

画像に基づく屋内シーン変遷の自動検知と 対話的イベント検索システム

Image-Based Automatic Detection of Indoor Scene Events and Interactive Inquiry

牧和宏, 島田伸敬, 白井良明

Kazuhiro MAKI, Nobutaka SHIMADA and Yoshiaki SHIRAI

立命館大学大学院 理工学研究科

maki@i.ci.ritsumei.ac.jp, {shimada, shirai}@ci.ritsumei.ac.jp

概要 屋内シーンで起こったイベントを検知・索引付けしておき、後からユーザが希望するシーンを検索できるシステムの提案を行う。シーン変化を検知し、持ち込まれた物体が持ち去られる時に検出されるであろうシーン変化を予想することにより、物体に隠れがあっても正しくイベントを識別する。また、ユーザが現場で直接物体や空間を指差しながら音声で問い合わせることができる、直観的なインタフェースを実装した。システムとユーザが対話的にシーンを検索することにより、システムの自動認識の誤りを訂正し、その誤りに関連した他の自動認識の誤りも訂正する。

1 はじめに

近年、犯罪数の増加と共に監視カメラや映像レコーダといった、セキュリティ用映像装置の需要が高まり急速に普及してきている。米国ではDARPA(Defence Advanced Research Projects Agency)の基、複数のカメラが協調して動作するビデオ監視システムの研究プロジェクトVSAM(Video Surveillance and Monitoring)が行われている[1]。このような屋外の監視映像を対象とするシステムでは、天候等による環境変動の影響を受けずに侵入物体を検出することが重要となる。

一方、室内環境を監視するシステムは、多数の人物の出入りなどで観測する画像が複雑になり移動物体の抽出は困難になってくる。加えて画像から人間の行動を理解する研究も盛んに行われ、室内の物体を操作したり椅子に座ったりしたことを検知するインテリジェントルームの研究や[2]、どこに何があるのか、誰が物を「持ち込んだ/持ち去った」のかを認識・管理するシステムの研究開発が行われている[3]。身に着けた画像センサの情報から、ユーザ自身の作業内容や物体の置き忘れなどを記憶・報告してくれるシステムの研究もある[4]。また部屋内のユーザの動作を認識することによって家電機器を操作するジェ

スチャインターフェースが研究されている[5]。その他にもWWWを用いて、ユーザがブラウザから見たいシーンの表示・検索ができるシステム[6]もある。シーン変化や室内の物体の種類を識別する技術は、画像認識の分野で研究されているものの[7, 8]、多様なシーンに対して全自動で誤りなく認識することは現状でまだ困難である。そういった避けがたい自動認識の誤りを、ユーザがシステムとインタラクティブにやり取りを行うことで誤りを訂正し、物を取ってきたり情報収集を行うサービスロボットが研究されている[9, 10]。これらの研究では、多量のデータ収集や情報処理を機械にまかせ、暫定的に出力された認識結果(冷蔵庫の中のどこにどんなものがあるか)をユーザが確認して、誤りがあつたり何かの影に隠れて見つからない場合には、「〇〇の後ろ」などのアドバイスをを行うことで認識誤りを回復し、作業を完遂する。このように、機械が不得意な場面での人間の識別能力は確かに正確ではあるが、監視カメラに保存された長時間の記録映像から目的のシーンを人手で探し出す作業は多大な労力を要し、見逃しの問題も抱えているため、安易にユーザが介入すれば解決するわけではない。

そこで、このような大量の映像データをシステム

が自動整理して効率よく保存しておき、ユーザは必要最低限の指示を行うだけで対話的に検索することができれば便利である。監視映像の検索性を向上させるには、自動画像処理によって重要な場面を検知し索引付けする技術が不可欠である。また自動認識の誤りがおこった場合に、ユーザが簡単に適切な指示を行って再検索できるインターフェースがあることが望ましい。

本研究では、監視映像下にかかる「物体の持ち込み/持ち去り」のイベントをある程度自動的に検知・整理しておき、後からユーザが「物体を持ち込んだ・持ち去った」シーンを特定できるシステムの開発を試みた。また、ユーザが放置された物体やかつて物体があった場所を直接現場で指差しながら、「これを持ってきたのは誰?」といった直観的な問い合わせができるようなインタフェースの実装も試みた。

2 システムの概要

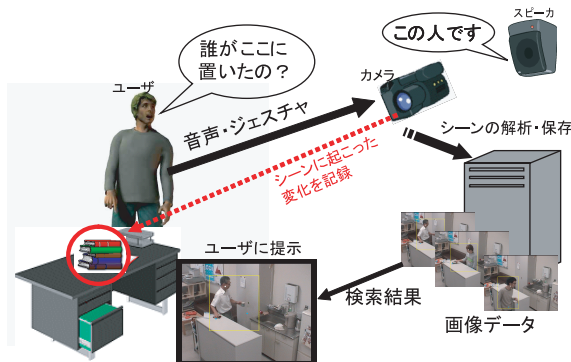


図 1: システムの概念図

図1にシステムの概念図を示す。システムの機能は、

- イベント検知部
- イベント解釈部
- ユーザインタフェース部

の3つからなる。

イベント検知部は、天井に設置されたネットワークカメラからの入力画像を処理するための計算機によって構成される。ネットワークカメラからの画像の取得にはHTTPプロトコルを用いている。常時、部屋の中の映像の処理を行い、映像中に人を検知すると“Human-database”にそのシーンを保存する。さらに、人間の動作に注目していて、人が何か物を「持ち込んだ/持ち去った」シーンを自動的に検知し“Event-database”に保存する。

イベント解釈部は、イベント検知部で“Event-database”に保存されたシーンに「いつ? どこで? だれが? 何をしたか?」を索引付けし、その情報をデータベースで管理する。イベント解釈部では、検知された画像の処理が必要だが、イベント検知部と並列に処理させることにより、イベント検知部がイベント解釈部の処理を待つことなく画像を取得できる。

ユーザインタフェース部は、マイクとスピーカ、計算機から構成される。イベント検知部と同一のネットワークカメラから画像を取得し、ユーザが現場で物体やかつて物体があった場所を指差しながら発話できるよう、ジェスチャ認識と音声発話の処理を行っている。ユーザから問い合わせがあれば、保存されている画像データを検索し、物を持ち込んだ・持ち去った人物の映っているシーンをユーザに提示する。なお音声対話は、音声認識エンジン Julius/Julian[15]を用いた対話システム [11] を拡張し実装した。

3 屋内シーン変遷の自動検知

3.1 人と物の検知

イベント検知部で行う物体検知には、文献 [12] の手法を用いる。本手法の概要は次の通りである。まず、急激な照明変化にロバストな背景画像を推定し、入力画像との差分をとり、前景領域を検出する。次に、前景領域から影を取り除き [13]、人の持つ特徴（大きさ、肌色、髪色）を用いて人領域を検出する。この時、人領域から顔、手、指先を検出する。人を検知したシーンは“Human-database”に保存しておく。人領域以外の領域を物体候補領域とし、ある一定区間同じ場所に観測された候補領域を物体領域とする。物体領域が見つければ、そのフレームから過去一定フレームをイベント発生区間とし、その区間の画像列をまとめて“Event-database”に保存する。図2に物体検知の様子を示す。図2-(a)は人が物を置いた時の画像である。このフレームでは人と物が離れていないので物が置かれたか検知できない。図2-(b)と(c)では物体候補領域が観測されている。この時点ではまだ連続観測されたフレーム数が一定値に達していないので、物体領域候補として観測を続ける。図2-(d)のフレームにおいて一定フレーム、物体候補領域を観測したので、初めて物体領域を検知したことを示している。一定フレーム数連続して観測された物体領域候補を物体領域とすることで、一瞬物体の前を横切る人がいても安定して検知することができる。検知されたイベントがどういったシーンであった

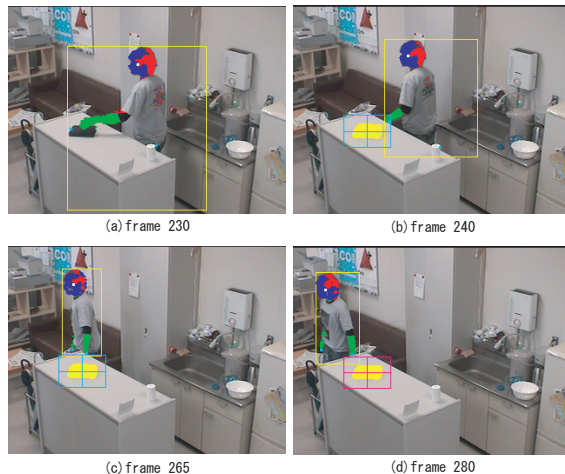


図 2: 物体検知の様子



図 3: 指差しジェスチャの例

かは、イベント解釈部で判定され、情報を付与する。

3.2 指差しジェスチャの認識

ユーザが音声でシステムに問い合わせた時、ユーザインタフェース部はユーザが指差しジェスチャを行っているかどうかチェックする。指先が体の中心線から十分離れている場合（図 3-(a)）、システムは指差ししていると認識する。指差しを検出したら、指先付近の領域をユーザが指定する領域とし、その場所の座標をシーン検索に用いる。ユーザが正面を指差ししていた場合は（図 3-(b)）、手領域の先端部分が狭くなっていれば指差ししていると判定し、ユーザが指定する領域を決定する。

4 レイヤ表現を用いた「持ち込み／持ち去り」イベントの識別

イベント検知部で検知したイベントは「持ち込み／持ち去り」の両方を含んでいる。イベント解釈部ではイベント検知が起こった時、どの物体が「持ち込み／持ち去り」だったのか識別を行なう。識別は、レイヤ表現を用いて事前に持ち込まれた物体が持ち去られる時に検知されるシーン変化を予想することにより行う。イベント検知が発生するたびに、シーンの記述を逐次データベースに書き込んでいく。シーン

の記述は、検知されたシーンに「いつ？どこで？だれが？何をしたか？」を索引付けするもので、イベント検知部で保存された画像列を代表する 1 画像をキーフレームとして、物体領域の面積、物体の重心座標、検知した時刻、シーン解釈結果を記述したものである。現在のシステムはシーン解釈部で誰がイベントを起こしたか分からないが、イベント検索時の対話の中で、結果画像に映っているのが誰かをユーザに尋ねたり、顔識別をさせることにより、シーンの記述に「だれが？」のインデックスを付与することができる。

また、シーンの記述以外に検知された物体が置かれる前の背景の画像パターンを覚えておき、これを用いて次に説明するイベント解釈アルゴリズムを実行する。

4.1 事前予想に基づくイベント解釈アルゴリズム



図 4: 持ち込み・持ち去りの例

「持ち込み／持ち去り」の判定は図 4 に示すように、ある物体 O_i が (a) で持ち込まれ、その後 (b) で持ち去られたシーンを見ると、検知される物体領域はほぼ同じと考えられる。そのため、イベント検知した時の物体領域と、それより過去に時系列に形状のマッチングを行う事で持ち込み／持ち去りを判定する。形状のマッチングは現在のマスクが過去と一致するか論理積 (AND) をとって調べる。論理積をとって 80% 以上の面積が真であれば一致とし「持ち去り」と記述する。図 4 の場合は (a) と (b) の物体領域の形状がほぼ一致したので、物体 O_i が持ち込まれその後、持ち去られたと記述できる。

日常生活のシーンでは机の上が散らかるなど、物体同士のオクルージョンは頻繁に発生し、物の形状が完璧に抽出できない場合がある。そこで物体同士のオクルージョン問題を解決するために、検知された物体領域を時系列のレイヤ構造として考える。移動物体の重なりを考慮した物体検出方法は、レイヤー型検知法 [14] があるが、この手法では物体のオクルージョン下で起こったことに対する処理、例えば、物体

を下から抜き取ると、物体の隠れている奥に物が置かれるといったシーンは対応できない。そこで、次状態の物体領域の予想を行う。次状態の物体領域は、持ち込まれた物体が持ち去られる時のシーン変化を物の重なりを考慮した上で予想する。具体的には、物体が持ち去られることにより得られる形状とその結果現れる背景のテクスチャを予想している。積み重なったレイヤの上から順に、画像中に見えている部分の形状で次状態の「持ち去り」予想画像を生成する。レイヤ表現に基づく事前予測によるイベント解釈アルゴリズムは以下のように定義した。

- Step1** ある物体領域と、全てのレイヤごとの予想した次状態の物体領域との前後関係の判定
- Step2** 同じ形状があれば、そのレイヤの物体領域を「持ち去り」領域としそのレイヤを削除する
- Step3** Step2でレイヤを削除したら、全てのレイヤの次状態の形状の予想を行い Step1に戻る
- Step4** 次状態の形状とマッチしなければ「持ち込み」領域とし、新規レイヤとして重なっているレイヤの最大値より1つ上のレイヤとして追加する
- Step5** レイヤ追加後に「持ち去り」によって検知される差分領域を予想し形状を生成する

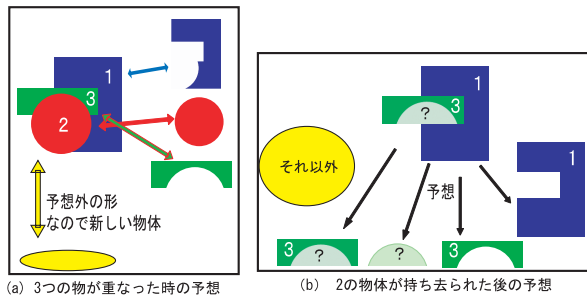


図 5: オクルージョン状態におけるの事前予想

図 5-(a) は番号の順に物体を検知したときの予想を示す。この時、予想されるのは3つの物体の重なりを考慮した形状か、それ以外の持ち込み時の物体形状のみとなる。図 5-(b) は (a) の状態から2の物体が持ち去られた時の状態を示す。予想で2の下レイヤにある領域をシステムは、見たことが無いので「？」の領域としておく。この領域と同じ階層のレイヤをつなぎ予想することでどの物体が持ち去られても対応することができる。

4.2 イベント解釈例

図 6 は、物が奥に隠れた形で持ち込まれた場合に、物体領域を予想しイベント解釈を行なう例である。

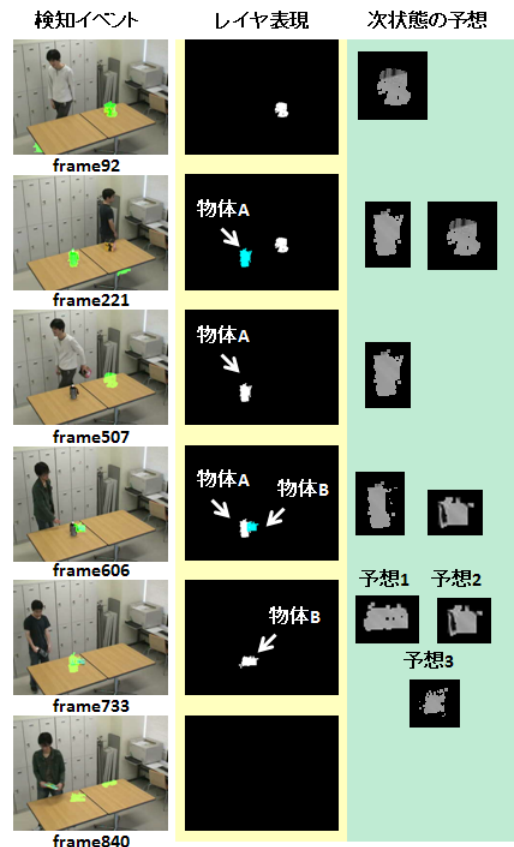


図 6: イベント解釈の成功例

frame92 で物体を検知し、何も物が置かれていないのでレイヤを追加する。frame221 では検知した領域は frame92 の領域とマッチしないので、物が持ち込まれ2つの物体があることを示している。frame507 で検知した領域は、frame221 で予想した形状とマッチするのでレイヤを削除する。frame606 では、過去に置かれた物の奥に物が持ち込まれる様子を示している。frame606 では、物体 A(frame221 の物体) は物体 B(frame606 の物体) の1つ下のレイヤとシステムは解釈している。図 7-(a) に示すように実際は

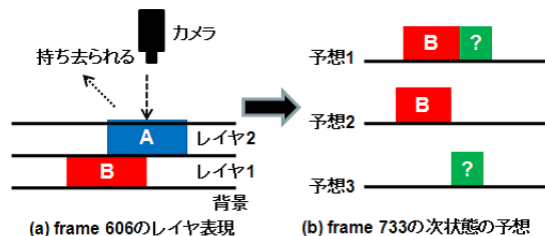


図 7: frame733 における次状態の予想

frame606 のレイヤ構造は物体 B の1つ上のレイヤに物体 A がある。frame733 で物体 A が持ち去られる事で現れる物体 B の領域は、システムは見たこと

が無い領域なので「？」の領域が現れる。「？」の領域は、ある物の奥に物が置かれた場合の例外措置として、

- 同レイヤ近辺の frame606 で検知した領域と結合した形状 (図 7-(b) の予想 1)
- 物体 B の形状 (図 7-(b) の予想 2)
- 「？」の領域の形状 (図 7-(b) の予想 3)

の 3 つを「持ち去り」時の形状として予想しておく。これにより、実際に frame840 で検知した物体領域が frame733 で予想した形状 (予想 1) とマッチしたので、物体 B と「？」が同一物体だと分かり、「持ち去り」と解釈してレイヤを削除することができる。

5 情報検索・提示のための ユーザインタフェース

5.1 対話インタフェースの構造

図 8 に対話を用いたシーン検索の概念図を示す。ユーザは指差しと音声発話によりシステムに問い合

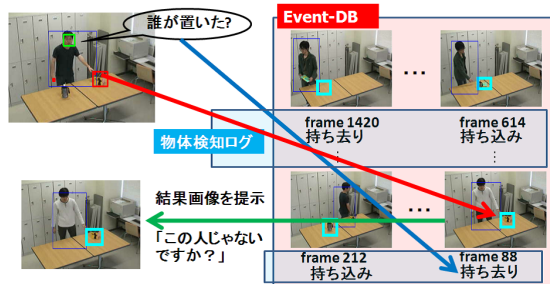


図 8: 対話を用いたシーン検索

わせる。ユーザが問い合わせると、システムは指差し場所付近で起こったイベントシーンを探し、検索結果の候補とする。さらに、音声認識により得られる情報 (イベントの種類や時間情報) を用いて、検索結果を絞り、最有力の検索結果をユーザに提示し、その結果が正しいか確認する。もし、ユーザから検索結果が間違っていると指摘されれば、ユーザが希望するシーンが見つかるまで他の検索結果を提示し、確認を続ける。

5.2 対話的シーン検索に基づく自動認識の誤りの訂正

イベント検知・解釈は自動画像処理によって行われるため、システムが誤認識を起こす場合がある。図 9 にシステムが誤認識した結果を示す。図 9-(a) でシステムは新しい物体 C を持ち込んだと判定したが、実

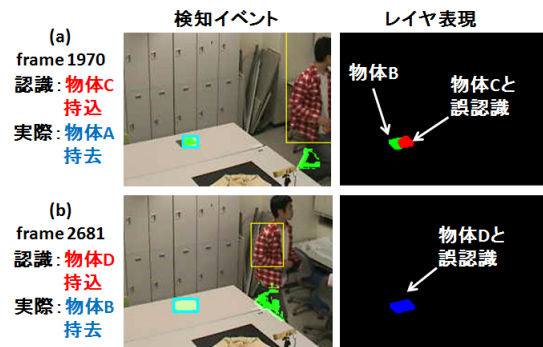


図 9: システムの誤認識結果

際は既に置かれていた物体 A を持ち去っている。これはシステムが物体 A を持ち込んだシーンを見逃したために誤認識になった。さらに、この誤認識のために図 9-(a) での持ち去り時の予想領域を間違い、図 9-(b) で物体 B を持ち去ったシーンを新しい物体 D を持ち込んだと誤認識してしまっている。

この状況で、ユーザが物体 A を持ち去ったシーンについて問い合わせた場合、システムはそのシーンを見つけては出ない。システムは "Event-database" から目的のシーンを見つけれない場合、そのシーンを見逃したかもしれないと考え、"Human-database" から目的のシーンを探す。なぜなら、多くのイベントは人によって起こされるので、"Human-database" に見逃したシーンが保存されているかもしれないからである。そこで、"Human-database" から目的のシーンを探すために、ユーザとのインタラクションを用いる。システムは "Event-database" のシーンの動画を順にユーザに提示しながら、ユーザが指定した物体が映っているシーンと映っていないシーンを教えてもらう。今回は、図 9-(a) の動画を見せた時に、ユーザはこのシーンが目的のシーン (物体 A を持ち去っている) と分かる。その旨をシステムに伝えれば、システムは図 9-(a) が持ち去りシーンと訂正でき、さらに持ち去られた物体が以前置かれたシーンを探す。

今回は、以前検知したシーンに該当しそうなものがないので、"Human-database" から見逃したシーンを探す。ユーザに該当物体が映っているシーンと映っていないシーンを教えてもらい、自動画像処理によって、指定物体が映っていないシーンから映っているシーンの間の期間で "Human-database" から目的のシーンを探し、frame872 にそれらしいシーンを検知することができ (図 10)、これを "Event-database" に追加する。追加することにより、それ以降のシーンの解釈に変更がないか検討し、最終的に図 11 のよう

な認識結果が得られる。

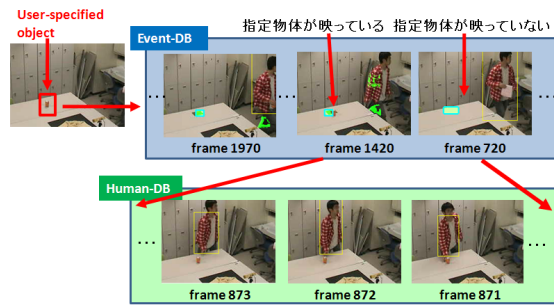


図 10: 見逃しシーンの検索例

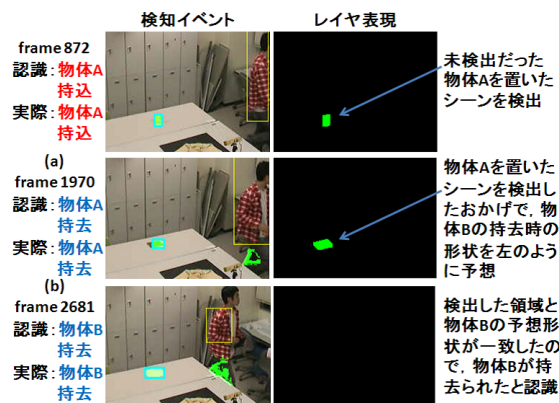


図 11: 訂正後のシステムの認識結果

6 まとめ

シーンに起こったイベントを自動検知し、解釈・保存するシステムを提案し、後からユーザが希望するイベントシーンを指差しと音声対話を使って簡単に問い合わせることができる検索システムを実装した。5.2節で示したように、ユーザとのインタラクションを用いることにより、システムの誤認識を訂正することは可能である。今後は、この訂正機能をうまく活用できるかどうか、長期間の実シーンに対して実験を行い、有効性や問題点を検討する必要がある。

参考文献

- [1] R. Collins, A. Lipton, T. Kanade, H. Fu-jiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa: "A system for video surveillance and monitoring: VSAM nal report", In Technical report CMU-RI-TR-00-12, Robotics Institute, CMU, May 2000.
- [2] 橋本秀紀, 新妻実保子, 佐々木毅: 空間知能化—インテリジェント・スペース—, 日本ロボット学会誌, Vol.23 No.06, pp.34-37, 2005.
- [3] 市川 徹, 山澤 一誠, 竹村 治雄, 横矢 直和: 高解像度全方位ビデオカメラを用いた遠隔監視システムにおけるイベント検出, 電子情報通信学会 技術研究報告, PRMU2000-213, pp.87-94, 2001.
- [4] 上岡隆宏, 河村竜幸, 河野恭之, 木戸出正継: I'm Here!: 物探しを効率化するウェアラブルシステム, ヒューマンインタフェース学会論文誌, Vol.6, No.3, pp.19-30, 2004.
- [5] 鈴木健一郎, 和田正樹, 梅田和昇: インテリジェントルームにおける家電機器操作の高度化, 日本機械学会ロボティクス・メカトロニクス講演会'06, 2P1-E21, 2006.
- [6] 藤吉弘亘, 榎本暢芳, 長谷川修, 金出武雄: "アクティビティモニタリング. 屋外監視映像の要約と WWW 上表示・検索システム.", 第 7 回画像センシングシンポジウム論文集, pp. 423.428(2001-6).
- [7] 松井康作, 浜田玲子, 井手一郎, 坂井修一: 監視映像におけるオブジェクト移動履歴検索, 第 67 回情報処理学会全国大会講演論文集, Vol.3, pp.79-80, 2005.
- [8] 榎原靖, 白井良明, 島田伸敬: 対話を用いた物体認識のための照明変化への適応, 電子情報通信学会論文誌 D-II, Vol.J87-D-II, No.2, pp.629-638, 2004.
- [9] 滝澤正夫, 榎原靖, 白井良明, 島田伸敬, 三浦純: サービスロボットのための対話システム, システム制御情報学会論文誌, Vol.16, No.4, pp.24-32, 2003.
- [10] 佐治禎基, 上野敦志, 武田英明: 移動のある物体の認識・管理を行うオフィスロボットの構築, 電子情報通信学会人工知能と知識処理研究会, 知能ソフトウェア工学研究会, 2000.
- [11] 小倉英樹, 片山憲昭, 島田伸敬, 白井良明: 複数の認識エンジンを併用したビデオ操作支援システム, 電子情報通信学会総合大会 A-19-5, 2006.3.
- [12] 片山憲昭, 島田伸敬, 白井良明: 部屋内シーン変遷の自動管理システム, ViEW2006 -ビジョン技術の実利用ワークショップ, 2006.
- [13] Daniel Grest, Jan-Michael Frahm and Reinhard Koch: "A Color Similarity Measure for Robust Shadow Removal in Real-Time", VMV (Vision, Modeling and Visualization), 2003
- [14] Hironobu FUJIYOSHI, Takeo KANADE: "Layered Detection for Multiple Overlapping Objects", IE-ICE TRANSACTIONS on Information and Systems Vol.E87-D No.12 pp.2821-2827, 2004.12
- [15] 大語彙連続音声認識システム Julius: <http://julius.sourceforge.jp/>

牧和宏: 平成 19 年立命館大学理工学部情報学科卒。現在, 同大学院理工学研究科博士前期課程在学中。

島田伸敬: 平成 7 年大阪大大学院工学研究科博士後期課程了。博士 (工学)。同年同研究科助手。平成 13 年同助教授を経て, 平成 16 年より立命館大学情報理工学部准教授。電子情報通信学会, 人工知能学会, IEEE 各会員

白井良明: 昭和 44 年東京大大学院工学系博士課程了。工学博士。同年電子技術総合研究所入所。昭和 63 年 4 月大阪大学工学部教授。平成 17 年 4 月立命館大学情報理工学部教授。IEEE, 電子情報通信学会, 人工知能学会, 日本ロボット学会などの会員。