# 熱画像と距離カメラのマルチモーダル情報統合に基づく動的頭部検出システムの検討

○中村圭佑, 中臺一博，中島弘史，Ince Gokhan
（(株）ホンダ・リサーチ・インスティチュート・ジャパン）

# Head Motion Detection System by Integrating Infrared Thermal Sensor and Distance Camera and its Evaluation

*Keisuke Nakamura, Kazuhiro Nakadai, Hirofumi Nakajima, Gokhan Ince
(Honda Research Institute Japan Co., Ltd.)

**Abstract**— This paper addresses a newly developed human-head detection system utilizing two different modalities, that is, an infrared thermal sensor and a Time-of-Flight (ToF) distance camera. A lot of previous work on human detection has been forcusing mainly on visual sensors, thus it has difficulties in dealing with environmental changes such as illumination. Our multimodal fusion approach is robust for such environmental change. Therefore, it suits for environmental scene analysis. The human detection algorithm is simple enough working in real time with maintaining the performance. It also has a function of human tracking by introducing a particle filter with a low computational cost. The experimental results with the multimodal approach successfully verified a novel performance improvement compared to single modality approach.
**Key Words:** Multi-modal integration, Human Detection, Particle Filter

## 1　序論

人・ロボットインタラクションの分野を中心として，視覚や聴覚等の異なるモダリティを統合し，ロバストに情報処理を行う研究が盛んに行われている [1]．人間は五感の情報を，状況に応じ，より情報量の多いものを選択的・統合的に処理し，様々なタスクを達成している．

我々はこれまで，音響信号処理のアプローチから，聴きたい音源への選択的注意を興味指数として具現化し，選択的に処理する手法 [2] を提案しており，[3] では，人間の音声のみに傾聴する統合システムを構築してきた．しかし，音響信号処理のみで人への選択的注意を実現する場合，音声が入力されなければ，人の検出ができないため，[3] では常に音声が入力されることを暗に仮定している．人間が途切れなく会話する状況は自然な環境では仮定できないため，音響信号以外の新たなモダリティによって常に人を検出することで，正しい選択的注意が実現できると考えられる．

本論文の目的は，新たなモダリティとして，熱画像と3D 距離カメラのマルチモーダル情報を統合した人体検出システムを提案評価することにある．

これまでにも人体検出システムは数多く提案されてきた．歩行者検出に関しては [8] が詳しく，多くは，画像センサまたは，赤外線による熱画像を単体で使用した手法となっている．その他，レーザスキャナを用いた手法 [4]，超音波レーダを用いた手法 [5] 等が報告されている．このようなユニモーダルな人体検出の問題点として，動的に特性の変化する実環境に対してロバストに人体を検出することができないことが挙げられる．

マルチモーダル情報統合による物体検出として，一般的なカメラと熱画像カメラ [6]，3D 距離カメラとステレオカメラ [7] を組み合わせた例が報告されている．こうした手法は複数の情報を同時に用いることにより，検出のロバスト性を向上させているが，ここで用いられている一般的なカメラやステレオカメラは，物体のテクスチャが測定しやすく，物体同定などには有利な反面，明暗の変化や，光源色の変化等にそもそも弱い．本稿のように人体検出のみを目的とする場合は，人間全体の形状や温度という比較的個体差なく観測できる特徴をロバストに抽出するため、熱画像と 3D 距離カメラによって，個体差の小さい体温と形状を情報統合した実時間人体検出システムを提案する．本稿では，人体でも露出していることが前提とできるため温度が高く，衣服等の状況変化によらない頭部の検出を扱うこととする．3D 距離カメラと熱画像の情報を用いることで，一般的なカメラやステレオカメラよりも光源の変化に対してロバストに人体の 3 次元位置を同時に計測できることが期待できる．

本稿では，さらに，実環境で移動する頭部検出に対応するため，移動のモデル化において自由度の高いパーティクルフィルタ（以降 PF と略す）を実装する．また，熱画像と距離カメラを単体で用いた場合の性能と比較し，本手法の有効性を示す．

## 2　マルチモーダル情報統合に基づく頭部検出

本節では熱画像と 3D 距離カメラそれぞれの画像処理と，統合手法について記す．Fig.1 に処理の流れを示す．

### 2.1　熱画像における処理

まず，入力熱画像を，最高温度 $T_H$[deg] と最低温度 $T_L$[deg] によって，以下のように二値化する．

$$\bar{f}(x,y) = \begin{cases} 1\ , \text{ if } \ \mathcal{F}(T_L) \leq f(x,y) \leq \mathcal{F}(T_H) \\ 0\ , \text{ otherwise} \end{cases} \quad (1)$$

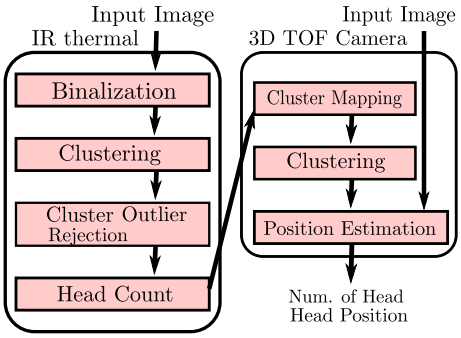
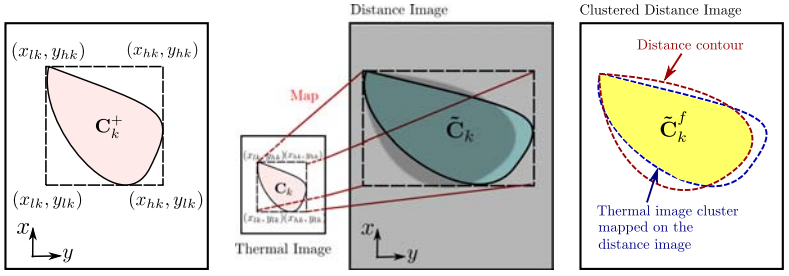



**Fig.**1 Flow of Image Processing




(a) Clustering on a thermal image (b) Cluster mapped on a distance image (c) Clustering on a thermal image

**Fig.**2 Image Processing for Head Detection

ここで，$\mathcal{F}(T)$ は検出対象物体の温度が $T$[deg] である時の，熱画像の輝度，$f(x,y)$ は画素インデックス $(x,y)$ の画素値を表す $(1 \leq x \leq X, 1 \leq y \leq Y$ とする$)$.

次に二値化画像 $\bar{f}(x,y)$ に対して階層的クラスタリングを行う．具体的には $\bar{f}(x,y)=1$ の画素に対し，上下左右に $N_{cls}$ 画素までを考えた正方画素領域 $\bar{f}(x+i, y+j)$ $(-N_{cls} \leq i \leq N_{cls}, -N_{cls} \leq j \leq N_{cls})$ を考え，$\bar{f}(x+i,y+j)=1$ なら同じクラスタとする（$N_{cls}$ は設計パラメータ）．この時，$k$ 番目のクラスタとみなされた $(x,y)$ の集合（$\mathbf{C}_k^+$）の最大値 $(x_{hk}, y_{hk})$ と最小値 $(x_{lk}, y_{lk})$ を 4 辺とする矩形ブロックを一つの矩形クラスタとする．矩形クラスタの例を Fig. 2(a) に示す．

次に，得られた矩形クラスタの中で不要クラスタを除外し 0 とする．不要クラスタの検出条件は以下とした．

- 矩形クラスタの画素数の上限 $((x_{hk}-x_{lk})(y_{hk}-y_{lk}) > R_{max})$．本稿で使用する 3D 距離カメラの最短計測距離よりも近くに頭が検出されている可能性が高い．
- 矩形クラスタの画素数の下限 $((x_{hk}-x_{lk})(y_{hk}-y_{lk}) < R_{min})$．本稿で使用する 3D 距離カメラの最長計測距離よりも遠くに頭が検出されている可能性が高い．またはノイズである可能性が高い．
- 矩形クラスタの縦横比 $((x_{hk}-x_{lk}) > \gamma(y_{hk}-y_{lk}))$．矩形クラスタの形状から明らかに頭部でない場合（本稿では評価実験において条件を満たしやすい横長クラスタの除外とした）．

ただし，$R_{max}, R_{min}, \gamma$ は設計パラメータ．ここで，除外後に残った矩形クラスタの範囲の中で値が 1 の画素の集合 $\mathbf{C}_k$ を以下で表す．

$$\mathbf{C}_k = \mathrm{arg}_{x,y}(\bar{f}(x,y)=1) \quad (2)$$
$$(x_{lk} \leq x \leq x_{hk}, y_{lk} \leq y \leq y_{hk}, k \in \mathbb{H})$$

ここで，$\mathbb{H}$ は残った矩形クラスタの集合である．

### 2.2 3D 距離カメラにおける処理

次に 3D 距離カメラでの処理について述べる．

まず，熱画像でのクラスタリングと不要クラスタの除外後に残った $\mathbf{C}_k$ を以下のように 3D 距離カメラの座標に写像する．

$$\tilde{\mathbf{C}}_k = \mathcal{G}(\mathbf{C}_k) \quad (3)$$

ここで，$\mathcal{G}$ は線形写像を表す．また，3D 距離カメラの画素 $(\tilde{x}, \tilde{y})$ での画素値を $\tilde{f}(\tilde{x}, \tilde{y})$ と表しておく．座標変換後の画像の例を Fig. 2(b) に示す．

ここで，$\bar{f}(x,y)=1$ $((x,y) \in \mathbf{C}_k)$ であっても，写像の誤差から，$\tilde{f}(x,y)$ $((x,y) \in \tilde{\mathbf{C}}_k)$ が必ずしも頭部と重なるとは限らないことに注目する．

このため，$\tilde{f}(x,y)$ $((x,y) \in \tilde{\mathbf{C}}_k)$ の全ての画素で，再度クラスタリングをすることで，座標変換後のクラスタの領域外不要画素を小さくする．クラスタリングの結果，最も画素数の多いクラスタを $k$ 番目の距離画像でのクラスタ $\tilde{\mathbf{C}}_k^f$ として採用する．このクラスタリングにより．Fig. 2(c) のように，距離画像上においても近い距離を持った画素のみの検出することができる．

最後に，得られたクラスタ $\tilde{\mathbf{C}}_k^f$ 内での距離画像の輝度値を平均することで，距離を導出し，クラスタの重心を求めることで，頭部の三次元位置を導出する．

## 3 PF を用いた動的頭部検出

2 章では，フレーム毎処理を述べた．実環境において，人間は常に動いており，センサに対して常に同じ方向を向いていると限らないため，クラスタの重心位置や形状は常に変化する．この場合，あるフレームでは頭部を検出できたとしても，センサノイズや形状変化の影響により，次のフレームでは頭部と検出できない場合が起こりうる．この状況に対応するためには，フレーム間相互の情報を使った追従処理によるロバスト化が必要である．

本稿では，複数同時頭部検出や，人間が常に等速運動している仮定を緩和するため，運動モデルの設計の自由度が高く，誤検出からの復帰が容易である利点を持つ PF を用いた．全章で観測された各クラスタに対して，粒子数 $N_p$ の粒子群を設ける．粒子群は，3 次元位置とその 3 次元位置を算出した時の分散を保持している．

PF の処理の流れは以下となる．

A1) 全ての粒子群にランダムウォーク運動モデルを仮定
A2) 既に存在する粒子群の中から，ユークリッド距離が $d_{max}$[m] であり，かつ最も近い位置を示しているクラスタを探索する（観測モデル）．
A3) (A2) で粒子群に対応するクラスタが発見された場合，その粒子群を観測情報でリサンプルする．
A4) (A2) で粒子群に対応するクラスタが発見されなかった場合，新たに粒子群を作成する．
A5) (A3) の粒子群の連続観測回数 $N_o$ を増やし，連続非観測回数 $N_u$ を 0 とする．
A6) (A4) の粒子群の $N_o$ を 0 とし，$N_u$ を増やす．
A7) $N_u \geq N_{u_{max}}$ となった場合，その粒子群は削除．

**Fig.**3 Experimantal Setup

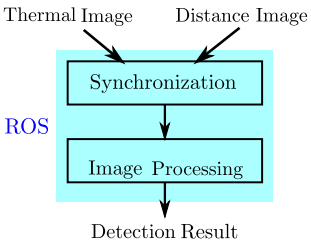



**Fig.**4 System Structure

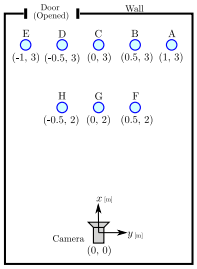



**Fig.**5 Downward Field View

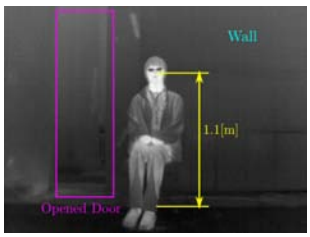



**Fig.**6 Camera View

A8) $N_o \geq N_{o_{max}}$ の粒子群のみを頭部とする．

$N_p, d_{max}, N_{o_{max}}, N_{u_{max}}$ がフィルタ設計の自由度であり，5 章で述べる．

## 4 システム構成

本章では，評価に用いたハードウェアとシステム構成について述べる．

熱画像センサは Apiste 製の FSV-1100 を用い（60[fps] で 320 × 240 ピクセルの画像を取得），3D 距離カメラには，Mesa imaging 製の SwissRanger SR4000（54[fps] で 176 × 144 ピクセルの画像を取得）を用いた．両センサを Fig.3 のように上下に合わせて設置し評価を行った．

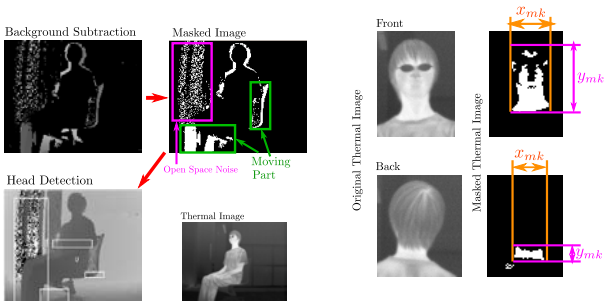
実時間処理のため，ソフトウェア側は全ての機能を，Willow Garage 社の提供するオープンソースの ROS[9] に実装した．システム構成図を Fig.4 に示す．図のように二つのセンサからの入力を ROS のノードで同期を取った後に頭部検出処理を行う．計算機環境としては Linux OS で 2.5 GHz Intel Core 2 Duo CPU と，2GB SDRAM を搭載した計算機を用いた．

## 5 評価実験

本章では 4 章で構築した熱画像と 3D 距離カメラの統合システムの評価を行う．まずは，統合させたことによる有効性を確認するため，人間が頭部を動かさず，その場で回転している状況下において，二つのセンサを統合したマルチモーダルなシステムを，ユニモーダルなシステム（熱画像のみの場合，3D 距離カメラのみの場合）と比較を行うことで性能比較・評価を行う．この評価では，画像処理の違いのみによる性能を比較するため，PF の処理を含めない評価を行う．次に，PF の実装の有効性を確認するため，人間の移動下で，2 章での処理は変えずに，PF の有無による頭部の追従性能の比較を行う．

### 5.1 静的頭部検出手法の性能比較

本節では，頭部位置を固定した時の，マルチモーダルとユニモーダルなシステムでの比較を示す．




(a) Distance Image (b) Thermal Image

**Fig.**7 Unimodal Clustering Approach

#### 5.1.1 実験条件

実験環境を Fig. 5 に示す．Fig. 3 のセンサ系を室内に設置し，そこから Fig. 5 の A-H の 8 個の各点で，頭部を検出する．顔の向きによる評価を行うため，人は頭が回転中心となるように回転椅子に座り，回転した状態で計測を行う．また，背景は，壁の有無によるロバスト性の評価を含めるため，壁面だけでなく，ドアを開け，遠方まで計測されるようにしている．

Fig. 6 が熱画像センサからの入力画像となる．センサ系は高さ 0.7[m] の台に設置され，人間側は頭部の中心がおよそ 1.1[m] となるように椅子の高さが調節されている．ゆえに，センサで計測される頭部の高さのノミナル値は 0.4[m] となる．

次に，2 章で示した処理の設計パラメータを設計する．予備実験などの結果より，$T_L = 33[\mathrm{deg}], T_H = 40[\mathrm{deg}], N_{cls} = 4, R_{max} = 2000, R_{min} = 50, \gamma = 2$ と設計した．

#### 5.1.2 比較対象の処理

次に比較対象のユニモーダルシステムの処理を示す．3D 距離カメラのみによる頭部検出の処理を Fig. 7(a) に示す．3D 距離カメラの場合は形状の抽出ができるが，人間とその他の物体の分類が困難であるため，“動作しているもの” を人間として検出することを想定し，距離画像から得た背景差分画像から頭部を検出することとする．よって検出アルゴリズムは以下となる．

B1) 距離画像の背景差分画像を得る．
B2) 背景差分画像をある閾値で二値化
B3) 画像をクラスタリングし，2.1 章と同様に不要クラスタを除外．

次に，熱画像のみによる頭部検出の処理を Fig. 7(b) に示す．熱画像は，体温を利用し，頭部を検出することができるが，一枚の熱画像から 3 次元位置を算出する必要がある．まず前段処理として 2.1 章と同じ手順でクラスタ $\mathbf{C}_k$ を得ておく．$\mathbf{C}_k$ の $x, y$ 方向の画素数をそれぞれ $x_{mk}, y_{mk}$ とする．$x_{mk}, y_{mk}$ からの三次元位置推定は以下で行われる．

C1) 頭部の大きさを $x$, $y$ 方向に 0.15[m], 0.25[m] と仮定．
C2) 得られたクラスタサイズ $x_{mk}$, $y_{mk}$ が 0.15[m], 0.25[m] とした時の距離計算を，$x, y$ 方向それぞれについて行う．
C3) 各 $x, y$ 方向で推定された距離の平均を採用．

**Table** 1 Result Comparison of Three Methods

| | | $\bar{e}_x$ [m] | $\bar{e}_y$ [m] | $\bar{e}_z$ [m] | $\bar{e}_d$ [m] | $\mathcal{E}$ | $\mathcal{I}$ | $\mathcal{S}$ |
|---|---|---|---|---|---|---|---|---|
| s | t | 0.182 | 0.068 | 0.648 | 0.695 | 0.123 | 0 | 0.549 |
| | d | 0.306 | 0.392 | 0.529 | 0.827 | 0.147 | 0.758 | 0.798 |
| | b | 0.054 | 0.027 | 0.085 | 0.117 | 0.134 | 0 | 0.006 |
| n | t | 0.126 | 0.066 | 0.499 | 0.536 | 0.112 | 0.031 | 0.331 |
| | d | 0.207 | 0.479 | 0.375 | 0.709 | 0.405 | 0.276 | 0.560 |
| | b | 0.043 | 0.028 | 0.072 | 0.096 | 0.112 | 0.031 | 0.006 |
| c | t | 0.186 | 0.103 | 0.781 | 0.822 | 0.340 | 0.134 | 0.440 |
| | d | 0.367 | 0.407 | 0.639 | 0.933 | 0.147 | 0.635 | 0.800 |
| | b | 0.043 | 0.035 | 0.092 | 0.117 | 0.339 | 0.129 | 0.017 |

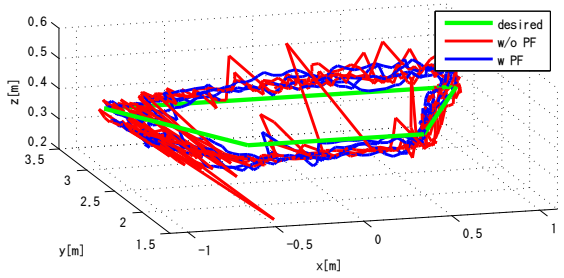



**Fig.**8 Comparizon with/without Particle Filter

#### 5.1.3 比較結果

本節では，2 章で提案したマルチモーダルなシステム（本章では方法 **b** と表す.）と，5.1.2 章での距離画像による検出システム（方法 **d**），熱画像による検出システム（方法 **t**）の比較を行う．

本評価では，多くの画像センサを使った手法が影響を受けやすい明暗の変化と衣服の変化を考慮し，実験環境で照明を点灯して被験者が半袖の衣服を着た時の結果（結果 **s**），実験環境を消灯した結果（結果 **n**），被験者が長袖で色の異なる衣服を着た時の結果（結果 **c**）を示す．

Table 1 は，Fig. 5 の A-H の各点で頭部を回転させた時の，頭部位置の $x, y, z$ 方向の推定誤差 $\bar{e}_x, \bar{e}_y, \bar{e}_z$，頭部位置のユークリッド距離の誤差 $\bar{e}_d$，頭部推定の削除誤り率 $\mathcal{E}$，挿入誤り率 $\mathcal{I}$，置換誤り率 $\mathcal{S}$ の A-H の各点の結果を平均したものである．ここで，頭部を検出できなかったフレーム $i$ の頭部個数 $N_{e_i}$，頭部を存在数以上に検出してしまったフレーム $i$ の頭部個数 $N_{i_i}$，頭部以外の物体を頭部として検出してしまったフレーム $i$ のクラスタ数 $N_{s_i}$，全フレーム数を $N_a$ として，$\mathcal{E} = \sum_i N_{e_i}/N_a$，$\mathcal{I} = \sum_i N_{i_i}/N_a$，$\mathcal{S} = \sum_i N_{s_i}/N_a$ と計算される．

表より，結果 **s**，**n**，**c** 全てにおいて方法 **b** の $\bar{e}_x, \bar{e}_y, \bar{e}_z, \bar{e}_d$ の推定誤差が小さく，また，明暗の変化や衣服の変化に影響を受けず，0.1[m] 程度の誤差で推定できていることがわかる．

また，結果 **d** の $\mathcal{I}, \mathcal{S}$ が他の手法よりも高いことがわかる．Fig. 7(a) では，背景差分画像の一例を示しているが，動作している部分と，ドアを開けた空間に大きな雑音を有して，頭部として検出していることがわかる．このため，複数の箇所を頭部として検出することによる $\mathcal{I}$，頭部とは別の物体を頭部とみなしたことによる $\mathcal{S}$ が増加したことがわかる．結果として，結果 **d** の頭部検出位置の誤差は大きくなる．

結果 **t** の頭部の位置の誤差が大きいのは，クラスタサイズに強く影響されていることが理由として考えられる．例えば，Fig. 7(b) のように，顔が前向きの場合と後ろ向きの場合は同じ位置に頭部があっても，検出されるクラスタサイズは大きく異なる．このため，(C1) の仮定で想定したクラスタと実際に検出されたクラスタに差異が生じ，前節の (C1)-(C3) の 3 次元位置の計算に誤差を与えたと考えられる．

また，方法 **t**, **b** では結果 **c** において，$\mathcal{E}$ が増加している．これは，襟のついた長袖の衣服を着たことによって，首が隠れ，後ろ向きの時に首が検出できなかったことによるものである．このように，衣服による影響は起こりうるため，動的な閾値設定や，熱画像や距離画像以外の情報を統合した手法等は今後の課題である．しかし，依然として，明暗の変化や衣服の色彩の変化によらないロバストな検出を実現することができた．

### 5.2 動的頭部検出手法の性能比較

次に，動的頭部において PF の有無による追従性能の比較を行う．実験では，Fig. 5 を，A, B, C, D, E, H, G, F の順に 3 回巡回し，頭部をトラッキングし，軌道のばらつきを評価するために目標軌道からの誤差の平均を算出した．3 章の PF の設計パラメータは $N_p = 100$, $d_{max} = 2$[m], $N_{o_{max}} = 5$, $N_{u_{max}} = 10$ とした．

目標軌道と推定軌道の比較結果を Fig. 8 に示す．図より，PF を実装した方が軌道のばらつきが少ないことが確認できる．実際に目標軌道からのユークリッド距離の誤差を各観測点で平均した値は，PF のある場合 0.0807[m]，無い場合 0.0837[m] となり，誤差の少ない軌道を実現でき，PF の有効性を確認することができた．

## 6 結論

本稿では熱画像センサと 3D 距離カメラのマルチモーダル情報を統合することによる頭部検出システムの提案と評価を行った．提案手法は，明暗の変化や衣服の変化にロバストな手法であり，高い精度で 3 次元位置を推定できることを確認した．また，動的頭部に対応するため，PF を実装し，その推定軌道誤差に評価から有効性を示した．

### 参考文献

[1] A. Jaimes *et al.*, “Multimodal human-computer interaction : A survey,” *Computer Vision and Image Understanding*, vol. 108, pp. 116–134, 2007.

[2] K. Nakamura *et al.*, “Intelligent Sound Source Localization for Dynamic Environments,” *IROS 2009*, pp. 664-669.

[3] K. Nakamura *et al.*, “音源への選択的注意を実現する音源同定と音源定位の統合システム,” *RSJ 2010*, to be presented.

[4] K.C. Frerstenberg *et al.*, “Pedestrian detection and classification by laserscanners,” *IEEE Intelligent Vehicles Symposium*, Paris, France, 2002.

[5] S. Milch *et al.*, “Pedestrian detection with radar and computer vision,” *Progress in Automobile Lighting*, 2001.

[6] T. Helene *et al.*, “Advanced surveillance systems: combining video and thermal imagery for pedestrian detection,” *Proc. of the SPIE*, vol. 5405, pp. 506–515, 2004.

[7] Z. C. Marton *et al.*, “Probablistic Categorization of Kitchen Objects in Table Settings with a Composite Sensor,” *IROS 2009*, pp. 4777–4784.

[8] D. Geronimo *et al.*, “Survey of Pedestrian Detection for Advanced Driver Assistance Systems,” *IEEE trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, 2010.

[9] M. Quigley *et al.*, “ROS: an open-source Robot Operating System”, *Open-Source Software workshop of ICRA 2009.*