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-

Activity Monitornig

- Web-page Data Summarization System for Video Surveillance -

藤吉　弘亘 [1]　　榎本　暢芳 [2]　　長谷川　修 [3]　　金出　武雄 [4]
Hironobu Fujiyoshi　Nobuyoshi Enomoto　Osamu Hasegawa　Takeo Kanade

[1] 中部大学 工学部情報工学科
College of Engineering, Chubu University

[2] 株式会社東芝 e-ソリューション社
e-Solutions Company, Toshiba Corporation

[3] 産業技術総合研究所 脳神経情報部門
Neuroscience Research Institute, AIST

[4] カーネギーメロン大学ロボティクス工学研究所
The Robotics Institute, Carnegie Mellon University

EMAIL: hf@cs.chubu.ac.jp

Abstract　本論文では，画像理解技術を用いた自動ビデオ監視システムの新しい一形態として，屋外の監視場所で起こる人と自動車の行動 (アクティビティ) を認識し，その結果と対象物体の情報を監視者に提供する WWW を用いた表示・検索システムについて提案する．本システムの処理は，対象物の検出・識別・アクティビティ認識からなる．最初に，システムは侵入物体を検出する．次に，その物体を人もしくは自動車に識別し，色や形状等の詳細情報を得る．最後に，識別結果と移動軌跡により “人間が自動車から降りた” 等のアクティビティを認識し，その結果を WEB サーバ上のデータベースに保存する．監視者であるユーザは，WWW を介してサーバ上のデータベースを参照することで，監視場所で起こったアクティビティの要約を確認することができる．CMU のキャンパスにテストシステムを構築し，実環境下におけるタスクを行い本システムの有効性を確認した．

1　はじめに

近年，犯罪発生率の急増に伴い，ビデオ監視システムに関する研究への期待が高まっている．従来の重要施設の入退出管理を目的としたビデオ監視システムは，監視カメラ映像を記録するものや，監視員が複数のカメラ映像を同時にモニタリングするものが多い．監視する範囲が広く 24 時間監視となると監視員への負担が大きくなり，問題とされている．これに対し，米国では DARPA(Defence Advanced Research Projects Agency) の下，画像理解技術を用いた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]．国内においても，分散協調視覚に関する研究が行われている [2]．これらのシステムは監視員の負担軽減と作業効率化に大きく貢献でき，新しいビデオ監視システムとして期待されている．

本論文では，画像理解技術を用いた自動ビデオ監視システムの新しい一形態として，屋外の監視場所で起こる人と自動車の行動 (アクティビティ) を認識し，その結果と対象物の情報を監視者に提供する WWW を用いた表示・検索システムについて提案する．本システムの処理は，物体検出・識別・アクティビティ認識からなる．最初に，システムは侵入物体を検出する．次に，その物体を人もしくは自動車に識別し，色や形状等の詳細情報を得る．最後に，識別結果と移動軌跡により “人間が自動車から降りた” 等のアクティビティを認識し，その結果を WEB サーバ上のデータベースに保存する．監視者であるユーザは，WWW を介してサーバ上のデータベースを参照することで，監視場

所で起こったアクティビティの要約を確認することができる．

以降，**2** ではアクティビティ認識のアルゴリズム，**3** ではテストシステムの概要，**4** ではシステムの動作例として WWW による表示と検索結果について述べる．

2　アクティビティ認識

人と自動車のアクティビティを認識するには，最初に監視領域への侵入物体を検出し，その種別を識別する必要がある．その後，識別結果とその移動軌跡による物体間のインタラクションによりアクティビティを認識する．

2.1　物体検出

侵入物体の検出には，検出すべき物体が存在しない背景画像を予め用意しておき，入力画像と背景画像の差分を計算する背景差分処理が多く用いられている [1]．人と自動車のアクティビティを認識するには，画像上の複数物体の重なりを検出する必要があるが，背景差分処理と領域クラスタリングを組み合わせた手法では重なった複数物体を１つの領域として検出する．

そこで，我々は複数物体の重なりを理解するレイヤー型検出法を提案した [3]．レイヤー型検出法は，ピクセル分析とリージョン分析の二つの処理からなる．ピクセル分析では，各ピクセルの輝度値の時間変化を観測し，その変化軌跡によりピクセルの状態を静もしくは動と判定する．ある時間幅の変化軌跡を用いることで，太陽光等の環境変化の影響を受けにくくなる．リージョン分析では，動とラベル化されたピクセル領域を移動物体と判定する．静とラベル化された領域は静止物体と判定し，背景上のレイヤーとして記憶する．一度停止した物体は，再び動き出すまでレイヤーとして登録されているため，レイヤー上を通過する移動物体を区別して検出することが可能となる．

図 1 に，停止した 2 台の自動車とその手前を移動する物体 (人) の検出例を示す．検出後，各物体はカルマンフィルタを拡張したトラッキングにより固有の ID が付けられる [1]．

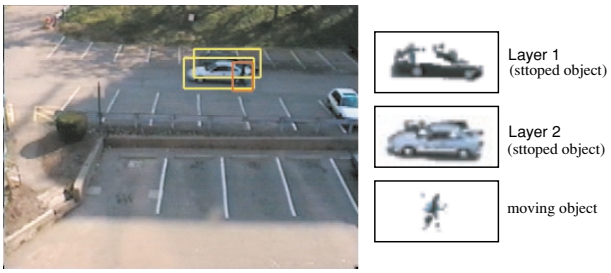


図 1: 物体検出例

2.2　車両と人の識別

次に，一映像中の車両と人物の識別とその色の推定，またユーザが定めた特定対象車両の検出を行う [4]．処理対象は屋外を移動するため，対象の「見え」は逐次変化し，不安定である．

2.2.1　種別（形状）の識別

まず学習用サンプル画像約 2000 枚に対し，オペレータが目視で種別 { セダン，バン，トラック，小型搬送車，人，その他 } のラベルをつけた (図 2 参照)．サンプル画像が予め定めた特定対象（実験では FedEx 社の集配車，郵便車，パトカーとした）である場合には，その旨のラベルも加えた．学習用の各画像からボックスの幅

図 2: 学習用サンプル画像例

と高さ (2 個)，濃淡画像部分の幅と高さ方向の 1,2,3 次モーメントの総計 (2 × 3=6 個)，濃淡画像部分の重心 (ボックス内の幅，高さ方向に 2 個)，濃淡画像部分の面積 (1 個) の 11 次元の特徴ベクトルを算出し，種別識別用の判別空間を構成する．種別の識別は，識別対象の画像の特徴ベクトルを種別識別用の判別空間に射影し，k 近傍法で判定して行った．

2.2.2 対象の色の推定

まず様々な条件下で撮影した色サンプル (晴天時の映像から約 1500，曇天時の映像から約 1000) にオペレータが自らの印象に基づいて 6 種類の色ラベルをつけ，それをシステムにそのまま学習させた．個々の色サンプル画像から 3 次元の色 [5] の特徴ベクトルを算出し，これをクラスとして線形判別空間を構成した．色の推定は，推定対象画像の色ベクトルをこの判別空間に射影し，k 近傍法で判定して行った．

2.2.3 最終判定と識別実験結果

物体検出は，入力映像中の対象に ID をつけてトラッキングする．そこで種別の識別では，種別の候補ごとに，トラッキングシーケンスを通じて最良の結果を求め，上位３つを順に種別の最終判定結果とした．色についても，各色の候補クラス毎にトラッキングシーケンスを通じて最良の結果を求め，上位３つを順に色の最終判定結果とした．図 3 は稼働中の本システムの

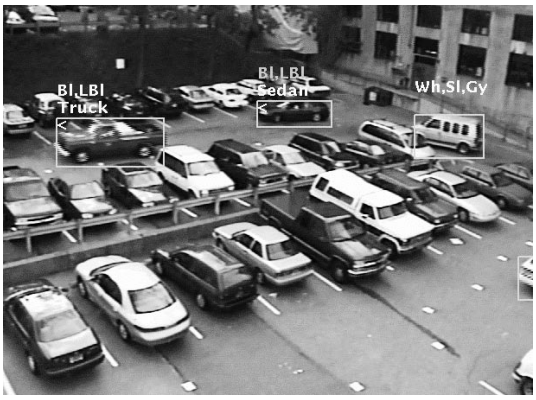


図 3: 識別処理中の出力画面例

処理画面例である．表 1 に，最終判定段階における性能評価実験の結果を示す．実験は，天候と太陽の位置の双方を考慮し，晴天／曇天の朝から日没の間の映像をオンラインでシステムに入力して行った．よって評価用入力映像にはいずれも学習サンプルはない．映像中に現れた処理対象は 180 個であった．実験では，システムによる種別と色の最終判定結果がともに目視による結果と一致した場合を正答とした．全正答率の平均は約 90% であった．この他に，本システムは特定対象車両の検出機能も備えている．

表 1: 識別実験結果

	H	S	V	T	M	O	To	E	%
H	67	0	0	0	0	7	74	7	91
S	0	33	2	0	0	0	35	2	94
V	0	1	24	0	0	0	25	1	96
T	0	2	1	12	0	0	15	3	80
M	0	0	0	0	15	1	16	1	94
O	0	2	0	0	0	13	15	2	87
								A	*90*

H:Human, S:Sedan, V:Van, T:Truck, M:Mule,
O:Others, To:Total, E:Error, A:Average

2.3 アクティビティ認識

ここでは，移動物体間の相互作用 (インタラクション) を含む行動 (アクティビティ) の推定について述べる．同様の研究として，Ivanov と Bobick[6] は，確率文脈自由文法 (SCFG) を用いて複数移動物体のインタラクションを推定したが，SCFG の遷移確率はマニュアルで設定していた．また Oliver らは CHMM[7] を用いた人物のインタラクション検出システムを開発したが，CHMM の学習にはシミュレータで発生させた合成データを用いた．これらに対し，我々のモデルでは各検出領域が内部状態として物体の種別，アクション，インタラクションの組を持つことを仮定し，それら内部状態の組同士の確率的関係に基づいてアクティビティを推定する．

ある領域（以下 Blob)$B^{(i)}$ と $B^{(j)}$ とが各フレームごとに検出され，フレーム間で追跡されるとすると，移動軌跡は Blob の系列 $B_0^{(i)},....B_{t-1}^{(i)},B_t^{(i)}$ と $B_0^{(j)},....B_{t-1}^{(j)},B_t^{(j)}$ とからなる．Blob 系列は，ある種別 $O^{(i)}$ と $O^{(j)}$ の対象物が，アクションの系列 $A_0^{(i)},....A_{t-1}^{(i)},A_t^{(i)}$ と $A_0^{(j)},....A_{t-1}^{(j)},A_t^{(j)}$ でインタラクションの系列 $I_0^{(i,j)},....I_{t-1}^{(i,j)},I_t^{(i,j)}$ にしたがって移動したものの観測量と考えられる．我々のゴールはこれら系列の組のうち，観測量としての Blob 系列を最尤に推定するものを求めることである．ここに各 Blob 系列に対し，時間 t での対象物種別 $O^{(i)},O^{(j)}$，Action 系列 $A_0^{(i)},...,A_t^{(i)},A_0^{(j)},...,A_t^{(j)}$，イ

ンタラクション系列 $I_0^{(i,j)}, ...I_{t-1}^{(i,j)}, I_t^{(i,j)}$ は，以下の条件付き確率を最大化するようなものとして推定できる．

$$\widehat{\left(S^{(i,j)}\right)}_0^t = argmax_{\left(S^{(i,j)}\right)_0^t} P\left(\left(S^{(i,j)}\right)_0^t \mid \left(B^{(i)}\right)_0^t, \left(B^{(j)}\right)_0^t\right) \quad (1)$$

ただし，

$$\begin{aligned}
&\left(S^{(i,j)}\right)_0^t \equiv \left\{O^{(i)}, O^{(j)}, \left(A^{(i)}\right)_0^t, \left(A^{(j)}\right)_0^t, \left(I^{(i,j)}\right)_0^t\right\}, \\
&\left(A^{(i)}\right)_0^t \equiv \left\{A_0^{(i)}, ...A_{t-1}^{(i)}, A_t^{(i)}\right\}, \\
&\left(I^{(i,j)}\right)_0^t \equiv \left\{I_0^{(i,j)}, ...I_{t-1}^{(i,j)}, I_t^{(i,j)}\right\}, \\
&\left(B^{(i)}\right)_0^t \equiv \left\{B_0^{(i)}, ...B_{t-1}^{(i)}, B_t^{(i)}\right\}.
\end{aligned}$$

本モデルにおける各内部状態の系列がマルコフモデルに従うという仮定を導入すると，式 1 の条件付き確率は一般に以下のように書ける．

$$\begin{aligned}
&P\left(O^{(i)}, O^{(j)}, \left(A^{(i)}\right)_0^t, \left(A^{(j)}\right)_0^t, \left(I^{(i,j)}\right)_0^t \mid \left(B^{(i)}\right)_0^t, \left(B^{(j)}\right)_0^t\right) \\
&\simeq \frac{P\left(B_t^{(i)} \mid O^{(i)}, A_t^{(i)}, I_t^{(i,j)}\right) P\left(B_t^{(j)} \mid O^{(j)}, A_t^{(j)}, I_t^{(i,j)}\right)}{P\left(B_t^{(i)}, B_t^{(j)}\right)} \\
&\quad \bullet P\left(I_t^{(i,j)} \mid O^{(i)}, O^{(j)}, A_t^{(i)}, A_t^{(j)}\right) \\
&\quad \bullet P\left(A_t^{(i)} \mid O^{(i)}, A_{t-1}^{(i)}, I_{t-1}^{(i,j)}\right) \\
&\quad \bullet P\left(A_t^{(j)} \mid O^{(j)}, A_{t-1}^{(j)}, I_{t-1}^{(i,j)}\right) \\
&\quad \bullet P(O^{(i)}, O^{(j)}, \left(A^{(i)}\right)_0^{t-1}, \left(A^{(j)}\right)_0^{t-1}, \left(I^{(i,j)}\right)_0^{t-1} \\
&\qquad \mid \left(B^{(i)}\right)_0^{t-1}, \left(B^{(j)}\right)_0^{t-1})
\end{aligned} \quad (2)$$

式 2 を計算するためには，条件付き確率 $P\left(B_t^{(i)} \mid O^{(i)}, A_t^{(i)}, I_t^{(i,j)}\right)$，$P\left(A_t^{(i)} \mid O^{(i)}, A_{t-1}^{(i)}, I_{t-1}^{(i,j)}\right)$，$P\left(I_t^{(i,j)} \mid O^{(i)}, O^{(j)}, A_t^{(i)}, A_t^{(j)}\right)$，および同時確率 $P\left(O^{(i)}, O^{(j)}, A_0^{(i)}, A_0^{(j)}, I_0^{(i,j)}\right)$ があらかじめ必要であるが，これらは監視場所のサンプル映像を用いて対象物タイプ，アクション，インタラクションの組の出現頻度を計測することで求められる．

認識時には，図 4 に示すトレリスを画面内の 2 つずつの Blob の組について生成する．そしてトレリス上の，ある時刻 t の，ある状態 Sk にたどりつくパスのうち式 2 の確率が最大になるものを選択し，最尤なイベント系列とする．図中で Sk は確率遷移の内部状態を表し，Blob-i のアクション，Blob-j のアクション，i と j とのインタラクションの組からなる．

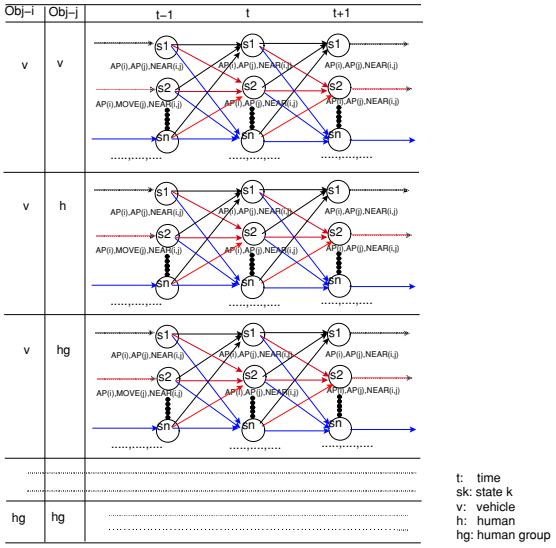


図 4: トレリス図

3 システム構成

CMU のキャンパスにテストシステムを構築した．本テストシステムは，屋外に設置してある監視カメラ，画像処理部，データベースと WEB サーバで構成される (図 5 参照)．画像処理部は，検出 (追跡)･識別･アクティビティ認識からなる．システムは，物体検出後その識別結果と画像チップを WEB サーバ上のデータベースに登録する．ここまでの処理は，PC(Pentium III-500MHz 4CPUs) にインプリメントされ約 10[fps] で動作している．同様に，アクティビティ認識の結果もデータベースに登録される．WEB サーバは，ユーザ (WEB ブラウザ) の要求に対して CGI(Common Gateway Interface) によりデータベースからアクティビティの要約結果を自動生成し表示する．

4 WWW による表示・検索結果

WWW によるアクティビティの要約結果の表示例を図 6 に示す．(a) は，アクティビティレポートの例である．アクティビティレポートは，“A Human got out of a Vehicle”, “A Vehicle parked” 等のイベント結果を時刻順に表示する．もし，ユーザがアクティビティに関与した物体の詳細を知りたいとき，物体の識別タイプや色

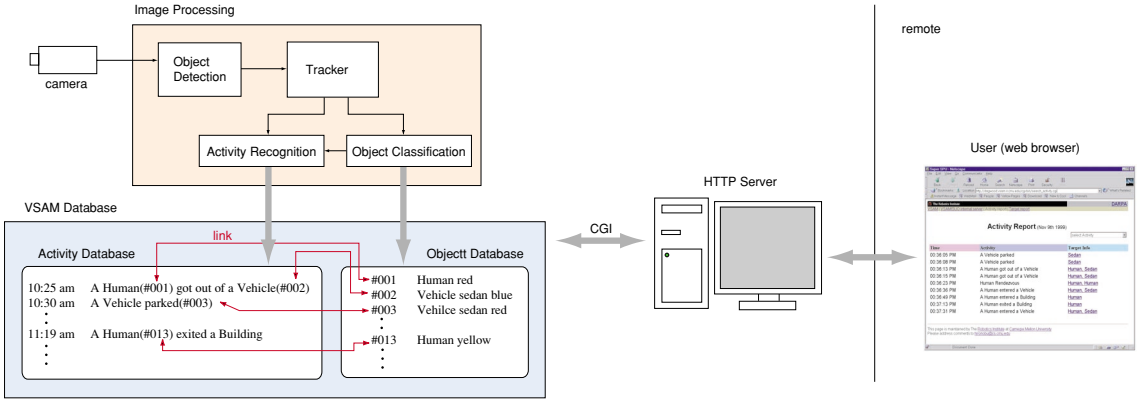


図 5: システム構成

等の情報とその画像チップをハイパーテキストリンクにより表示することができる．(b) はオブジェクトリポートの例である．オブジェクトリポートは，システムによって検出された全ての物体の一覧を表示する．観測された物体数が多いとき，ユーザは全ての物体を確認することが不可能であるため，識別タイプごとに表示させることができる．また，ユーザがある特定の物体をマウスクリックにより指定すると，システムはデータベースに登録された他の物体との類似度を識別タイプと色情報から計算し，類似度の高い順に表示する．これにより，監視領域内の別の場所や異なる時間帯に観測された同一対象物体の行動を知ることができる．

5 まとめ

本論文では，監視領域における人と自動車のアクティビティを認識し，その結果を WWW を用いて表示・検索するシステムについて提案した．システムの特徴としては，監視領域内で検出された物体の識別情報とアクティビティの要約を知ることができる．本システムは，テロ防止対策や交通量調査等への応用が考えられる．

謝辞 本システムの開発に助言して頂いた Dr. Robert T. Collins, Dr. Alan J. Lipton を始めとする VSAM プロジェクトのメンバーに感謝致します．

参考文献

[1] R. Collins, et al.: “A system for video surveillance and monitoring: VSAM final report”, *Technical report CMU-RI-TR-00-12, Robotics Institute, CMU*, May 2000.

[2] T. Matsuyama: “Cooperative distributed vision”, *Proc. of DARPA Image Understanding Workshop*, volume 1, pp. 365–384, November 1998.

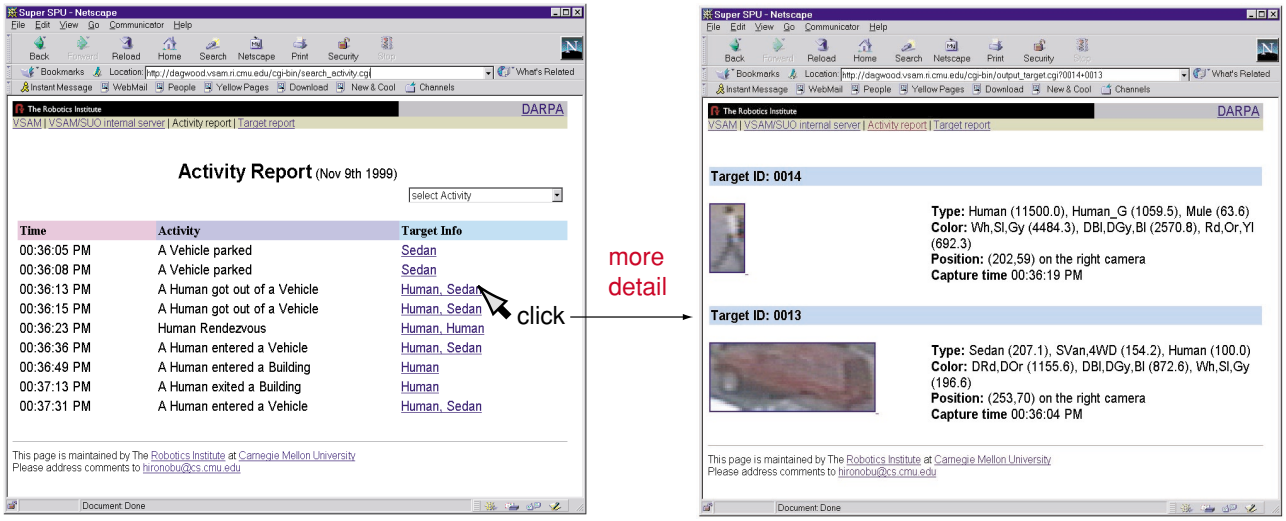
[3] 藤吉弘亘, 金出武雄: “複数物体の重なりを理解するレイヤー型検出法”, 第 7 回画像センシングシンポジウム論文集, 2001.

[4] 長谷川修, 金出武雄: “一般道路映像中の対象物のオンライン識別”, 第 7 回画像センシングシンポジウム論文集, 2001.

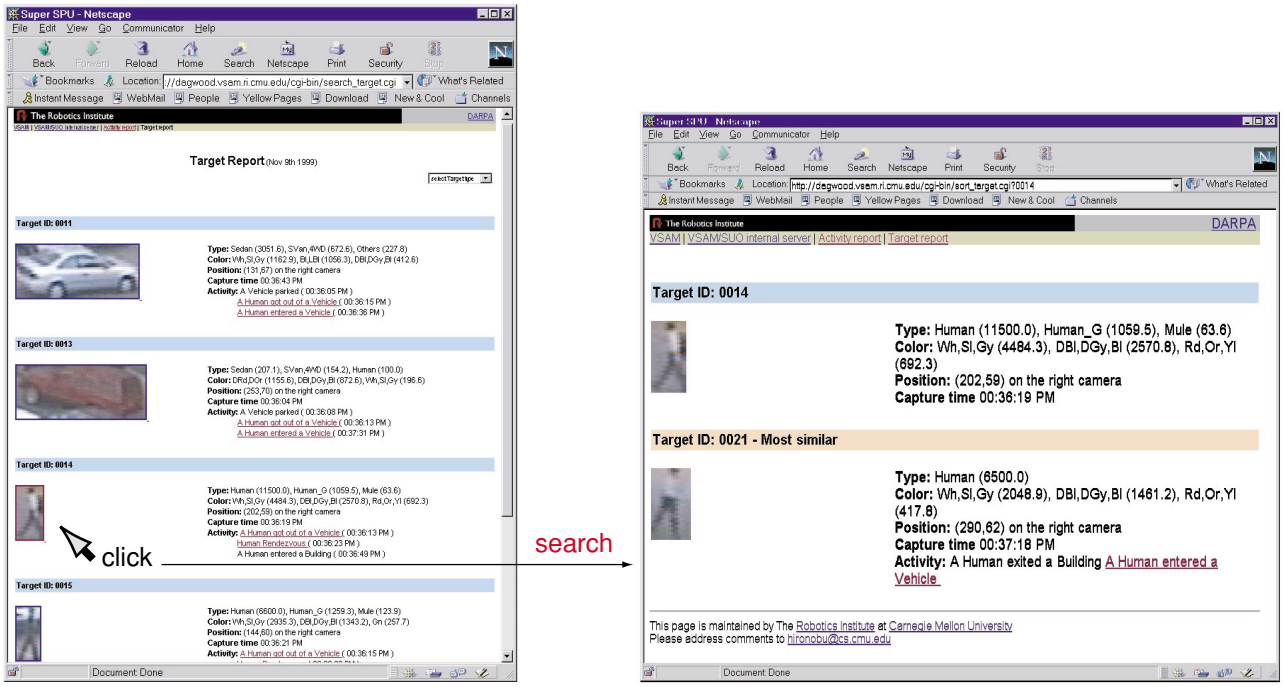
[5] Y.Ohta et al.: “Color Information for Region Segmentation”, *Computer Graphics and Image Processing*, vol.13, no.3, pp.222-241, 1980.

[6] Y.A. Ivanov, et al.: “Recognition of visual activities and interactions by stochastic parsing”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):852–872, August 2000.

[7] N. Oliver, et al.: “A Bayesian Computer Vision System for Modeling Human Interactions”, *Proc. of ICVS’99* , Gran Canaria, Spain, Jan 1999.



(a) Activity report for web page.



(b) Object report for web page.

図 6: WWW による表示と検索例