Thermal and 3D Kinect Sensor Fusion for Robust People Detection using Evolutionary Selection of Supervised Classifiers

L. Susperregi[1], E. Jauregi[2], B. Sierra[2], J. M. Martínez-Otzeta[1], E. Lazkano[2] and A. Ansuategui[1]
[1]*TEKNIKER-IK4, Autonomous and Smart Systems Unit., Eibar, Spain*
[2]*Department of Computer Science and Artificial Intelligence, University of Basque Country, Donostia/San Sebastián, Spain*

Keywords: Computer Vision, Machine Learning, Robotics, 3D People Detection, Estimation of Distribution Algorithms.

Abstract: In this paper we propose a novel approach for combining information from low cost multiple sensors for people detection on a mobile robot. Robustly detecting people is a key capability needed for robots that operate in populated environments. Several works show the advantages of fusing data coming from complementary sensors. Kinect sensor offers a rich data set at a significantly low cost, however, there are some limitations using it in a mobile platform, mainly that Kinect relies on images captured by a static camera. To cope with these limitations, this work is based on the fusion of Kinect and thermopile array sensor mounted on top of a mobile platform. We propose the implementation of evolutionary selection of people detection supervised classifiers built using several computer vision transformation. Experimental results carried out with a mobile platform in a manufacturing shop floor show that the percentage of wrong classified using only Kinect is drastically reduced with the classification algorithms and with the combination of the three information sources.

1 INTRODUCTION

Service robots, now and in the near future, performing tasks as assistants, guides, tutors, or social companions in human populated settings such as museums, hospitals, etc. pose two main challenges: by the one hand, robots must be able to adapt to complex, unstructured environments and, on the other hand, robots must interact with humans. While interacting with the environment, the robot must navigate, detect and avoid obstacles (Morales et al., 2011). A requirement for natural Human Robot interaction is the robot's ability to accurately and robustly detect and localize the persons around it in real-time. This problem is a challenging one, quite difficult when a low cost camera is the only available sensor (Yao and Odobez, 2011).

This article describes the realization of a human detection system based on low-cost sensing devices. Recently, research on sensing components and software lead by Microsoft provide useful results for extracting the human kinematics (Kinect motion sensor device (Kinect,)).

Within this article, the service proposed by the mobile robot is to approach the closer person in the room, i.e. to approach the person to a given distance and to verbally interact with him. This "engaging" behaviour can be useful in potential robot services such a tour guide, health care or information provider. Once the target person has been chosen, the robot plans a trajectory and navigates to the desired position. To accomplish this the robot must be able to detect human presence in its vicinity and it cannot be assumed that the person faces the direction of the robot since the robot acts proactively.

Kinect offers a rich data set at a significantly low cost. While the Kinect is a great addition to robotics there are some limitations. First, the depth map is only valid for objects more than 80cm away from the sensing device. Second, the Kinect uses an IR projector with an IR camera which means that sunlight could affect negatively, taking into account that the sun emits in the IR spectrum. Third, Kinect rely on the detection of human activities captured by a static camera. In mobile robot applications the sensors setup is assumed to be embedded in the robot that is usually moving. As a consequence the robot is expected to evolve in environments which are highly dynamic, cluttered, and frequently subjected to illumination changes. To cope with this, this work is based on the hypothesis that the combination of Kinect and thermopile array sensor (low cost Heimann HTPA thermal sensor, (HTPA,)) can significantly improve the robustness of human detection. Thermal vision

Susperregi L., Jauregi E., Sierra B., María Martínez-Otzeta J., Lazkano E. and Ansuategi A..
Thermal and 3D Kinect Sensor Fusion for Robust People Detection using Evolutionary Selection of Supervised Classifiers.
DOI: 10.5220/0004397601020110
In *Proceedings of the 10th International Conference on Informatics in Control, Automation and Robotics* (ICINCO-2013), pages 102-110
ISBN: 978-989-8565-71-6

helps to overcome some of the problems related to colour vision sensors, since humans have a distinctive thermal profile compared to non-living objects and there are no major differences in appearance between different persons in a thermal image. Another advantage is that the sensor data does not depend on light conditions and people can also be detected in complete darkness. Therefore it is a promising research direction to combine the advantages of different sensor sources because each sensing modality has complementary benefits and drawbacks.

This article outlines the design and development of a multimodal human detection system. The chosen approach is:

- To combine machine learning paradigms with computer vision techniques in order to perform image classification: first we apply transformations using computer vision techniques and afterwards we perform classification using machine learning paradigms.
- To combine the resulting classifiers obtained by this new image classification paradigm. Appart of using all the classifiers obtained (paradigms $\times$ transformations), we use a new aproach in multi classifier construction in which a previous selection of classifiers is performed.

We have experimented in a real manufacturing shop floor where machines and humans share the space in performing production activities. Experiments seem promising considering that the percentage of wrong classified using only Kinect's detection algorithms is drastically reduced.

2 RELATED WORK

People detection and tracking systems have been studied extensively due to the increase of demand of advanced robots that must integrate natural human-robot interaction capabilities in order to perform some specific tasks for the humans or in collaboration with them. As a complete review on people detection is beyond the scope of this work, an extensive work can be found in (Schiele, 2009), we focus on most related work.

People detection solutions that can be used on mobile robots should cope with several requirements:

- Camera and other sensors are usually not static since they are mounted on a moving platform. As a consequence, many algorithms aim at the surveillance applications are not applicable.
- fast (real-time). The computational load of the usedalgorithms should be low in order to perform real-time detection.
- non-invasive (normal human activity is unaffected).

To our knowledge, two approaches are commonly used for detecting people on a mobile robot. One, vision based techniques, and another approach, combining vision with other modalities, normally range sensors such as laser scanners or sonars like in (Guan et al., 2007). Methods for people detection in colour images extract features based on skin colour, face, clothes and motion information such as (Bellotto and Hu, 2010). All methods for detecting and tracking people in colour images on a moving platform face similar problems and their performance depends heavily on the current light conditions, viewing angle, distance to persons, and variability of appearance of people in the image.

Most existing combined vision-thermal based methods, in (St-Laurent et al., 2006; Hofmann et al., 2011; Johnson and Bajcsy, 2008; Thi Thi Zin and Hama, 2011), concern non-mobile applications in video monitoring applications, and especially for pedestrian detection where the pose of the camera is fixed . Some works, (Gundimada et al., 2010), show the advantages of using thermal images for face detection. They suggest that the fusion of both visible and thermal based face recognition methodologies yields better overall performance.

As yet, however, there is hardly any published work on using thermal sensor information to detect humans on mobile robots. The main reason for the limited number of applications using thermal vision so far is probably the relatively high price of this sensor. (Treptow et al., 2005) shows the use of thermal sensors and grey scale images to detect people in a mobile robot. A drawback of most of these approaches is the sequential integration of the sensory cues. People are detected by thermal information only and are subsequently verified by visual or auditory cues.

Most of the abovementioned approaches have mostly used predefined body model features for the detection of people. Few works consider the application of learning techniques. (Arras et al., 2007) proposes to use supervised learning (AdaBoost) to create a people detector with the most informative features . (Mozos et al., 2010) builds classifiers able to detect a particular body part such as a head, an upper body or a leg using laser data.

Combination of classifiers has been widely used as a useful approach in several machine learning tasks (Kuncheva, 2004). In the field of people detection several authors have used this approach, like (Oliveira et al., 2010),that use histograms of oriented gradients

(HOGs) and local receptive fields (LRFs), which are provided by a convolutional neural network, and are classified by multilayer perceptrons (MLPs) and support vector machines (SVMs) combining classifiers by majority vote and fuzzy integral.

3 PROPOSED APPROACH

We propose a multimodal approach, which can be characterized by the fact that all used sensory cues are concurrently processed. The proposed detection system is based on a Kinect motion sensor device for the XBOX 360 and a HTPA thermal sensor developed by Heimann, (HTPA,), mounted on top of a RMP Segway mobile platform, which is shown in Figure 1.

Figure 1: The used robotic platform: a Segway RMP 200 provided with the Kinect and the thermal sensor.

We aim at applying a new approach to combine machine learning paradigms with computer vision techniques in order to perform image classification. Our approach is divided into three phases: transformation using computer vision techniques, classification using machine learning paradigms and optimal combination of classifiers using a previous classifier selection by means of EDA (Estimation of Distribution Algorithms) (Müehlenbein and Paaß, 1996).

1. Computer vision transformations. In order to have different views of the images, different modifications over the original pictures are performed. The main goal of this phase is to have variability in the aspect the picture offers, so that different values are obtained for the same pixel positions. As it has been mentioned before, we aim at using three input images (colour, depth, temperature) to construct a classifier. To enrich the input database, we have decided to build some variants using the matrix obtained in the original images, applying computer vision related transformations. In this way, and for each of the three data sources, a set of equivalent images is obtained, and a set of databases are constructed, one for each of the transformation used.

   To achieve this, we combine some standard image related algorithms (edge detection, gaussian filter, binarization, and so on) in order to obtain different views of the images, and afterward, we apply some standard machine learning classifiers taking into account the pixel values of the different modifications of the pictures. From the original training database collected, a new training database is obtained for each of the computer vision transformation used, summing up a total of 24 databases for each device.
2. In the classification phase, the system learns a classifier from a hand-labeled dataset of images (abovementioned original and transformations). As classifiers we use of five well known ML supervised classification algorithms with completely different approaches to learning and a long tradition in different classification tasks: IB1, Naive-Bayes, Bayesian Network, C4.5 and SVM.
3. Then, the goal of our fusion process is to maximize the benefits of each modality by intelligently fusing their information, and by overcoming the limitations of each modality alone.

- Considering the large number of possible classifiers combinations (24x5 for each sensor) we attempt to get an optimal solution making a selection of a subset of classifiers which obtain better result from the accuracy point of view. An evolutionary algorithm called Estimation of Distributions Algorithm (EDA) is used to perform the selection.

3.1 Data Sources

s stated before, two kind of data sources are used coming from the Kinect sensor and the thermopile array.

Kinect 3D Images. Kinect provides 3D images, it uses near infrared light to illuminate the subject and the sensor chip measures the disparity between the information received by the two IR sensors. It provides a 640x480 distance (depth) map in real time (30 fps). In addition to the depth sensor the Kinect also provides a traditional 640x480 RGB image.

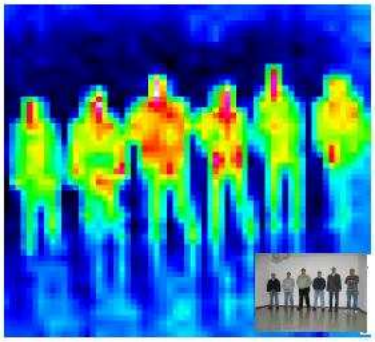

Figure 2: Image thermopile.

Thermal Images. The HTPA allows the measurement of temperature distribution of the environment, where very high resolutions are not necessary, such as person detection, surveillance of temperature critical surfaces, hotspot or fire detection, energy management and security applications. The sensor only offers a 32x31 image that allows a rough resolution of the temperature of the environment as it is shown in Figure 2. The benefits of this technology are low costs, the very small power consumption, small size, as well as the high sensitivity of the system.

3.2 Computer Vision Transformations

Three image type data taken in parallel (image, distance, temperature) are used to build a classifier whose goal is to identify whether a person is in the viewscope of the robot or not. Figure 3 shows an example of the three different images obtained; each image is considered as a gray scale one, and the value of each pixel, position in the matrix, is considered as a predictor variable within the Machine Learning database construction, summing up $n \times m$ features, being m the column number and n the row number in the image. Each image corresponds to a single case in the generated database.

In order to have different views of the images, different modifications over the original pictures are performed. The main goal of this phase is to have variability in the aspect the picture offers, so that different values are obtained for the same pixel positions.

We have selected some of the most common transformations offered by related software, in order to show the benefits of the proposed approach making use of simple algorithms. Table 1 presents the transformations used, as well as a brief description of each of them. It is worth to point out the fact that any other CV transformation could be used apart from the selected ones.

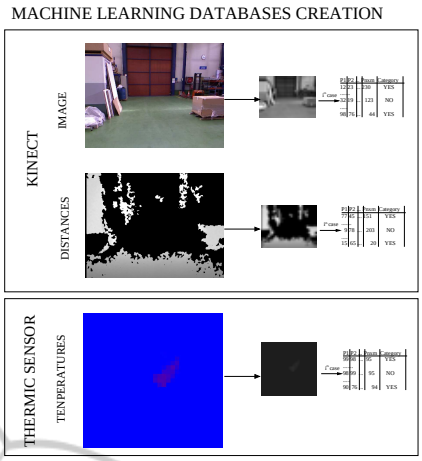

Figure 3: Image preprocessing and training database creation.

Table 1: Used image transformations.

Transform	Command	Effect
Transf. 1	Convolve	Apply a convolution kernel to the image
Transf. 2	Despeckle	Reduce the speckles within an image
Transf. 3	Edge	Detect edges in the image
Transf. 4	Enhance	Apply a filter to enhance a noisy image
Transf. 5	Equalize	Perform histogram equalization
Transf. 6	Gamma	Perform a gamma correction
Transf. 7	Gaussian	Reduce image noise and levels
Transf. 8	Lat	Local adaptive thresholding
Transf. 9	Linear-Str.	Linear with saturation histogram stretch
Transf. 10	Median	Apply a median filter to the image
Transf. 11	Modulate	Vary the brightness, saturation, and hue
Transf. 12	Negate	Negate the image
Transf. 13	Radial-blur	Radial blur the image
Transf. 14	Raise	Create a 3-D effect
Transf. 15	Selective-blur	Blur pixels within a contrast threshold
Transf. 16	Shade	Shade the image
Transf. 17	Sharpen	Sharpen the image
Transf. 18	Shave	Shave pixels from the image edges
Transf. 19	Sigmoidal	Increase the contrast
Transf. 20	Transform	Affine transform image
Transf. 21	Trim	Trim image edges
Transf. 22	Unsharp	Sharpen the image
Transf. 23	Wave	Alter an image along a sine wave

3.3 Machine Learning Classifiers

As classifiers we use five well known ML supervised classification algorithms (Mitchell, 1997) with completely different approaches to learning and a long tradition in different classification tasks: IB1, Naive-Bayes, Bayesian Network, C4.5 and SVM. Then, the goal of our fusion process is to maximize the benefits of each modality by intelligently fusing their information, and by overcoming the limitations of each modality alone.

IB1. The IB1 (Aha et al., 1991) is a case-based, Nearest-Neighbor classifier. To classify a new test sample, all training instances are stored and the nearest training instance regarding the test instance is found: its class is retrieved to predict this as the class of the test instance.

Naive-Bayes. The Naive-Bayes (NB) rule (Cestnik, 1990) uses the Bayes theorem to predict the class for each case, assuming that the predictive genes are independent given the category. To classify a new sample characterized by d genes $\mathbf{X} = (X_1, X_2, ..., X_d)$, the NB classifier applies the following rule:

$$c_{N-B} = \arg\max_{c_j \in C} p(c_j) \prod_{i=1}^{d} p(x_i|c_j)$$

where c_{N-B} denotes the class label predicted by the Naive-Bayes classifier and the possible classes of the problem are grouped in $C = \{c_1, \ldots, c_l\}$.

Bayesian Networks. A Bayesian network, belief network or directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph (DAG). Probabilistic classifiers give to the new case the most likely class for the observed data. In this paper we have used Bayesian Networks as classification models (Sierra et al., 2009).

C4.5. The C4.5 (Quinlan, 1993) represents a classification model by a decision tree. It is run with the default values of its parameters. The tree is constructed in a top-down way, dividing the training set and beginning with the selection of the best variable in the root of the tree.

Support Vector Machines (SVM). SVM are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyperplane in that space, one which maximizes the margin between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are pushed up against the two data sets (Meyer et al., 2003).

3.4 Combination of Classifiers

In order to finally classify the targets as human or non human, the estimation of the Kinect based classifiers has to be combined with the estimation of the thermal

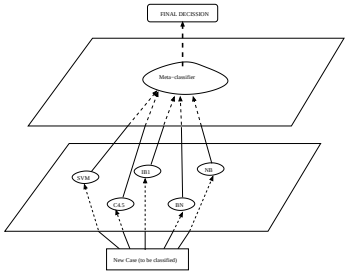

Figure 4: Stacked Generalization schemata.

based classifier. After building the individual classifiers ($5 \times 24 = 120$ for each sensor) the aim is at combining the output of the different classifiers to obtain a more robust final people detector.
Two approaches are performed and compared:

1. Stacked generralization approach: standard multiclassifier to combine the 360 classifiers
2. Classifier Subset Selection Stacked. A selection of some of the classifiers is done first, and then the combination is performed among the selected classifiers.

Stacked Generalization. The last step is to combine the results of the classifiers obtained for the three sensors (colour, distance, temperature). To achieve this, we use a bi-layer Stacked Generalization approach (Wolpert, 1992; Sierra et al., 2001) in which the decision of each of the 360 single classifiers is combined by means of another method, the so called meta-classifier. Figure 4 shows the typical approach used to perform a classification with this multiclassifier approach. It has to be noticed that the second layer classifier could be any function, including a simple vote approach among the used classifiers.

We have used this multiclassifier to combine the different classifiers learned in each type of image. It is worth to notice that this is done for comparison reasons only, as our proposal is to use some of those classifiers only, to reduce computational load and, at the same time, to increase the obtained accuracy.

Classifier Subset Selection Stacking. The new multiclassifier paradigm, which extends the Staking Generalization approach is shown in figure 5. As it can be seen, we added to the multiclassifier an intermediate phase in which a subset of the classifiers belonging to the first layer are selected. The criterion to make the selection depends on the goal of the classification task, and we have decided to use the classification accuracy in our case.

The way the classifiers are selected (and discarted) is not unique; due to our previous experience, we de-

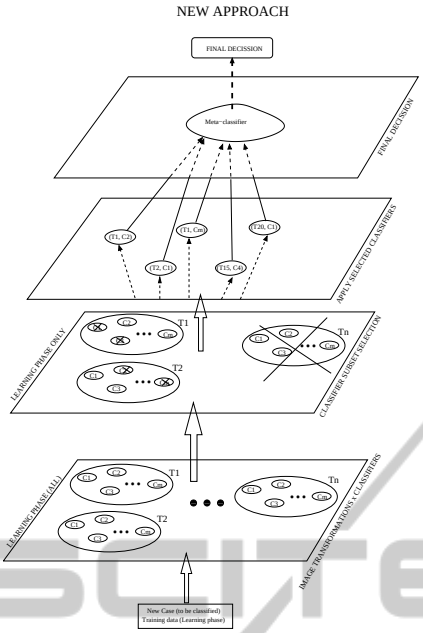

Figure 5: Classifier Subset Selection Stacking.

cided to use Estimation of Distribution Algorithms to perform the so called Classifier Subset Selection (CSS) which reduce the number of classifiers to be used in the final model, decreasing in this way the computational payload while increasing the obtained accuracy.

3.4.1 Estimation of Distribution Algorithms

Estimation of distribution algorithms (EDAs) have successfully been developed for combinatorial optimization (Inza et al., 2000). They combine statistical learning with population-based search in order to automatically identify and exploit certain structural properties of optimization problems.

4 EXPERIMENTAL SETUP

The manufacturing plant is a real manufacturing shop floor where machines and humans share the space in performing production activities. The shop floor in Figure 6 can be characterized as an industrial environment, with high ceilings, fluorescent light bulbs, high windows, etc. The lighting conditions are very changing from one day to another and even in different locations along the path covered by the robot.

Method. These are the steps of the experimental phase:

1. Collect a database of images that contains three data types that are captured by the two sensors: 640x480 depth map in real time (30 fps), 640x480 RGB image, 32x31 thermopile array.
2. Reduce the image sizes from 640×480 to 32×24, and convert colour images to gray-scale ones.
3. For each image, apply 23 computer vision algorithms, obtaining 23 new databases for each image type. Thus, we have 24 data sets for each image type.
4. Build 120 classifiers, applying 5 machine learning algorithms for each image type training data sets (5×24).
5. Apply 10 fold cross-validation using 5 different classifiers to each of the previous databases, summing up a total of $3 \times 24 \times 5 = 360$ validations.
6. Select a combined classifier among its 360 different models using two approaches: (1) a multiclassifier to combine all the classifiers learned in each type of image; (2) Classifier Subset Selection stacking approach .

Figure 6: Manufacturing plant.

Training Data Sets. The training data set is composed of 1064 samples. The input to the supervised algorithms is composed of 301 positive and 764 negative examples. The set of positive examples contains people at different positions and dressed with different clothing in a typical manufacturing environment. The set of negative examples is composed of images without people in the image and with other objects in the environment such as machines, tables, chairs, walls, etc.

To obtain the positive and negative examples the robot was operated in an unconstrained indoor environment (the manufacturing plant). At the same time, image data was collected with a frequency of 1Hz. During robot motion the images were hand-labeled as positive examples if people was visually detected in the image, and as negative examples otherwise.

5 EXPERIMENTAL RESULTS

Performance of the people detection system is evaluated in terms of detection rates and false positives or negatives. In order to make a fast classification – real time response is expected– we first transform the colour images in gray-scale 32×24, and reduce as well the size of the infrared images to 32×24 size matrix. Hence we have to deal with 768 predictor variables, instead of $307200 \times (3$ colours) of the original images taken by the Kinect camera.

First of all, we have used the five classifiers using the reduced original databases (32×24 for Images and Distances, 31×31 for thermal pictures). Table 2 shows the 10 fold cross-validation accuracy obtained. The best obtained result is 92.11% for the thermal images original database, and using SVM as classifier. The real time Kinect's algorithms accuracy among the same images was quite poor (37.50%), as the robot was moving around the environment and the Kinect has been made to be used as a static device. As a matter of fact, that has been the origin of the presented research.

Table 2: 10 Fold cross-validation accuracy percentage obtained for each classifier using original images.

Data source	BN	NB	C4.5	K-NN	SVM
RGB	89.20	71.74	82.63	90.89	85.35
Depth	86.29	68.64	83.29	90.89	84.04
Thermal	89.67	86.10	87.79	91.74	**92.11**

The same accuracy validation process has been applied to each image transformation on each image format. Table 3 shows the results obtained by each classifier on the transformed 23 image databases. The best result is obtained by the C4.5 classifier after transforming the images using Transformation 7 (Gaussian one).

After performing the validation over the distance images, the results shown in Table 4 are obtained. The best result is obtained again by the C4.5 classifier after transforming the images using Transformation 7 (Gaussian one), with a 92.82 accuracy.

Finally, the classifiers are applied to the thermal images, obtaining the results shown in Table 5. In this case we obtain the best result (93.52) for the SVM classifier, and for two of the used transformations (Transf. 8 –Lat– and Transf. 9 –Linear-strech–). Moreover, the obtained results are identical for both paradigms, so there are redundant algorithms and, if selected, only one of them can be used in the final combination obtaining indistinct results.

Table 3: Images: 10 fold cross-validation accuracy percentage obtained for each classifier using each of the proposed transformations.

Images	BN	NB	C4.5	K-NN	SVM
Transf. 1	89.20	71.74	90.89	82.63	85.35
Transf. 2	87.89	72.30	90.99	84.41	86.29
Transf. 3	83.19	74.84	87.98	75.87	81.41
Transf. 4	88.92	71.92	90.89	82.44	86.20
Transf. 5	86.76	71.64	89.77	80.47	80.66
Transf. 6	87.98	71.36	90.89	83.29	86.29
Transf. 7	87.79	64.79	**91.83**	85.92	84.79
Transf. 8	76.81	78.03	85.07	71.36	76.90
Transf. 9	88.54	73.90	91.17	81.31	84.98
Transf. 10	87.98	69.48	90.70	82.82	84.69
Transf. 11	85.54	72.96	91.55	82.07	85.26
Transf. 12	88.92	71.74	90.89	82.63	85.35
Transf. 13	88.73	68.64	90.99	82.63	85.45
Transf. 14	88.83	71.74	90.89	83.76	85.54
Transf. 15	89.20	71.74	90.89	82.63	85.35
Transf. 16	83.85	75.12	86.38	77.93	81.78
Transf. 17	89.77	71.46	90.23	83.00	82.44
Transf. 18	88.73	71.55	90.61	82.35	85.35
Transf. 19	88.17	70.61	91.46	82.82	86.10
Transf. 20	89.11	70.99	90.80	82.63	84.98
Transf. 21	89.20	71.74	90.89	82.63	85.35
Transf. 22	88.83	71.36	90.33	82.35	82.72
Transf. 23	88.73	72.30	90.80	83.85	85.82

5.1 Final Combination

The last step is to combine the results of the classifiers obtained, 120 by each sensor. To do that, we firstly use a Stacking classifier (Wolpert, 1992) in which the decision of each single classifier is combined by means of another classifier (ths so called metaclassifier). Table 6 shows the obtained results. As it can be seen, the best obtained accuracy is 95.31%, using a Bayesian Network as metaclassifier. It significantly improves the result of the best single classifier (93.52 for the Thermal images).

It is worth to mention that the best classifier combination obtained used a total of 51 single classifiers, and that all the three sensors are used, i.e., that transformations and related single classifiers to each of the sensors have been selected. Although the number of classifiers could be seen as high for a real time image processing, it has to be taken into account that we use small size images which are fast transformed, and that the classifiers, once constructed, give the classification result in miliseconds. A classifier parallelization could be used also to obtain a faster answer, as all of the single classifiers can be executed independently, but it is not realy necessary in this case.

Table 4: Distances: 10 fold cross-validation accuracy percentage obtained for each classifier using each of the proposed transformations.

Distances	BN	NB	C4.5	K-NN	SVM
Transf. 1	86.29	68.64	90.89	83.29	84.04
Transf. 2	86.38	68.45	91.27	83.38	82.91
Transf. 3	83.66	78.87	87.23	78.97	81.60
Transf. 4	86.10	68.54	90.89	82.91	83.29
Transf. 5	85.35	70.80	90.89	80.38	81.97
Transf. 6	86.38	70.33	90.61	82.25	83.76
Transf. 7	85.92	66.95	**92.86**	85.26	84.23
Transf. 8	83.19	73.62	84.04	73.15	78.40
Transf. 9	85.26	67.70	90.33	83.00	83.19
Transf. 10	85.54	68.92	92.30	85.16	85.35
Transf. 11	84.69	68.26	90.99	81.50	82.35
Transf. 12	86.67	68.64	90.89	83.38	84.04
Transf. 13	85.35	68.08	92.21	82.54	83.29
Transf. 14	86.57	68.73	90.89	83.76	84.13
Transf. 15	86.29	68.64	90.89	83.29	84.04
Transf. 16	83.66	78.69	87.14	80.38	85.35
Transf. 17	85.63	71.27	90.52	82.25	81.50
Transf. 18	85.63	66.20	89.77	82.72	82.54
Transf. 19	86.48	70.05	90.89	83.85	83.94
Transf. 20	86.67	69.01	90.70	83.29	83.85
Transf. 21	85.45	70.33	91.36	83.29	82.82
Transf. 22	85.73	71.08	90.42	81.78	81.60
Transf. 23	85.92	68.64	91.27	80.47	83.10

Table 5: Thermal sensor: 10 fold cross-validation accuracy percentage obtained for each classifier using each of the proposed transformations.

Thermical images	BN	NB	C4.5	K-NN	SVM
Transf. 1	89.67	86.10	91.74	87.79	92.11
Transf. 2	90.99	84.32	92.39	91.46	92.58
Transf. 3	89.30	86.67	90.80	86.29	92.39
Transf. 4	89.11	83.85	92.49	89.39	90.33
Transf. 5	85.73	84.60	92.77	90.33	85.63
Transf. 6	89.67	85.92	91.74	87.79	91.83
Transf. 7	86.57	82.16	89.67	87.79	89.95
Transf. 8	89.11	85.92	91.64	84.04	**93.52**
Transf. 9	90.80	88.08	92.39	87.89	**93.52**
Transf. 10	84.98	81.97	86.29	80.56	85.63
Transf. 11	71.74	71.74	71.74	71.74	71.74
Transf. 12	89.77	85.63	91.74	87.79	92.11
Transf. 13	90.05	84.69	92.77	90.14	91.08
Transf. 14	89.11	86.01	91.08	87.89	91.83
Transf. 15	89.67	86.10	91.74	87.79	92.11
Transf. 16	89.48	86.85	91.17	90.33	89.95
Transf. 17	89.67	87.23	91.74	87.04	90.99
Transf. 18	89.11	85.63	91.55	85.63	89.86
Transf. 19	89.67	85.07	91.83	87.79	91.83
Transf. 20	89.77	86.01	91.74	87.79	92.68
Transf. 21	83.57	47.89	84.41	82.54	72.02
Transf. 22	89.77	85.82	91.92	87.79	91.17
Transf. 23	90.05	85.45	92.02	90.33	91.27

Table 6: Multiclassifier combination: 10 fold cross-validation accuracy percentages obtained.

Metaclassifier	BN	NB	C4.5	K-NN	SVM
Results (360 classifiers)	*95.31*	94.93	94.27	94.93	94.27
Results (CSS)	**98.87**	96.53	96.06	98.12	97.37

6 CONCLUSIONS AND FUTURE WORKS

This paper presented a people detection system for mobile robots using using 3D camera and thermal vision and provided a thorough evaluation of its performance. The system uses a combination of Computer Vision and Machine Learning paradigms. This approach was designed to manage three kind of input images depth, color, and temperature to detect people. We showed that the detection of a person is improved by cooperatively classifying the feature matrix computed from the input data, where we made use of Computer Vision transformations and supervised learning techniques to obtain the classifiers. Our algorithm performed well across a number of experiments in a real manufacturing plant. This work serves as an introduction to the potential of multi-sensor fusion in the domain of people detection in mobile platforms. In the near future we envisage:

- To extend to other scenarios. The approach will be extended toward a museum scenario.
- To develop trackers combining/fusing visual cues using particle filter strategies, including face recognition, in order to track people or gestures.
- To integrate with robot's navigation planning ability to explicitly consider human in the loop during robot movement.
- To use other single classifier paradigms, and other transformations
- To use more complex computer vision approaches (SIFT, SFOP and so forth)

ACKNOWLEDGEMENTS

The work described in this paper was partially conducted within the ktBOT project and funded by KUTXA Obra Social, the Basque Government Research Team grant and the University of the Basque Country UPV/EHU, under grant UFI11/45 (BAILab).

REFERENCES

Aha, D., Kibler, D., and Albert, M. (1991). Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Arras, K. O., Martinez, O., and Burgard, M. W. (2007). Using boosted features for detection of people in 2d

range scans. In *In Proc. of the IEEE Intl. Conf. on Robotics and Automation*.

Bellotto, N. and Hu, H. (2010). A bank of unscented kalman filters for multimodal human perception with mobile service robots. *International Journal of Social Robotics*, 2(2):121–136.

Cestnik, B. (1990). Estimating probabilities: a crucial task in machine learning. In *Proceedings of the European Conference on Artificial Intelligence*, pages 147–149.

Guan, F., Li, L., Ge, S., and Loh, A. P. (2007). Robust human detection and identification by using stereo and thermal images in human-robot interaction. *International Journal of Information Acquisition*, 4(2):1–22.

Gundimada, S., Asari, V. K., and Gudur, N. (2010). Face recognition in multi-sensor images based on a novel modular feature selection technique. *Inf. Fusion*, 11:124–132.

Hofmann, M., Kaiser, M., Aliakbarpour, H., and Rigoll, G. (2011). Fusion of multi-modal sensors in a voxel occupancy grid for tracking and behaviour analysis. In *In: Proc. 12th Intern. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Delft, The Netherlands.

HTPA. Heimann sensor. http://www.heimannsensor.com/index.php.

Inza, I., Larrañaga, P., Etxeberria, R., and Sierra, B. (2000). Feature subset selection by bayesian networks based optimization. *Artificial Intelligence*, 123(1-2):157–184.

Johnson, M. J. and Bajcsy, P. (2008). Integration of thermal and visible imagery for robust foreground detection in tele-immersive spaces. In *Information Fusion, 2008 11th International Conference on*.

Kinect. Kinect sensor. http://en.wikipedia.org/wiki/Kinect.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley and Sons, Inc.

Meyer, D., Leisch, F., and Hortnik, K. (2003). The support vector machine under test. *Neurocomputing*, 55(1):169–186.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

Morales, N., Toledo, J., Acosta, L., and Arnay, R. (2011). Real-time adaptive obstacle detection based on an image database. *CVIU*, 115(9):1273–1287.

Mozos, O. M., Kurazume, R., and Hasegawa, T. (2010). Multi-part people detection using 2D range data. *International Journal of Social Robotics*, 2(1):31–40.

Müehlenbein, H. and Paaß, G. (1996). From recombination of genes to the estimation of distributions. In *Lecture Notes in Computer Science: Parallel Solving from Nature IV*, volume 1411, pages 178–187. Springer Verlag.

Oliveira, L., Nunes, U., and Peixoto, P. (2010). On exploration of classifier ensemble synergism in pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 11(1):16–27.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Schiele (2009). Visual People Detection - Different Models, Comparison and Discussion. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Sierra, B., Lazkano, E., Jauregi, E., and Irigoien, I. (2009). Histogram distance-based bayesian network structure learning: A supervised classification specific approach. *Decision Support Systems*, 48(1):180–190.

Sierra, B., Serrano, N., Larrañaga, P., Plasencia, E. J., Inza, I., Jiménez, J. J., Revuelta, P., and Mora, M. L. (2001). Using bayesian networks in the construction of a bi-level multi-classifier. A case study using intensive care unit patients data. *Artificial Intelligence in Medicine*, 22(3):233–248.

St-Laurent, L., Prvost, D., and Maldague, X. (2006). Thermal imaging for enhanced foreground-background segmentation. In *The 8th Quantitative Infrared Thermography (QIRT) conference*, Padova, Italie.

Thi Thi Zin, Hideya Takahashi, T. T. and Hama, H. (2011). Fusion of infrared and visible images for robust person detection. *Image Fusion*. InTech, Available from: http://www.intechopen.com/articles/show/title/fusion-of-infrared-and-visible-images-for-robust-person-detection.

Treptow, A., Cielniak, G., and Duckett, T. (2005). Active people recognition using thermal and grey images on a mobile security robot. In *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*, Edmonton, Canada.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5:241–259.

Yao, J. and Odobez, J.-M. (2011). Fast human detection from joint appearance and foreground feature subset covariances. *Computer Vision and Image Understanding*, 115(10):1414–1426.