Wearable Vision Sensor

Masato TODA†, Kazuhiko SUMI†, and Takashi MATSUYAMA†

† Graduate School of Informatics, Kyoto University　Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
E-mail: †{masatoda,sumi,tm}@vision.kuee.kyoto-u.ac.jp

Abstract　The aim of this research is to study a new method to obtain an image sequence and recognize human action through a wearable vision sensor. It will be an important component to realize a computer system, which will support a human at work or in his/her daily life. Although numerous efforts have been proposed for supporting a human achieving a task interactively, none could offer a support autonomously without prompting the human to be supported. To realize such a smart system, the essential function is to estimate interest and intention of the observed human. A new approach of acquiring a common sight with the user and estimating his/her action with his/her intention is proposed. It employs the gazing direction and recognizes gazing state and the target of interest. First we describe active image acquisition. In this method, gazing state is decided from gaze direction in a given period. The gazing state and the depth of the interesting target are used to control pan, tilt, and zoom of the active stereo camera to achieve optimal field of view capturing the interesting target. Then we describe classification of human manipulation action in a stereo image captured by the wearable vision sensor. Manipulation task is classified by the combination of hand-object-background relationships, hand motion relative to the interesting target, position of gazing point, and gazing state. Employing gazing point and state enables to classify apparently similar actions into meaningful states, each of which relate to a different supportive information to the person wearing the camera.

Key words　gaze direction, sight, wearable, stereo, camera control, action recognition

1. はじめに

人がコンピュータとが話したいときに自然と会話が始まり，助けが欲しいときにコンピュータの方から声をかけてくれるとしたら大変便利である．例えば，人が道具を用いて作業を行なう場合に操作方法や注意事項についての情報を提示したり，人が物に興味を持った場合にその物体についての詳細な情報を提示したり，また，後で再び参照するための映像や作業を記録することなどが期待される．これまで多くのインタラクティブシステムが提案されてきたが，その多くはあらかじめ決められた作業を支援するものだったり，システムと人とが質問応答を行って目的を絞り込むものだったりしていた．しかし実際に使うことを想定してみると，コミュニケーションするために多くの説明を要するようなシステムは利用者にとって負担である．人が今，会話を求めているか，会話する余裕があるか，助けを必要としているか，何をしているか，何に興味があるのか，行動の意図は何か，というような人の状態を推定することができれば，このような問題が解決でいるであろう．このような人の状態や意図・興味の認識には，音声や映像情報のほかその人の習慣や癖，置かれている状況や背景知識などを駆使せねばならないが，本研究では，これらの情報の中でも重要な役割を占める映像情報を，人と同じ視点で撮影するウエアラブルカメラと視線検出装置とを用いて取得し，認識することを目的とする．

近年，人の行動や状態などを認識するために，カメラによって撮影された映像を解析・認識するコンピュータビジョンの技術が多く活用されている．コンピュータビジョンを用いた人の行動や状態の認識は使用するカメラの構成によって，環境内にカメラを設置し，周囲から人を観察する環境埋め込み型ビジョンと，人がカメラを装着し，装着者と同じ視点から環境や人の体の一部を観察するウェアラブルビジョンとの二つのアプローチに分類できる．

環境埋め込み型ビジョンの優れた点は，人の行動を知るために環境を最適化できる点である．環境内で行なわれる定型的な作業に合わせてカメラ配置・視野の設定をあらかじめ行なうことができるため，認識のために必要な情報を必要となる精度で獲得することが可能である．崎田らは人の視線情報と環境埋め込み型ビジョンと組み合わせることで人とロボットとの協調作業を可能にすることを試みている [3]．作業場や店舗など人の行動の内容や目的があらかじめ決まっている環境では有用である．しかしながら，環境埋め込み型ビジョンは，システム自体が環境内に設置されているため，人を観察できる空間や見る方向が限られてしまい汎用性に乏しく，さらに，体や手，物体による遮蔽のため，行動認識のための重要な素材となる情報が獲得できない可能性があり，行動認識を困難にしやすい．

一方で，ウェアラブルビジョンの優れた点は，人と機械が視野を共有することによって，人が注意を向けた対象の映像を人とほぼ同じ見え方で取得することができることである．その映像を認識することによって，人が，環境中の何に着目していたか，また，その対象に対してどのような意図・注意・感情を持っていたかを推定できる可能性が高く，また，取得した映像を再び人の目の視点で提示できるためインタフェイス性に優れている．ウェアラブルビジョンの問題点として，人がカメラを装着することに負担がかかるという点が挙げられるが，今後の技術の進歩に伴い，より人の負担が少ないデバイスが開発されることが期待できる．そのため，本研究では，ウェアラブルビジョンを用いている．

これまでウェアラブルビジョンを用いて様々な研究がなされている．池田ら [2] は，視野映像の変化をもとに，装着者の位置や運動の推定を行なっている．また，青木ら [11] は，装着者の手の移動方向を検出することにより，装着者の動作の記述を試みている．これらの研究により，人の移動や動作を認識することが可能になった．しかし，これらの認識だけでは，人の行動の意図や興味を認識するには不十分である．人の行動の意図や興味を認識するためには，システムが人の注意などの心的状況を示唆する情報を獲得・認識する必要がある．そこで，中山ら [1] や榊原ら [9] は，視野映像と視線情報を用いて人が注目した注視領域を獲得された全視野映像から抽出する研究を行なった．これらの研究により，システムが視野映像を獲得するだけではなく，映像内のどの領域に人が注目しているのかを知ることが可能になったが，カメラの視野や解像度の制限により，注視領域を十分認識できるだけの大きさで捉えることができない場合がある．また，築澤ら [4] は，人が把持物体に注目している場合を想定し，把持物体の 3D ディジタル化を行なうことで，映像の記録だけではなくその形状の記録も可能にしている．しかし，入力には，あらかじめ適切な視野と解像度で撮影され，開始から終了まで切り出されたビデオシーケンスが使われており，視野の最適制御を行ない空間的にカメラの視野を選択することと，一連の映像を，人の行動や注目物体の変化をキーにして，時間的に意味のある区間に分割することは，まだ未解決な問題として残っている．

そこで本論文では，人の視線情報とウエアラブルカメラの映像から人の注視点の３次元位置と視線状態の認識を行う手法，注視点情報に基づき人の注目した対象を高解像度で撮影する手法，および，視点・手・対象物体の関係から人の把持行動を意味のあるパターンに分類する手法とを提案する．本研究では，図 1 に示す装着型能動視覚センサを使用する．装着型能動視覚センサは，アイ・マーク・レコーダと呼ばれる視線測定装置 (NAC 社 EMR-8) とパン・チルト・ズーム制御が可能なアクティブステレオカメラ (SONY EVI-G20×2) からなる頭部装着型のセンサである．従来の頭部装着型の視覚センサと比べ，装着型能動視覚センサは，次のような点で優れている．

- 人の視線情報を得ることで，画像中の人が見ている点を特定できる．
- カメラのパン・チルト・ズーム制御を用いて，視野中心・視野範囲・解像度を最適化できる．
- ステレオカメラを用いることで，映像だけではなく，その奥行き情報も取得することができる．

本論文では，このような装着型能動視覚センサの利点を活かし，注目対象の映像を最適な視野と解像度で獲得し，認識する．

図 1　装着型能動視覚センサ

本論文の構成は以下の通りである．まず 2 章で，人の視線の振る舞いについてレビューし，ウエアラブルビジョンにとって有用な視線状態を定義する．次に，3 章で，注目対象映像の取得について述べ，実験によりその有効性を示す．4 章では，注目対象映像の理解について述べ，実験によりその評価を行ない，最後に 5 章で結論をまとめる．

2.　視線の振る舞いに関する考察

人の眼球運動については古くから多くの研究が見られるが，たとえば読書のように何か目的を持った作業をしているとき，その動きは大きく，固視 (fixation または停留ともいう)，追従眼球運動 (pursuit), 跳躍眼球運動 (saccade) に分類される [5]. 固視は平均して 250msec 継続し，跳躍は数 10msec で完了する．跳躍時には網膜像は流れているが，その像は知覚されず [6], [7], 固視と追従の場合に視覚情報を処理していると考えられている．一方，ウエアラブルカメラによる視野映像の取得と理解を考えたとき，人の視線に完全に追従して数百 msec 単位で細かく視野が移動することは効果的ではない．視線に完全に追従してしまうと映像は跳躍運動に支配されて大きく揺れてしまうからである．そこで，固視と跳躍を大きなまとまりのある単位として取り出すことを考える．視線に動きを用いた画像の見方の分析 [5] によれば，一つの興味ある対象上に固視する点が複数存在し，それらの間を，跳躍している場合が多い．そこで，本研究ではこのような複数の固視する点をまとめて，一つの興味対象として捉えることを考える．

そのために，予備実験として次のような実験を行なった．被験者は図 1 に示す装着型能動視覚センサを装着し，被験者の手前 2m に置かれたボードに描かれた絵をみて，何が描かれていたかをできるだけ覚えようとするよう指示された．このような状態で，視線の動きを 1 分間記録し，終了後，被験者に同じ絵を見せて興味があったところを自己申告してもらい，固視の継続時間を興味の有無によって分類し，ヒストグラムを作成した．図 2 にその結果を示す．

この結果，興味のある対象の上では，小さな跳躍を挟んだ固視が 1.5 sec 以上継続するのに対し，興味の定まっていない場合には，固視はたかだか 1.5 sec しか継続しない．そこで，本研究においては，視線状態を興味ある対象がなく画像中の多くの場所を見ている状態と興味ある対象を連続的に固視している状態とに分け，それぞれ次のように呼ぶことにする．

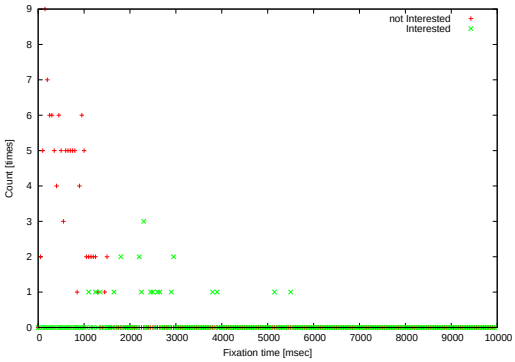


図 2　固視時間の分布 (緑：興味あり，赤：興味なし)

- 注目対象模索状態：様々な方向に視線を向けて，興味を持つ対象や行為を行なうために有用となる対象を探し続けている状態
- 詳細情報獲得状態：対象を認識し，その対象から何らかの感情を喚起させられている状態

現実には視線検出装置は 60Hz の周期で視線方向に相当する座標データ (u, v) を出力するが，その間に瞬きがあるとデータがとぎれたり，座標が大きく跳躍したりする場合もある．そこで，上に述べた視線の状態を表すために，連続して一定の範囲に視線がとどまっているかどうかを基準に，次のようにして決定する．まず，時刻 t_1 における視線方向 $\hat{s}(t_1)$ を次のように決定する．

$$\hat{s}(t_1) = \begin{cases} s(t_1) & \ldots \quad if\ |s(t) - s(t_1)| < \Delta s \\ & \quad (\forall s(t),\ t_1 - 0.25 < t <= t_1) \\ \text{invalid} & \ldots \quad otherwise \end{cases} \quad (1)$$

ただし，視線状態 $\hat{s}$ は，固視時には有効な値 $s(t)$ を持ち，跳躍または計測不能時には無効値 $invalid$ を持つと定める．$s(t)$ は時刻 t における有効な視線ベクトル $(u(t), v(t))$ を示し，Δs は一定範囲内に滞留していることを示すしきい値であり，本研究では 0.5 deg を用いた．式中 t の範囲を示す値 0.25 は固視が継続するとされる代表的な時間間隔を採用している．$\hat{s}$ を用いて，視線状態 S を次のように定義する．

$$S(t_1) = \begin{cases} 1 & \ldots \quad if\ |\hat{s}(t) - \hat{s}(t_1)| < \Delta s \\ & \quad (\forall \hat{s}(t),\ t_1 - 1.3 < t <= t_1) \\ 0 & \ldots \quad otherwise \end{cases} \quad (2)$$

ただし，視線状態 S において 1 が詳細情報獲得状態を示し，0 が注目対象模索状態または計測不能を示す．式中 t の範囲を示す値 1.3 は図 2 に示した実験から得られた固視が継続するとされる代表的な時間間隔 1.5 sec に対して，瞬きの平均値 0.2 sec 分のマージンを見込んだ値として決定した．

3.　視線方向と対象までの距離に応じたアクティブ撮影

人の注視点および注視対象は，その人の行為の意図や持っている興味を強く反映しているため，システムがそれらを共有することは，人の行動の意図や興味を理解するために非常に有用

であると考えられる．本章では，システムが以下の機能を持つことで，詳細な注視対象映像を取得し，より詳細な対象の理解を可能にすることを考え，これらの機能を，人の注視点の 3 次元位置抽出による注視対象位置の特定と 2 章で述べた視線状態の認識を行ないながら，最適な画像撮影パラメタを決定することで実現する手法を提案する．

- 視野内の注視点を測定する．
- 広い視野を持つ
- 注視点周辺では対象の詳細情報を得るために高い解像度を持つ．

人の注視点の 3 次元位置を知ることは，システムがその人が今環境中の何に注目しているかを知るだけではなく，人間の目と位置が異なるカメラ視点からの対象の方向や距離が計算できるため，注目対象の高解像度映像を取得するパン・チルト・ズーム制御を行なうことに非常に役立つ．従来，視線はその方向だけで記述されることが多かったが，近年，3 次元的な視点の位置を計測しようという試みがみられる [1], [8]．本論文では，視線測定装置から視線方向を得て，視線直線上でステレオ視を行なうことにより，任意形状の注視対象に対して高速でかつ正確な 3 次元注視点抽出を行なう．

ここで課題となるのは，カメラの限られた視野と解像度の中で，上記の広い視野映像の取得と注視対象の高解像度映像の取得とを両立させるためのカメラのパン・チルト・ズーム制御法である．本論文では，視線状態により，視線に追従させる制御と注視対象を高解像度で抽出する制御とを切り替えることで解決する．

3.1 3 次元注視点抽出法

人の注視点とは，その人の視線直線上の一点であり，注視点において，視線直線と見ている対象とが交わる．この 3 次元注視点位置を視線情報とステレオ画像を用いて抽出する．

視線測定装置から得られる視線直線をステレオ画像に投影する．この投影された直線は，ステレオ計測におけるエピポーラ線と等価である．(以下，この直線をアイ・エピポーラ線と呼ぶ)．片方の画像におけるアイ・エピポーラ線上の 1 点は，3 次元空間中の視線直線中の 1 点を示し，もう片方の画像におけるアイ・エピポーラ線上の 1 点と対応関係にある．ステレオ画像内で注視点にあたるアイ・エピポーラ線上の点の組では，そこには対象があるため類似度が高いものが写り，他の組では異なるものが写る (図 3)．この性質を利用し，ステレオ画像上の全ての点の組に対して，ブロックマッチングを行なうことでその点同士の類似度を比べ，最も類似度の高い点の組を注視点として抽出する．

この操作により，注視点の左右画像上の位置および，3 次元位置を知ることができる．また，本手法では，探索範囲を視線直線に限定することにより，計算量を抑えることができる．

3.2 カメラのパン・チルト・ズーム制御

3.2.1 映像取得モードの決定法

2 章で定義した人の視線状態には，2 つの状態があり，それぞれの状態では，人の対象にたいする興味の度合いが大きく異なるため，システムに求められる動作も変化する．そこで，視線状態により，カメラの制御を選択的に実行する．

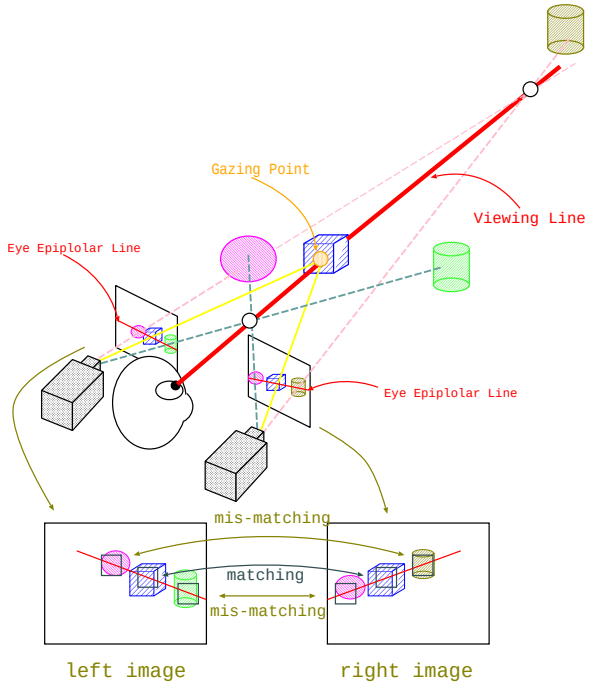


図 3　3 次元注視点抽出法

- 視線状態が注目対象模索状態の場合，カメラを最広角の状態で視線方向に追従するように角度制御を行い，見ている対象を含んだ広範囲の環境映像を取得する．
- 視線状態が詳細情報獲得状態の場合，それまでに得られた注視点に安定するように角度制御を行い，注視点の位置に応じたズーム制御を行なうことで，対象の全体像を大きく写した映像を取得する．

3.2.2 撮影パラメタ決定法

a) カメラのパン・チルトパラメタ

取得されたビデオ映像は，再び表示して利用することも考えると，急激な視野の移動は避けて，滑らかに視野を移動させるべきである．そこで，本論文では，角速度制御を行なうことにより，滑らかなカメラ運動を行なう．視線状態により，制御目標として視線方向と注視点方向を選択し，時刻 t におけるパン，チルトそれぞれの制御角速度 $v_p(t), v_t(t)$ を，目標値との誤差 $\epsilon_p(t), \epsilon_t(t)$ から，制御量更新周期 T を 0.5s とした PID 制御により，次式にように決定する．ただし，P, I, D は，定数である．

$$v_p(t) = P\frac{\epsilon_p(t)}{T} + \frac{I}{T}\sum_{t=0}^{t}\epsilon_p(t) + D\frac{\epsilon_p(t)-\epsilon_p(t-1)}{T^2} \quad (3)$$

$$v_t(t) = P\frac{\epsilon_t(t)}{T} + \frac{I}{T}\sum_{t=0}^{t}\epsilon_t(t) + D\frac{\epsilon_t(t)-\epsilon_t(t-1)}{T^2} \quad (4)$$

b) カメラのズームパラメタ

カメラのズームパラメタを決定する際，問題となるのは，視線状態が詳細情報獲得状態での注視対象全体の高解像度映像を取得するためのズームパラメタの決定法である．本論文では，問題を簡単にするために，注視対象の大きさを 1 辺 200mm の立方体を考える．そして，注視点の 3 次元位置情報を得ることで，投影中心から注視対象までの距離 D[mm] を知り，大きさ

S[pixel^2] の画像中で対象が p の割合を占めるように，次式のようにカメラの焦点距離 f[pixel] を計算することにより，最適なズームパラメタを決定する．

$$f = \frac{Sp}{200^2} D \tag{5}$$

本論分では，S を 640×480，p を 0.4 とし，計算を行っている．

3.3 実験による評価

3.3.1 注視点抽出実験

被験者に指示した点を見てもらい，本論文で提案した手法を元に注視点の 3 次元位置を抽出した．

被験者の注視点位置とシステムが抽出した注視点位置の誤差は，画像上においては平均で約 8pixel，3 次元空間中においては約 3800mm の距離の注視点に対して 64mm であった．

また，計算時間が 87ms(11.5frame/sec) であり，本システムでは，3 回の画像撮影 (15frame/sec) で 2 回の 3 注視点抽出が行える．この計算時間は，人の平均固視時間 (約 250ms) と比較すると十分早く，人の注視行動を測定することに十分適した時間であるといえる．

3.3.2 システムとしての動作実験

本論文で提案した注目対象映像取得システムを試作し，その動作を検証した．システムは，視線測定装置から視線情報を受け取り固視と跳躍を分別するモジュール EMR Receiver，ステレオ画像と視線情報から 3 次元注視点を抽出するモジュール Depth Finder と視線状態を認識しカメラ制御パラメタを決定するモジュール Sight Tracker and Mode Switcher からなり，松山ら [10] の提案したダイナミックメモリと呼ばれる通信インタフェースを用いることでモジュール間の情報通信の同期性を保ったまま，それぞれのモジュールを並列に動作させた (図 4)．

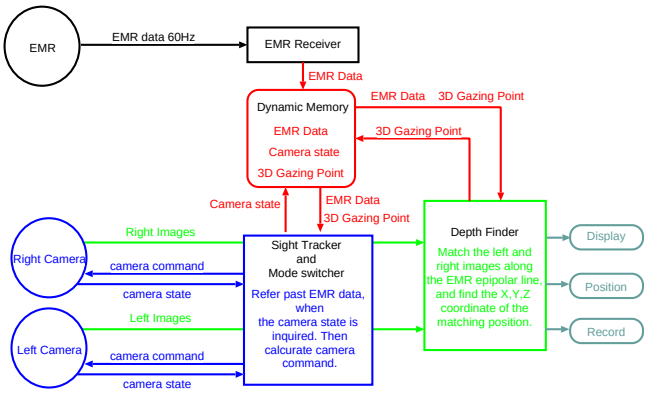
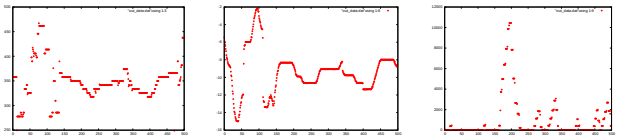


図 4　System Architecture

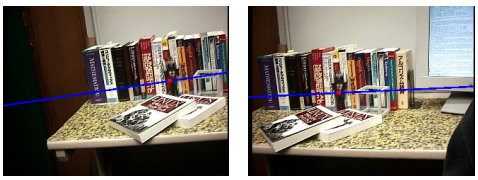
観測された視線情報と制御によるカメラパラメタの時間的変化の一例を図 5 に示す．カメラ角度の変化については，視線が変化する場合は，視線方向になめらかに追従し，視線の変化が少ない場合は，安定した動きを行なっている．また，視線方位の変化に対するカメラ角度の遅延は，最大で 1.2 秒であり，人の視線状態が詳細情報獲得状態になるまでに，注視対象をカメラの視野角内に収めることができている．カメラズーム値の変化については，視線の変化が少ない場合に，対象までの距離に応じたズームアップをしている．

実験結果の一例として，注視点抽出実験結果を図 6，7 に示す．図中の青い線はアイ・エピポーラ線，赤い点は抽出した注視点である．



(a) 視線情報 (垂直方向)　(b) チルトパラメタ (右カメラ)　(c) ズームパラメタ (右カメラ)

図 5　視線情報とカメラパラメタの時間的変化



左画像　右画像

図 6　興味対象模索状態での注視点抽出

左画像　右画像

図 7　詳細情報獲得状態での注視点抽出

実験結果より，以下のことが言える．

- カメラの視野は人の視野に比べて狭いが，カメラの回転制御を行なうことにより，広範囲にわたり測定できる．
- 目とカメラの位置が違うために，視線方向の情報だけでは，カメラの視野内に見ているものを収めるのが難しいが，カメラ制御を行ないながら注視点の 3 次元位置を得ることにより，注視している対象の画像をズームアップした映像を得ることができる．

これらのことより，本論文で提案した視線状態の認識と注視点抽出を元にしたカメラ制御により，人が注目した対象の映像を，広範囲に渡り，高解像度で得ることができると考えられる．

4. 視線を用いた把持行動の詳細分類

人の行なう一連の動作の中での動作対象に対する興味や意図の持ち方は変化する．したがって，人と協調して動作するシステムには，どの対象に人は注目しているのか知るだけではなく，その物体をどのように捉えているのかも知る機能が必要になる．本論文では，様々な人の行為の内,「持つ」・「動かす」・「観察する」などの人の興味や意図が強く現れる様々な動作を含み，そ

れらに応じたシステムの動作選択が必要となる把持行為に着目し，これらの動作を認識することを考える．

これまでのウェアラブルビジョンにおける把持行為の動作認識では，人の視野映像が入力された時に，手と物体の接触情報 (接触・非接触) や手の動き情報 (静止・運動) を観測し，把持行為における動作を,「掴む」・「持つ」・「動かす」・「無関係」の 4 つの種類に分類することが可能である．しかし,「持つ」動作には,「観察する」・「作用させる」などの意識的に異なる動作が含まれ，これまでの動作認識ではその人の行動の意図や興味を認識することは難しい．

そこで本論文では，これまで観測されている情報に加え，人の注意や興味などの主観を示唆する情報の一つである視線情報を用い，人の把持行為のシーケンスの中から，人が意識的に行ないシステムが働きかけを行なうのが望ましいと考えられる動作を検出・記述する手法を提案する．

本論文で観測する情報を図 8 に示す．これらの情報を組み合わせることにより，72 通りの動作を検出できる．しかし，人をサポートするシステムが人の行動の意図や興味を認識する観点から見ると，72 通りのうち，起こり得ないもの，人の意図が不明確でシステム側から見ると情報が不足しているもの，人が無関心でシステムに動作を求めないだろうと容易に推測されるものなど有効でない動作がある．本論文では，このような動作を除き，人間の動作とその時の注意を一意的に定めることができる図 9 に示す 7 通りの動作を検出する．各動作と観測される情報との関係を表 1 に示す．

- 手と対象の接触情報

モード 0: 接触していない (持っていない)
モード 1: 接触している (持っている)

- 手の動き情報

モード 0: 静止している
モード 1: 対象の方へ近付いている
モード 2: 対象から遠ざかっている
モード 3: 注視点の方へ近付いている
モード 4: 注視点から遠ざかっている
モード 5: その他の方向へ動いている

- 注視点位置情報

モード 1: 対象上にある
モード 2: 手の上にある
モード 0: その他の位置にある

- 視線状態情報

モード 0: 注目対象模索状態である
モード 1: 詳細情報獲得状態である

図 8　本論文で観測する情報

4.1　動作モードの分類法

本論文では，装着型能動視覚センサから得られる視線情報とステレオ画像から，観測する 4 つの情報のうち手と対象の接触情報，手の動き情報，注視点位置情報の観測をコンピュータビジョン分野における基礎的な技術を組み合わせることにより行なう．そして，観測された情報を表 1 と照らし合わせることで，人間の行動を分類する．処理の手順を以下に示す．

動作 1：　対象を手に持っていないが，注目している
動作 2：　対象に注目しながら掴もうとしている
動作 3：　対象の持ち方に注意しながら掴もうとしている
動作 4：　対象を手に持ち，対象を観察している
動作 5：　対象を手に持ち，扱い方に注意している．
動作 6：　対象を，対象に注意を払いながら動かしている
動作 7：　対象を，移動先に注意を払いながら動かしている
動作 0：　その他の動作をしている

図 9　本論文で記述する動作

表 1　記述する動作と検出する情報の関係

	手と対象の接触	注視点位置	手の動き	視線状態
動作 1	非接触	把持対象	静止	詳細情報獲得
動作 2	非接触	把持対象	把持対象	詳細情報獲得
動作 3	非接触	手	把持対象	詳細情報獲得
動作 4	接触	把持対象	静止	詳細情報獲得
動作 5	接触	手	静止	詳細情報獲得
動作 6	接触	把持対象	他の場所	詳細情報獲得
動作 7	接触	その他	注視点	詳細情報獲得

（1）　手と把持対象領域の検出

ステレオ画像から距離画像を生成し，背景距離情報と比較することで手または把持対象領域を検出する．検出された領域から肌色検出を用いて手領域と把持対象領域を分離する．

（2）　手と把持対象の接触情報の観測

画像上に置ける手と把持領域の連結性と領域での連結部における距離画像での連続性を評価することで，手と把持対象が接触しているか判別する．

（3）　手の動き情報の観測

画像における手領域の時間的変化を元に画像上の手の動きベクトル $\mathbf{u}$ を求め，$\mathbf{u}$ と距離画像での手領域の Depth の変化から 3 次元空間中の手の動きベクトル $\mathbf{U}$ を算出する．手の重心から把持対象および 3 章で提案した手法で得られる注視点の 3 次元位置へのベクトル $\mathbf{H}_o, \mathbf{H}_e$ と $\mathbf{U}$ のなす角から，手の動きを図 8 の 6 つのモードに分類する．

（4）　注視点位置情報の観測

注視点の 3 次元位置情報と，検出された領域および距離画像を比較することにより，図 8 の 3 つのモードに分類する．

（5）　動作の分類

上記のように検出された手と把持対象の接触情報，手の動き情報，注視点位置情報および 2 章で提案した視線状態情報の 4 つの情報と表 1 を照らし併せることで，その時点での人の動作や注意を図 9 の動作に分類する．

4.2　実験による評価と考察

本章で提案した手法を評価するために実験を行なった．被験者が行う作業として，以下に示す作業を設定し，被験者が実際に行なっているシーンを装着型能動視覚センサを用いて撮影し，行為の中から定義した 8 つの動作を検出した．

- 直方体を指定された穴へ差し込む動作を行なう
- 複雑さを負荷するため，効き手とは逆の手片手で行なう

なお，実験環境として，テクスチャが十分あり，ステレオ画像の対応点づけがしっかりと行なえる環境を選択した．

手および把持対象領域検出処理の 1 例を図 10 に示す．検出結果を見れば分かるように手と対象の領域 (図 10(f) の黄色と水色の領域) が正しく検出されていることが分かる．

(a) 右撮影画像　(b) 左撮影画像　(c) 距離画像

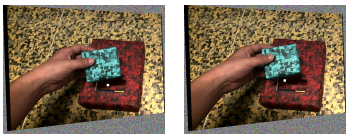
(d) 手/対象領域検出結果　(e) 肌色検出結果　(f) 手と対象領域検出結果

図 10　手および把持対象領域検出

次に，手と把持対象の接触情報の検出処理の例を図 11 に示す．左側の画像では，接触していないと判断され，右側の画像では，接触していると判断された．

図 11　手と把持対象の接触情報検出

次に，手の動き情報検出の例を図 12 に示す．視野映像が入力された時，その直前の入力画像と比較した結果，手の移動ベクトルと手と注視点を結ぶベクトルのなす角 θ_e は，$\cos\theta_e = 0.985$ という値が得られ，手は注視点方向へ近づいていると判断され，システムが手の動きを正しく検出できている．しかし，手がひねるなどの回転運動を行っている場合，全体的な移動はしていないが移動ベクトルが検出され，誤った判断をしてしまう場合があった．



(a) 入力画像　(b) 参照画像

図 12　手の動き情報検出

注視点位置情報の検出処理の例を図 13 に示す．左側の画像では，注視点は，対象上にあると判断され，右側の画像では，その他の場所にあると判断された．これらの結果より，本手法により，手と把持対象の接触情報を正しく検出できていることがわかる．

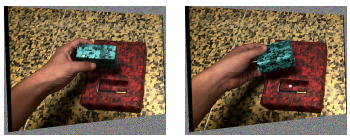


図 13　注視点位置情報の検出

システムが検出および記述した動作の時間的変化を図 14 に示す．各グラフの縦軸の値は，図 8 のモード番号および図 9 の動作番号に対応する．この検出結果と実験後被験者が記述した動作の時間的変化を比較した．全観測データ 550frame 中，システムが検出した動作と被験者が記述した動作が一致した frame 数は，331frame であり全体の 60.2%であった．

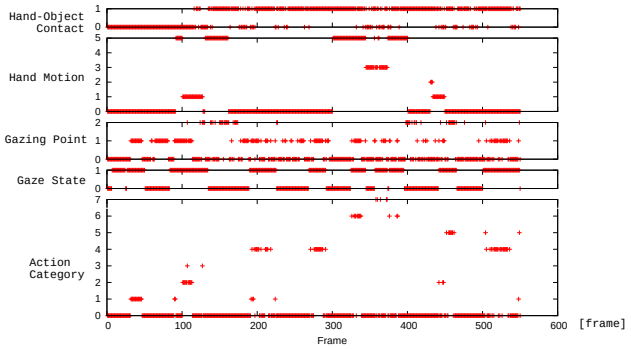


図 14　把持行為における動作の検出および記述結果

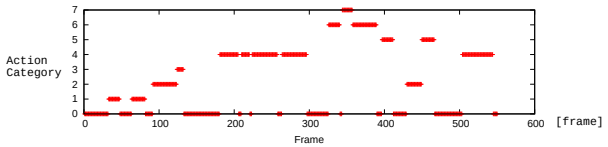


図 15　把持行為における動作の記述 (被験者の主観)

この検出エラーについて，詳細を表 2 に示す．被験者が行為に注意を払っていないと記述した動作 (図 15 中の値が 0) 全 219frame の内，システムが何らかの注意を払っていると判断した frame(図 14Action Category 中の値が 1〜7) が 19frame，被験者が注意を払って動作している記述した動作 (図 15 中の値が 1〜7) 全 331frame の内，システムが他の注意を払って動作していると判断した frame (図 14Action Category 中の値が 1〜7) が 10flame，動作に注意を払っていないと判断した frame(図 14Action Category 中の値が 0) が 189flame であった．

表 2　動作認識エラーの分布 (全 550frame)

		システムの観測結果	
		動作 1〜7	動作 0
人の主観	動作 1〜7	10frame	189frame
	動作 0	19frame	×

実験結果から分かるように，システムが行った動作記述結果と被験者が記述した動作の一致度が 60%であり，全体として高い数字が得ることができなかったが，人が動作に注意を払っていない時にシステムが注意を何らかの払って動作をしている

と検出したり，人が注意を払って動作している時にシステムが他の注意を払って動作していると検出するエラーは，全体の約5%と低い値になっている．この結果は，本論文で提案する手法を用いて，人の動作認識を行った場合に，人を支援するシステムが，人の必要としない働きかけを勝手に行う可能性が低いことを示している．従来の動作認識では人が必要とする情報を定めること難しかったが，本論文で提案する視線状態と注視点の位置情報を付加した動作認識によって，人を支援するシステムが，人が作業を行う中で必要になった情報を的確に提示することが可能になるといえる．

動作認識エラーの約 9 割が，被験者の主観では注意を払っていると記述された動作 (図 15 において値が 1〜7) に対して，システムが動作に注意を払っていないと判断したもの (図 14Action Category 中の値が 0) であった．この理由として視線状態の認識エラーが挙げられる．本論文では，人の眼球運動の中で固視の継続時間に着目し，人の視線状態を分類したが，実験では，手や対象の動きに合わせて眼球を動かす追従眼球運動が多く見られ，試作システムではこの状態を，注目対象模索状態と判断してしまっていた．今後，追従眼球運動を認識する機構を構築し，追従眼球運動を行っている状態における人の興味の持ち方を認識することで，動作記述の一致度を高くすることが可能になると考えられる．他の理由としてステレオマッチングのエラーや肌色検出のエラーが挙げられるが，これらは実装上の問題であり，今後，解決することが可能な点といえる．

また，本実験は，処理速度を考慮せず，オフラインで行った．1frame あたりの処理時間は約 1 分 30 秒であり，大半は距離画像生成処理時間であった．人を支援するシステムには，実時間応答性が必要であり，現在のシステムはこの要求を満たしていない．しかし，本システムに，これまでに様々な場所で研究されている実時間距離画像生成アルゴリズムを導入することにより，把持行動の詳細分類の実時間を行うことが可能になると考えている．

5. 結　　論

人間の行動や興味の理解のために，装着型能動視覚センサを用いた人の把持行為における人の動作の記述に取り組み，以下の 3 つの手法を提案し，実験による評価を行なった．

（1） 人の視覚行動について分析を行ない，視線状態を 2 つの状態に分類した．

（2） 人の視線状態の認識と注視点の 3 次元位置を計測を元にしたカメラのパン・チルト・ズーム制御を行なうことにより，人の意図や興味の対象を，広範囲に渡り，ズームアップした映像を獲得する手法を提案した．

（3） 人の把持行為に注目し，獲得映像から注視点・手・把持対象の位置情報を認識し，それらの位置関係やその変化により人の動作解析を行なった．従来の対象や手の動きだけではなく，視線・視点という動作の主体の情報を付加することで，人の意図や興味を含んだ動作の記述を行なえることを示した．

今後の課題として以下のことが挙げられる，

- 手の運動の詳細な検出

本論文で提案した手法では，手の運動について，移動運動の検出を行ったが，ひねりなどの回転運動や指を曲げるなどの局所的な運動については考慮しなかった．今後このような手の運動を詳細に検出する必要がある．このような検出を行なうためには，手のモデルを計算機内に持ち，獲得された映像とフィッティングを行なうなどの，より高度な処理が必要になる．

- 人の眼球運動の認識

本論文では，人の固視時間に着目し，人の注意の状態を推測した．今後，追従眼球運動などを認識する機能を構築する必要がある．

- 行動の詳細なカテゴリー化

手の運動の検出がより詳細になると，行動のカテゴリー化がより細分化できると考えられる．また，今どこに注目しているかという情報だけでなく，その時間的変化を認識することにより，例えば「比較している」などの行動が記述できるのではないかと考えられる．

- 記述した動作を元にしたシステムの動作選択

記述された動作を元にシステムに期待されている動作を推測し動作する人間支援システムを構築する．

謝辞

本稿の一部は２１世紀ＣＯＥプログラム「知識社会基盤構築のための情報学拠点形成」によるものである．また, 本研究の遂行に当っては科学研究費補助金特定領域研究 13224051 の支援を受けた．

文　　献

[1] A.Sugimoto,A.Nakayama and T.Matsuyama:“Detecting a Gazing Region by Visual Direction and Stereo Cameras”,Proc.of ICPR2002 Vol.3,pp.278-282,2002.

[2] 池田, 杉本, 井宮: “注視点対応とオプティカルフローを利用したカメラ運動の推定”, 情報処理学会コンピュータビジョンとイメージメディア研究会, SIG-CVIM-138, pp. 105–112, 2003.

[3] 崎田健二，小河原光一，木村浩，池内克史：“視線を利用した人間とロボットの協調作業”, 第 21 回日本ロボット学会学術講演会講演論文集, 2003.

[4] Sotaro Tsukizawa, Kazuhiko Sumi, Takashi Matsuyama.: “3D Digitization of a Hand-Held Object with a Wearable Vision Sensor”, ECCV Workshop on HCI, Prague, Czech Republic, pp.129-141., May 2004.

[5] 苧坂良二，古賀一男，中溝幸夫 (編), “眼球運動の実験心理学,” 名古屋大学出版会，1993.

[6] 池田光男：“目は何を見ているか–視覚系の情報処理”，平凡社，1988

[7] 乾敏郎：“視覚情報処理の基礎”，サイエンス社，1991

[8] 満上育久, 浮田宗伯, 木戸出正継：“視線情報を用いた注視点の 3 次元位置推定”, 電子情報通信学会技術報告書 PRMU，Vol.102, NO.334, pp.7 ー 12，2003

[9] 榊原章仁，浮田宗伯，木戸出正継：“視線履歴を用いた注視領域抽出”，電子情報通信学会技術報告書 PRMU，Vol.102, NO.554, pp.1-6, 2003.

[10] T.Matsuyama, S.Hiura, T.Wada, K.Murase, A.Yoshioka: “Dynamic Memory: Architecture for Real Time Integration of Visual Perception, Camera Action, and Network Communication”, Proc. of Comuter Vision and Pattern Recognition Conference, pp.728-735, 2000.6

[11] 青木茂樹，高橋昌史，大西正輝，小島篤博，福永邦雄：“ウェアラブルカメラによる動作認識とテキスト表現”，電子情報通信学会技術研究報告，IE2003-37, PRMU2003-67, MVE2003-49, pp.59-64, July 2003