グラフィクスとCAD 105−13
(2001. 11. 9)

VSAM：画像理解技術を用いたビデオ監視システムプロジェクトについて

藤吉 弘亘[†], 金出 武雄[††]

[†] 中部大学工学部情報工学科,
[††] カーネギーメロン大学ロボット工学研究所

あらまし: 米国カーネギーメロン大学 (CMU) で行われているビデオ監視システムに関する研究プロジェクト VSAM (Video Surveillance and Monitoring) は，複数のカメラを用いた屋外監視システムの構築を目指している．カメラセンサは，動画像理解技術により屋外カメラ映像から実時間で物体の検出・追跡・識別を行い，ネットワークを介して物体情報を監視センタに送信する．各カメラセンサからの情報はサイトモデルと呼ばれる共通の 3D データ空間で統合され，マップ上に表示される．これにより，ユーザは，人や自動車がどこを移動しているか等の状況を実時間でモニタすることが可能となる．ユーザがシステムにタスクを与えると，複数のカメラセンサは協調してトラッキングし，一台のカメラでは許容できない広範囲における侵入物体の行動軌跡を知ることが可能となる．本稿では，VSAM プロジェクトの概要と動画像理解技術について述べ，評価実験により本システムの有用性を示す．

VSAM : Video Surveillance and Monitoring Project

Hironobu FUJIYOSHI[†], Takeo KANADE[††]

[†]Department of Computer Science, College of Engineering, Chubu University,
[††]The Robotics Institute, Carnegie Mellon University

Abstract: The Video Surveillance and Monitoring (VSAM) team at Carnegie Mellon University (CMU) has developed an end-to-end, multi-camera surveillance system that allows a single human operator to monitor activities in a cluttered environment using a distributed network of active video sensors. Video understanding algorithms have been developed to automatically detect people and vehicles, seamlessly track them using a network of cooperating active sensors, determine their 3D locations with respect to a geospatial site model, and present this information to a human operator who controls the system through a graphical user interface. The goal is to automatically collect and disseminate real-time information to improve the situational awareness of security providers and decision makers. The feasibility of real-time video surveillance has been demonstrated within a multi-camera testbed system developed on the campus of CMU. This paper presents an overview of the issues and algorithms involved in creating this semi-autonomous, multi-camera surveillance system.

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1997年より2000年の3年間，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) が行われた [1]．カーネギーメロン大学 (CMU) では，キャンパスに12台のカメラを配置しテストシステムを構築した [2]．システムは，動画像理解技術により検出した侵入物体を複数のカメラが協調してトラッキングし，その状況をリアルタイムで監視員に提示する．これにより，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．

本稿では，画像理解技術を用いた自動ビデオ監視システムとして，VSAMプロジェクトの概要，動画像理解技術，監視結果の表示方法を中心に述べる．

2 システムの構成とその特徴

VSAMシステムの特徴とその構成について述べる．

2.1 分散協調による状況理解

1台のカメラセンサで監視できる範囲は限られているので，実時間で広域監視を行うためには，複数のカメラセンサを効果的に配置し，ユーザ（監視員）からのタスクをネットワークに接続された複数のカメラセンサが協調して動作する必要がある．VSAMプロジェクトでは，このようなシステムの実現を目指している．

VSAMシステムの概要を図1に示す．個々のセンサは，屋外カメラからの映像より実時間で物体の検出・追跡・識別を行い，自動検出した物体情報を監視センタに送信する．各カメラセンサからの情報はサイトモデルと呼ばれる共通の3Dデータ空間で統合され，マップ上に表示される．ユーザは，人や自動車がどこを移動しているか等の広域の動的状況を実時間でモニタすることが可能となる．ユーザがシステムにタスクを与えると，複数のカメラセンサは協調して動作するため，1台のカメラでは許容できない広範囲における侵入物体のトラッキングを可能とし，その行動軌跡を知ることができる．さらに，コンピュータグラフィックス (CG) を用いて合成映像を生成することにより，監視領域で実際に起った動的シーンを仮想的に再現することが可能となる．

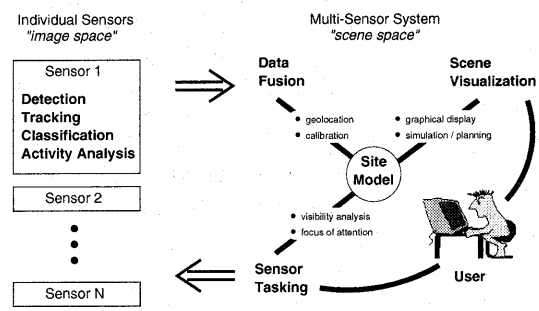


図 1: システムの概要

2.2 サイトモデル

VSAMシステムの特徴の一つは，サイトモデル (3D地形データ) の使用である．サイトモデルを用いることで，単眼カメラでも対象物体の三次元位置の推定が可能となる．米国における地形データは，USGS (U.S.Geological Survey) からDEM (Digital Elevation Map) データを購入することができるが，その解像度は1ピクセル $= 30m^2$ と低い．そこで，VSAMプロジェクトではCMUのキャンパスと周辺の地形データに建物や道路等の情報を加えたより高解像度のDEMデータを作成した (図2参照)．これをサイトモデルと呼ぶ．また，GPSを用いて測定した複数のランドマークより各カメラの位置を計算した [3]．図2(c) は，実際のカメラビューをサイトモデルと計算したカメラ位置から合成したものである．

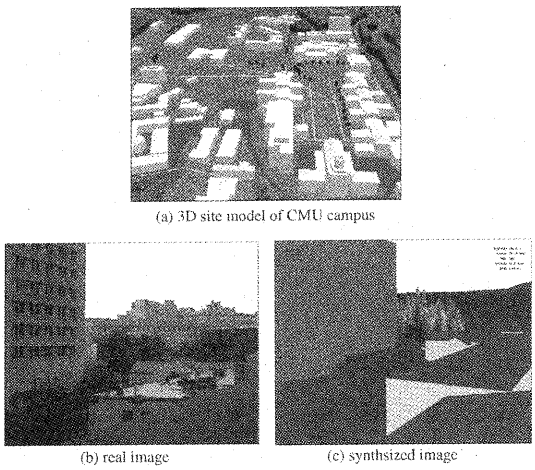


図 2: サイトモデル

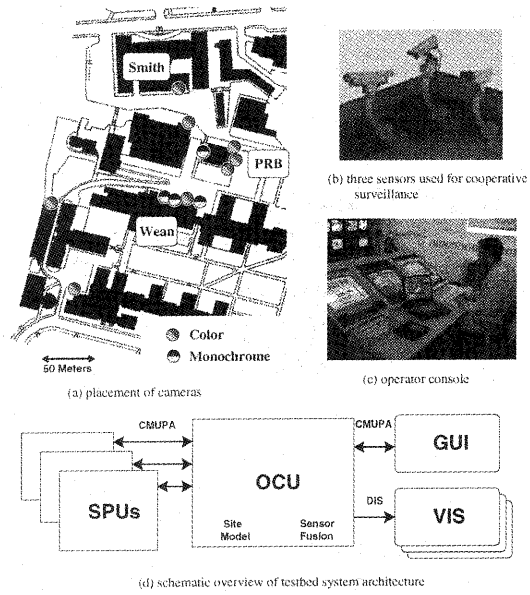


図 3: システムの構成

2.3 システムの構成

CMU キャンパスの約 $0.4km^2$ の領域を監視対象とし，12 台のカメラを建物の屋上や壁面に死角領域が少なくなるように配置した (図 3(a) 参照)．テストシステムは，図 3(d) に示すように，SPUs (Sensor Processing Units), OCU (Operator Contorol Unit), GUI (map-based Graphical User Interface), VIS (Visualization nodes) から構成される．

SPU は，pan/tilt/zoom 機能を持つアクティブカメラと動画像処理を行う PC で構成されたセンサである．SPU はカメラ映像から自動的に物体を検出・識別し，その結果を記号データとしてネットワークを介して OCU に送信する．

OCU は全ての SPU からの情報を受け取りデータベースに登録し，GUI はマップ上にその結果を表示する．

3 動画像理解技術

動画像理解技術として，物体検出，識別，アクティビティ認識のアルゴリズムと評価実験結果について述べる．

3.1 物体検出・追跡

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 [4]．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一つの領域として検出する．

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 [5]．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 4 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．比較的交通量が多い屋外駐車場の 2 箇所を視野とするカメラで撮影した約 8 時間のビデオ映像を用いて，本検出法の評価実験を行った結果，背景差

分法は約 83%，レイヤー型検出法は 92%の検出率を得た．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 [2].

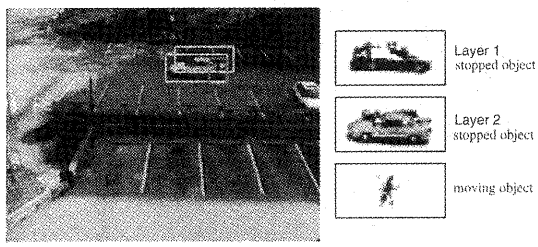


図 4: 物体検出例

3.2 物体識別

検出した物体は，以下の要因によりその「見え」は逐次変化し不安定であるため，人と自動車の識別は容易でない．

- 屋外環境を対象物体が移動
- 天候による照明変化
- カメラの位置

VSAM システムでは，12 台のカメラを見えの大きく異なるカメラ毎にグループ化し，それぞれの識別器を作成することで対処した．識別にはニューラルネットワークを用いた．また，物体の詳細情報として対象の色，車種 (形) を識別するために判別空間による手法を用いた．

3.2.1 ニューラルネットワークを用いた識別

見えの大きく変わるカメラ毎に約 1000 枚の学習サンプルを作成した．検出した画像から計算した形状の複雑度 (周囲長^2/面積)，面積，縦横比，カメラのズーム倍率をニューラルネットワークへの入力パラメータとした．出力は，人 (一人), 人のグループ (二人以上), 自動車, その他の 4 クラスとした．ニューラルネットワークは 3 層の階層型で，学習にはバックプロパゲーション法を用いた．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，トラッキングシーケンスを通じて各フレーム毎に出力されたニューラルネットワークの結果から各クラスのヒストグラムを作成し，最大となったクラスを最終識別結果とした．小規模なニューラルネットワーク (入力層 4, 中間層 16, 出力層 4) で構成しているため，リアルタイムでの処理を実現している．

3.2.2 種別 (形状) の識別

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．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その長方領域の幅と高さ (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 ($2 \times 3 = 6$ 個)，濃淡画像部分の重心 (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 [6].

3.2.3 対象物の色の推定

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 [7] 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．

図 5 は稼働中の本システムの処理画面例である．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により評価を行った結果，種別と色の平均識別率は約 90%であった．

3.3 アクティビティ認識

監視領域で起った事象を知るためには，人と自動車のインタラクションを認識することが必要となる．このような行動認識は，テロ防止対策や駐車場管理等への応用が考えられる．

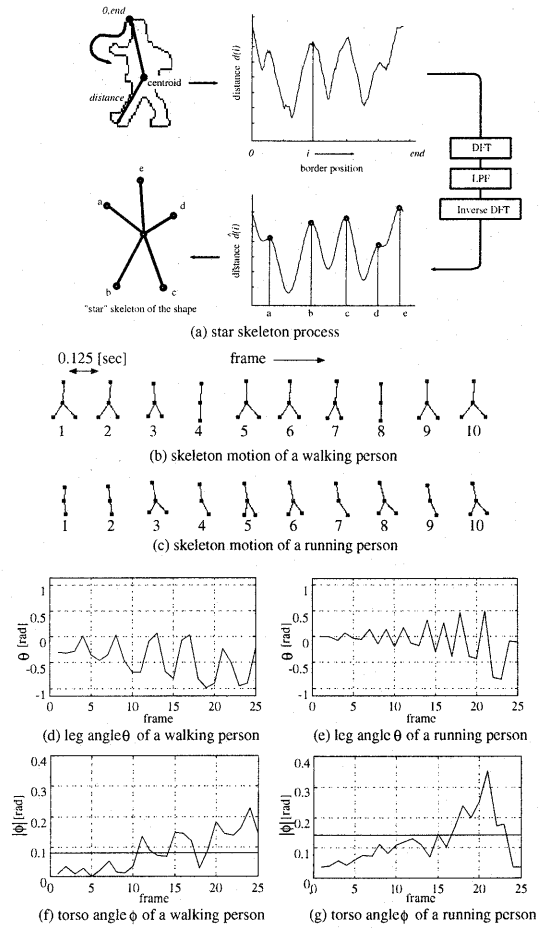
図 5: 物体識別例

3.3.1 人の動き分析

人の動きを認識するには，人体の各部位を追跡する手法が多く提案されている [8]．我々は，リアルタイムで検出した人の画像領域から，スター型スケルトンを抽出する手法を提案した [9]．スター型スケルトンは，検出領域の重心と輪郭線上の自動的に抽出された特徴点で構成される (図 6 参照)．スケルトン形状から頭部と足部の特徴点を決定し，姿勢の前傾度合いと足部の時間的変化を周波数分析することにより，人間が歩いているか走っているかを識別することが可能となる．

3.3.2 人と自動車の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) を推定する．物体検出·識別によって得られた情報を基に，各検出領域が内部状態として物体の種別，アクション (Appearing, Moving, Stopped, Disappearing)，インタラクション (Near, MovingAwayFrom, MovingTowards, NoInteraction) 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．詳細については，文献 [10] を参照いただきたい．これにより，以下の 6 種類のアクティビティを認識することが可能となる．これらの結果の表示方法は **5.3** に述べる．



(a) star skeleton process

(b) skeleton motion of a walking person

(c) skeleton motion of a running person

(d) leg angle θ of a walking person

(e) leg angle θ of a running person

(f) torso angle φ of a walking person

(g) torso angle φ of a running person

図 6: スター型スケルトンによる動きの分析

(1) human entered a vehicle.
(2) human exited a vehicle.
(3) human entered a building.
(4) human exited a building.
(5) a vehicle parked.
(6) human rendevouz.

4 対象物体の 3 次元位置の推定

検出された対象物体の 3 次元位置の推定は，通常複数のカメラを用いたステレオ手法が多く用いられるが，監視領域が広く物体数が多い場合，常に一つの物体を 2 台以上のカメラセンサでトラッキングできるとは限らない．VSAM では，監視領域の地形データであるサイトモデ

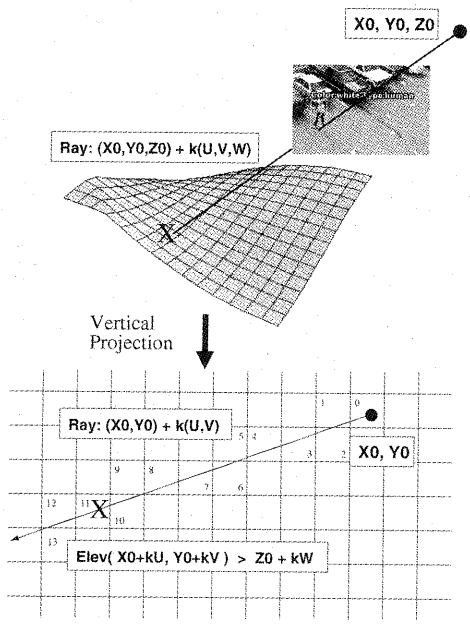


図 7: 位置の推定

ルを用意しているため，対象物体 (人や自動車) が地面上に接しているという仮定により，画像上の検出直方領域の下辺部の中心点を通過する 3 次元空間上の光線が地形データ (DEM) の地面と交差する点を対象物の 3 次元位置と推定できる (図 7 参照)．Leica 社製のレーザトラッカーセオドライトを用いて，本手法の位置推定精度の評価を行った．評価実験は，駐車場を二周したときのセオドライトによる測量座標とカメラセンサにより光線交差法を用いて自動推定した座標間の距離を測定した．実験の結果，カメラから対象までの距離が約 65m のとき，平均誤差が 0.6m と良好な精度を得た．

5 動的シーンの表示

監視システムでは，リアルタイムに情報をユーザに伝える必要がある．また，ユーザが一度に複数のモニタを監視することは難しいため，複数の情報源からの情報を一括して提示する必要がある．ここでは，動画像理解技術により得られた情報 (識別クラス, 3D 位置座標) をユーザに提示する手法として 2 次元平面地図を用いたマップ型 GUI, CG による 3D 動的シーン表示，WWW を用いたアクティビティの要約表示について述べる．

5.1 マップ型 GUI

SPU から集められた対象物体の情報を 2 次元平面地図の上にリアルタイムで表示するマップ型 GUI を開発した (図 8 参照)．マップ上には，現在の全ての対象物の位置に物体を示す記号 (人は丸，自動車は四角) とカメラの状態として位置と FOV (Field of view) が表示される．ユーザがマウスによりマップ上のカメラをクリックすると，そのライブカメラ映像をモニタすることができる．マップ型 GUI は広域に配置された複数のカメラからの情報を一括して表示するため，監視領域の状況の把握が容易となる．

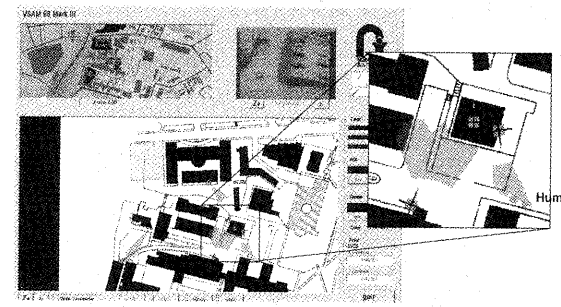


図 8: マップ型 GUI

5.2 CG による動的シーンの 3D 表示

監視結果は全てデータベースに登録されているため，それらのデータとサイトモデルにより CG を用いて動的な合成映像を自動生成しユーザに提示した．CG を用いることで実際に起った動的シーンを，任意の視点からの映像として再現することができる．交通事故等を様々な角度から再現することで事故の検証に有効であると考えられる．その他の応用例としては，サッカー等のスポーツにおけるゲームシーンの解析と再現が挙げられる．

5.3 WWW によるアクティビティの要約表示

人と自動車のアクティビティと識別結果は，全てデータベースに登録されている．ユーザ

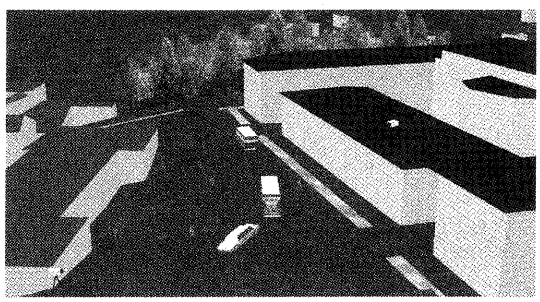


図 9: CG による動的シーンの再現

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．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 [11].

WWW によるアクティビティの要約結果の表示例を図 10 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等の情報とその画像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，識別タイプ毎に表示することができる．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長時間の行動を推定することができる．

6 まとめ

本稿では，複数のカメラを用いたビデオ監視システムのプロジェクト VSAM について，動画像理解技術と監視結果の表示方法を中心に述べた．VSAM の特徴は，動画像理解技術とサイトモデルを組み合わせることにより広域の実時間監視が可能となる点である．VSAM プロジェクトの技術課題や本稿では省略した複数のセンサによる協調動作等については，参考文献や Web site(http://www.cs.cmu.edu/~vsam/) を参照いただきたい．

謝辞 CMU VSAM Project: Robert T. Collins, Alan J. Lipton, David Duggins, Osamu Hasegawa, Nobuyoshi Enomoto の諸氏に深く感謝する．

参考文献

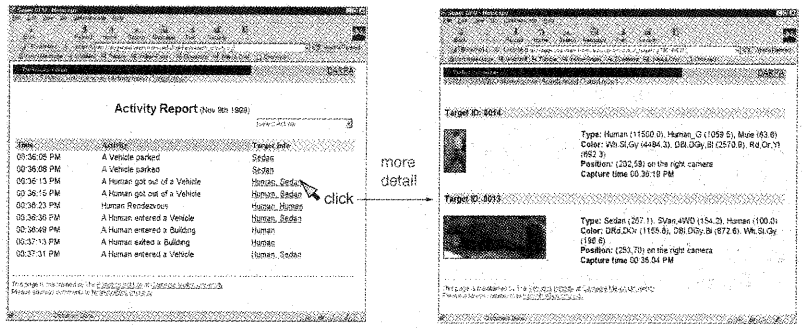
[1] VSAM: “Section I, video surveillance and monitoring”, Proc. of the 1998 DARPA Image Understanding Workshop, Vol.1, pp. 1–400 (Nov. 1998).

[2] R. Collins, A. Lipton, H. Fujiyoshi and T. Kanade: “Algorithms for cooperative multi-sensor surveillance”, Proc. of IEEE Special Issue on ”Video Communications, Processing and Understanding for Third Generation Surveillance Systems”, (Oct. 2001).

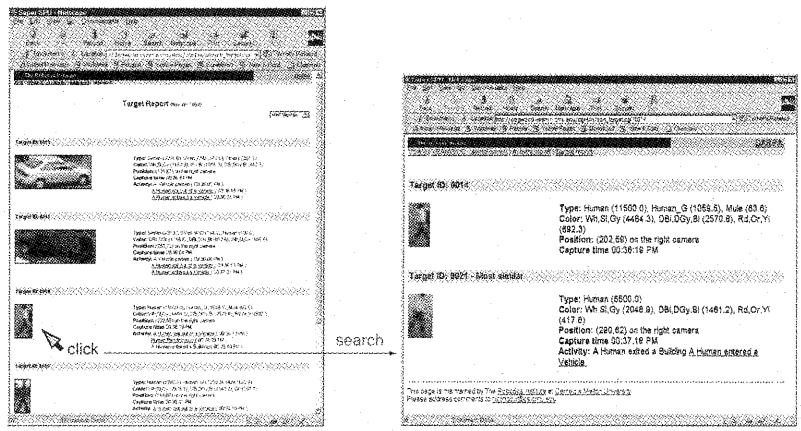
[3] R. Collins and Y. Tsin: “Calibration of an outdoor active camera system”, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 528–534 (Jun. 1999).

[4] A. Lipton, H. Fujiyoshi and R. Patil: “Moving target detection and classification from real-time video”, Proc. of the 1998 Workshop on Applications of Computer Vision, pp. 8–14 (Oct. 1998).

[5] 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, pp. 369–374 (2001-6).



(a) Activity report



(b) Object report

図 10: WWW による表示と検索例

[6] 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, pp. 221–226 (2001-6).

[7] Y. Ohta and T. Kanade: "Color information for region segmentation", Computer Graphics and Image Processing, Vol. 13, No. 3, pp. 222–241 (1980).

[8] J.K. Aggarwal and Q. Cai. "Human motion analysis: A review", Computer Vision and Image Understanding, Vol. 73 No. 3, pp. 428–440 (Mar. 1999).

[9] H. Fujiyoshi and A. Lipton: "Real-time human motion analysis by image skeletonization", Proc. of the 1998 Workshop on Applications of Computer Vision, pp. 14–21 (Oct. 1998).

[10] N. Enomoto, T. Kanade, H. Fujiyoshi and O. Hasegawa. "A method for monitoring activities of multiple objects by using stochastic model", IEICE Transactions on Information and Systems, Vol. 84-D No. 12 (Dec. 2001).

[11] 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).