# Robust People Detection and Tracking from an Overhead Time-of-Flight Camera

Alvaro Fernandez-Rincon, David Fuentes-Jimenez, Cristina Losada-Gutierrez,
Marta Marron-Romera, Carlos A. Luna, Javier Macias-Guarasa and Manuel Mazo
*Department of Electronics, University of Alcalá, Ctra. Madrid-Barcelona, km. 33,600, 28805-Alcalá de Henares, Spain*
{*alvaro.fernandez, david.fuentes, losada, marta, caluna, macias, mazo*}*@depeca.uah.es*

Keywords: People Detection, Tracking, Time-of-Flight, ToF Camera.

Abstract: In this paper we describe a system for robust detection of people in a scene, by using an overhead Time of Flight (ToF) camera. The proposal addresses the problem of robust detection of people, by three means: a carefully designed algorithm to select regions of interest as candidates to belong to people; the generation of a robust feature vector that efficiently model the human upper body; and a people classification stage, to allow robust discrimination of people and other objects in the scene. The proposal also includes a particle filter tracker to allow people identification and tracking. Two classifiers are evaluated, based on Principal Component Analysis (PCA), and Support Vector Machines (SVM). The evaluation is carried out on a subset of a carefully designed dataset with a broad variety of conditions, providing results comparing the PCA and SVM approaches, and also the performance impact of the tracker, with satisfactory results.

## 1 INTRODUCTION

In the last years, automatic people detection and tracking in a non-invasive way (without adding turnstiles or other contact systems for access control) has received a lot of attention because of its different applications such as access control, video-surveillance or behavior analysis.

In this paper, we propose a system for robust and reliable detection and tracking of multiple people from depth image sequences, acquired using an overhead ToF camera. The proposal works properly even if the number of people is high or if they are close to each other.

There are several works in the literature that propose different approaches for people detection. The first works (Ramanan et al., 2006; Jeong et al., 2013), are based on the use of an RGB camera. These proposals obtain suitable results under controlled conditions, but they do not work properly in scenarios with occlusions. In order to reduce the occlusions, other approaches use a camera in an overhead position (Antic et al., 2009; Cai et al., 2014). Other works (Dan et al., 2012; Del Pizzo et al., 2016) use the fusion of RGB and depth information (obtained using a Kinect® sensor (Sell and O'Connor, 2014)) in order to improve the detection.

However, using RGB images can imply an invasion of users' privacy, since there is information that could allow knowing the identity of the people in the scene. This can be a relevant issue in applications where there are privacy preservation requirements, due, among others, to legal considerations. Because of that, in the last few years, researchers have looked for alternatives in order to preserve the users' privacy. Some of them propose the use of overhead depth sensors or 2.5D cameras , based on Time of Flight (ToF) (Bevilacqua et al., 2006; Stahlschmidt et al., 2014; Jia and Radke, 2014) or structural light (Zhang et al., 2012; Galčík and Gargalík, 2013; Rauter, 2013; Zhu and Wong, 2013; Del Pizzo et al., 2016)for people detection and tracking, preserving their privacy.

The works described in (Zhang et al., 2012; Stahlschmidt et al., 2014; Jia and Radke, 2014; Del Pizzo et al., 2016) allow people detection preserving the users' privacy but, since these works do not include a classification stage, they cannot discriminate between people and other objects in the scene. Because of that, these proposals generate an important number of false positives in realistic scenarios.

Other approaches (Galčík and Gargalík, 2013; Rauter, 2013; Zhu and Wong, 2013) incorporate a classification stage in order to reduce the number of false positives. The strategies described in (Galčík and Gargalík, 2013), and (Zhu and Wong, 2013) obtain a descriptor based on the human head and shoul-



Fernandez-Rincon A., Fuentes-Jimenez D., Losada-Gutierrez C., Marron-Romera M., Luna C., Macias-Guarasa J. and Mazo M.
Robust People Detection and Tracking from an Overhead Time-of-Flight Camera.
DOI: 10.5220/0006169905560564
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 556-564
ISBN: 978-989-758-225-7

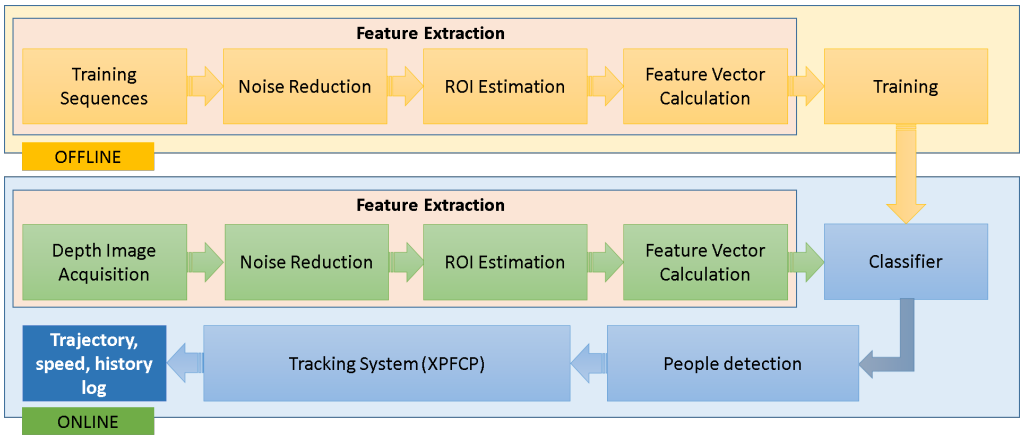



Figure 1: General System Architecture.

ders structure. These proposals allow people detection and are able to discriminate between people and other objects in the scene, but their detection rates drop significantly if people are close to each other.

Regarding the tracking of multiple people, multiple approaches have been developed during the last decades (Jia et al., 2008). Among them, the main alternatives can be divided into three groups: using an estimator for each object to follow (Isard and Blake, 1998), using a single estimator based on an extended state vector (MacCormick and Blake, 2000), and using a single multimodal estimator (Marron et al., 2005; Marron et al., 2010). Since there can be several people detected in any scene, it is necessary to implement an association algorithm in order to improve the reliability of the tracking process. There are different alternatives for this task, being the most widely used those based on Maximum Likelihood (ML), Nearest Neighbor (NN) and Probabilistic Data Association (PDA) (Bar-Shalom et al., 2011).

The structure of the paper is as follows: Section 1 provides a general introduction and a review of the literature, Section 2 describes the main modules of the system architecture, Section 3 includes the experimental setup, results and discussion, and Section 4 contains the main conclusions and future work.

# 2 SYSTEM DESCRIPTION

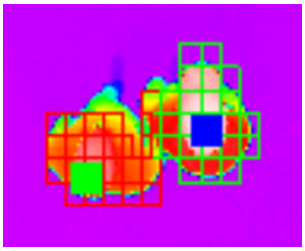
Figure 1 shows the general architecture of the proposed system. Its main modules will be described next, and we will devote more attention to the Feature Extraction strategy, and also to the Tracking Module, as they are the most relevant due to their novelty (the former) and the specific adaptations carried out (in the later).

## 2.1 Feature Extraction

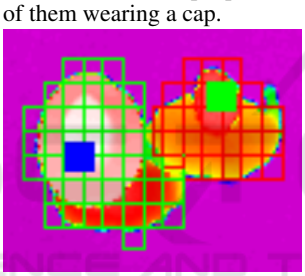
The people detection process includes an offline stage in charge of generating (training) the models to be used in the people classification stage, and the online process includes the following modules:

1. **Depth Image Acquisition (height acquisition).** The ToF camera is located in an overhead position at a $h_{camera}$ height from the floor, and its optical axis is perpendicular to the floor plane. To obtain the height matrix **H**, we subtract $h_{camera}$ to each pixel height of the depth image acquired by the camera.
2. **Noise Reduction.** One of the fundamental drawbacks in ToF cameras is the high noise level that is present in the depth image. This noise is especially significant if there are moving objects in the scene, reaching a great number of invalid pixels along the objects edges. To reduce the noise and the number of invalid pixels, we have implemented a noise filtering algorithm that includes two stages. In the first one, the invalid pixels are corrected using the mean value of the nearest valid pixels. We consider as invalid pixels those which are flagged by the camera as invalid pixels, and those with a height greater than the maximum height for a person (220 cm). Then, a nine element mean filter is applied to the height matrix **H** to smooth the detected surfaces.
3. **Regions of Interest (ROI's) Estimation.** In this work, we use a local maxima detection algorithm to select which regions in the height matrix **H** correspond to people or other objects. In case there are several people or objects in the scene, this algorithm must determine which pixels belong to each of them. The ROI's are defined as the pixels around each detected local maximum that belong

to the same object. Since the body parts of interest for this solution are the head, neck and shoulders, we impose the criterion that the height difference between the highest point on the head and the shoulders, should not be greater that an *interest height* $h_{interest}$. Taking into account anthropometric considerations (Matzner et al., 2015), we selected $h_{interest} = 40cm$. Finally, contour analysis (Luna et al., 2016) is performed to assign pixels to each ROI. Figure 2 shows two examples of ROI estimation in scenes with two people, some of them wearing accessories.



(a) Scene with two people, one of them wearing a cap.



(b) Scene with two people, one of them wearing a hat.

Figure 2: Examples of ROI's estimation.

4. **Feature Vector Calculation.** The feature vector components will be related to the pixel associated to the person surface in different height levels within the corresponding ROI. In this work, the feature vector is composed of six components (Luna et al., 2016). Five of these features will be related to the visible people or objects surfaces at different heights, and the sixth component will correspond to the relationship between the higher and lower diameters of the top surface, providing an idea on the eccentricity of the person head. The detail of the feature vector calculation is shown in Algorithm 1.

First, features related to the pixels associated to the head, neck and shoulders surfaces are calculated. To do this, we divide the $h_{interest}$ in 20 slices with a slice height $\Delta h$ (in this work, $\Delta h = 2cm$), counting the number of pixels found in each slice $s_i$, and building a vector $\mathbf{s} = \{s_1, s_2, \ldots, s_{20}\}$.

The components of the $\mathbf{s}$ vector are very sensitive to the appearance changes of a person (hair style, hair length, neck height, etc.), the person height, and, additionally, the effects of noise on the distance measures. To minimize the noise measurement errors, the first three components of $\mathbf{s}$ (spanning $6cm$) are integrated in component $\varphi_1$ of the feature vector $\varphi$. If the maximum value of the components $s_{1,2,3}$ is $s_3$, we assume that $s_1$ is corrupted by noise, and it is not taken into account. In this case, $\varphi_1$ will integrate $s_{2,3,4}$. The feature vector components $\varphi_{2,3}$ (corresponding to the head region too), and $\varphi_{4,5}$ (corresponding to the shoulders) integrate three $s_i$ values.

As the number of pixels associated to each component $\varphi_j$ depends of the person height, it is necessary to normalize them. To carry out the normalization, the relationship between the maximum height *hmax* and $\varphi_1$ was calculated. As an initial approximation, a quadratic relationship has been defined:

$$\widehat{\varphi} = a_0 + a_1 hmax + a_2 hmax^2 \qquad (1)$$

where $a_0$, $a_1$ and $a_2$ are the coefficients to estimate.

The Levenberg-Marquardt algorithm was used for the determination of those coefficients, using a sample set of people with heights between 140 cm and 213 cm. The final estimated values are $a_2 = 0.138$, $a_1 = -36.94$, and $a_0 = 2997$.

The normalized components ($\varphi_{1,2,3,4,5}$) of the feature vector provide information on the top view surfaces of people and objects, but initial experiments on people detection showed the need to also include more information related to the overhead geometry of the head. So, a sixth component $\varphi_6$ has been added to the feature vector. This component is calculated as the relationship between the major and minor axes of the region located $6cm$ below the maximum height ($s_{1,2,3}$). In Algorithm 1 the function that calculates $\varphi_6$ is referred to as $rba\{ROI_k(x_n, y_n), hmax_k\}$.

Figures 3, 4 and 6 show several examples of real depth frames for different situations, including the profile of the feature vectors obtained for selected elements.

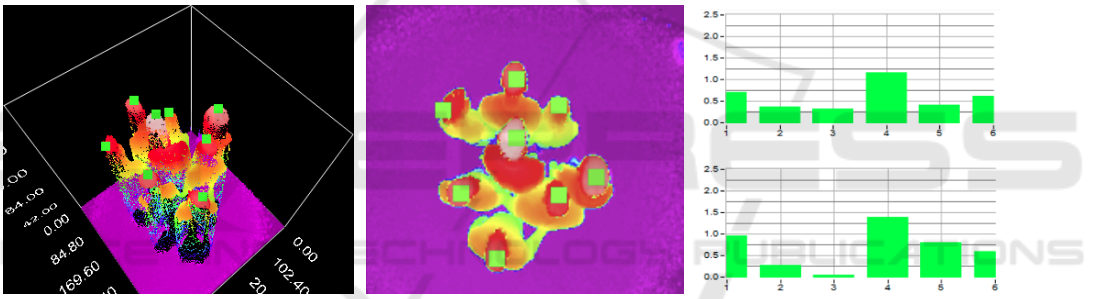
## 2.2 People Class Selection

Prior to defining the required classes that will be used to classify the detected ROIs as corresponding to a person or not, we designed a dataset that was meant to consider people with different heights, hair styles and colors, complexions, and wearing or not accessories that could heavily affect the feature vector components (wearing hats, caps, etc.).

From the study of the acquired data, we initially decided to define two classes, corresponding to peo-

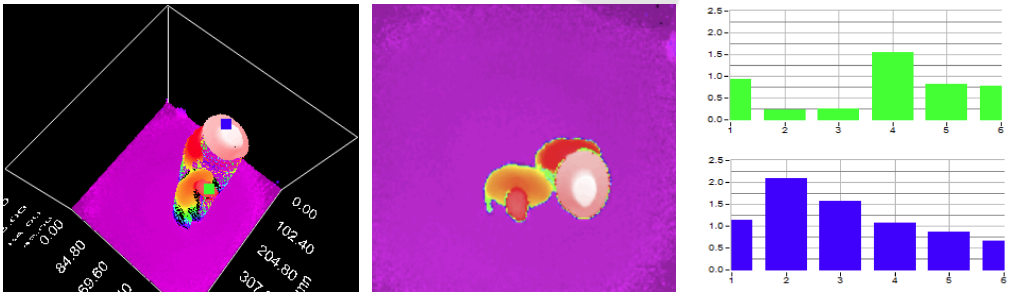
**Input** : height matrix **H**, number of *ROI's K*, $ROI_{1...K}(x,y), hmax_{1...K}, \widehat{\varphi}_1$
**for** *k=1..K* // Find feature vector for each ROI
**do**
  **for** *n=1..N* // N is number of pixels belonging to $ROI_k$
  **do**
    i = 1 + $round\{(hmax_k - ROI_k(x_n,y_n))/\Delta h\}$
      // $(x_n,y_n)$ are the coordinates of pixel $n$ belonging to $ROI_k$ and $\Delta h = 2cm$
    $s_i = s_i + 1$ // $s_i$ is the number of pixels in slice $i$, where $i = 1,\ldots,20$
  $u = argmax_{1\le i\le 3}\{s_i\}$ // Find the maximum value of $s_i$ where $i = 1,\ldots,3$
  **for** *j=1..3* **do**
    **if** $u < 3$ **then**
      $\varphi_j = \sum_{k=1}^{3} s_{k+3(j-1)}/\widehat{\varphi}_1$ // Calculate $\varphi_{1,2,3}$ taking $s_1$ into account
    **else**
      $\varphi_j = \sum_{k=2}^{4} s_{k+3(j-1)}/\widehat{\varphi}_1$ // Calculate $\varphi_{1,2,3}$ without taking $s_1$ into account
  $u = argmax_{10\le i\le 16}\{s_i\}$ // Find the maximum value of $s_i$ where $i = 10,\ldots,16$
  **for** *j=1..2* **do**
    $\varphi_{j+3} = \sum_{k=u-1}^{u+1} s_{k+3(j-1)}/\widehat{\varphi}_1$ // Calculate $\varphi_{4,5}$
  $\varphi_6 = rba\{ROI_k(x_n,y_n), hmax_k\}$ // Calculate $\varphi_6$
**Output**: feature vector $\varphi$

Algorithm 1: Algorithm for Feature Vector Calculation.



(a) 3D point cloud measures. (b) 2D depth map. (c) Sample feature vector values.

Figure 3: Example of a scene with eight people. In Subfigure (c), top graphic corresponds to a person 165cm tall and long hair, and the bottom graphic to a person 202cm tall and short hair.

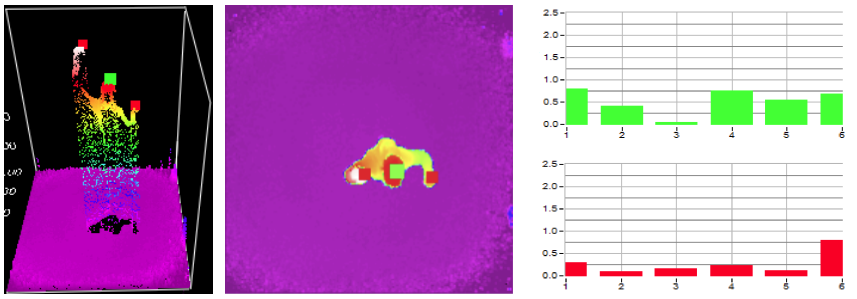


(a) 3D point cloud measures. (b) 2D depth map. (c) Sample feature vector values.

Figure 4: Example of a frame with two people, one of them wearing a hat. In Subfigure (c), top graphic corresponds to the person without hat, and the bottom graphic to a person wearing a hat.

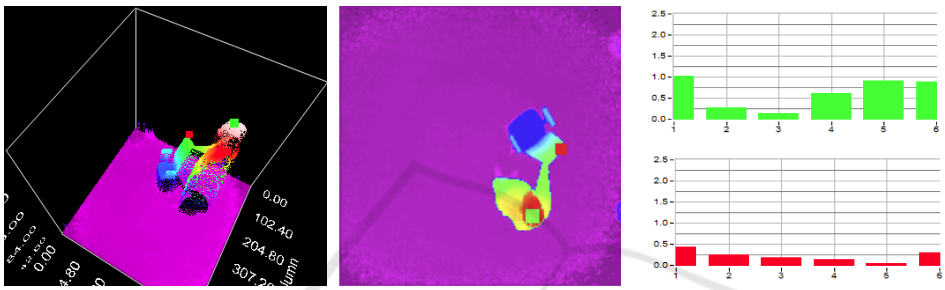
ple with or without accessories (classes 1 and 2, respectively). Some examples of training ROIs for people without accessories are shown in Figure 7, while Figure 8 shows some examples of ROIs for people with accessories (hats and caps in this case).

When we introduced the use of the SVM classifier (more on this below), a new class was added to be able to model a general "non-people" class, comprising partial people ROIs, out of ROIs areas, chairs, floor areas, fists from people in the sequences, etc.



(a) 3D point cloud measures. (b) 2D depth map. (c) Sample feature vector values.

Figure 5: Example of a frame with one person moving his fists up and down. In Subfigure (c), top graphic corresponds to the person, and the bottom graphic to the detected fists.



(a) 3D point cloud measures. (b) 2D depth map. (c) Sample feature vector values.

Figure 6: Example of a frame with one person pushing a chair. In Subfigure (c), top graphic corresponds to the person, and the bottom graphic to the detected chair.

Figure 7: Image samples for the *people without accessories* class (class 1).

Figure 8: Image samples for the *people with accessories* class (class 2).

Figure 9 provides some examples of training material for the non-people class.

Figure 9: Image samples for the *non-people* class (class 3).

## 2.3 Classifier

Two approaches were selected in order to classify each feature vector as corresponding or not to a person: Principal Component Analysis (PCA), and Support Vector Machines (SVM), which will be briefly described next.

### 2.3.1 PCA based Classifier

Our first approach was using a classifier based on Principal Component Analysis (PCA) (Shlens, 2014; Jiménez et al., 2005), due to its simplicity and robustness. This strategy required an offline estimation of the models for each class, prior to the online classification process.

In the offline process, the two transformation matrices required in the PCA strategy are calculated. To do so, a number of training vectors were used, associated to different people representative of each of the two people classes.

The transformation matrices for each class are formed by the eigenvectors associated to the highest eigenvalues of the corresponding scatter matrices (Shlens, 2014; Jiménez et al., 2005). In our case, three eigenvectors have been chosen, following the criterion that the average normalized residual quadratic error (RMSE) is higher than 90%.

In the classification process (online process), the feature vector of each ROI is calculated, and for each class, the difference between this vector and the average vector class is projected in the transformed space.

The projected vector will then be recovered in the original space. The Euclidean distance between the projected and recovered vectors is computed, and referred to as the *reconstruction error*. This process is applied for each of the two classes.

Finally, a feature vector is classified as corresponding to a person if its reconstruction error is lower than a given threshold for any of both transformations (classes). The threshold for each class was determined experimentally for each class, calculated from the average value of the reconstruction error and its its standard deviation (evaluated on the training subset).

### 2.3.2 SVM based Classifier

As an alternative to the PCA classifier described above, we also addressed the use of a SVM as the final people classifier(Burges, 1998), also requiring a supervised training stage.

We initially planned to use a binary SVM (to distinguish between people and non people), but the relatively bad results we obtained in preliminary experiments, lead us to use a multiclass SVM (Burges, 1998; Hsu and Lin, 2002), in which we included classes for people with and without accessories, in addition to the non-people class.

The SVM models were trained from manually selected areas covering a broad range of conditions in what respect to people and non-people characteristics, and their distribution along the recording area.

Preliminary experiments were run in order to decide the SVM kernel type, and the optimal values for the $C$ and *gamma* coefficients. The final configuration used was a radial kernel with $C = 0.5$ and $gamma = 0.00015$.

## 2.4 Tracking System

As shown in Figure 1, the global system includes a final tracking stage, that is executed from the results of the people detector final classifier. This tracking process allows to obtain each detected person trajectory along the video sequence, i.e., its position and speed at each time $t$.

The resulting data from the people detector inform about the number of persons $P_t$ detected in the corresponding image $I_t$, as well as their position $(x_{p,t}, y_{p,t})$, with $p_t = 1..P_t$. These data are used as in a probabilistic filtering and tracking process based on a single particle filter that is thus used for multimodal modeling of the dynamics hypotheses of the people in the image.

A constant speed model is used to perform the probabilistic filtering and tracking, whose state (2) and output (3) equations show that the state vector includes information about both the person position and speed hypothesis $\mathbf{x}_{p,t} = (x_{p,t}, y_{p,t}, v_{y_{p,t}}, v_{y_{p,t}})$, while the output vector just includes the position hypothesis, as it is generated by the people detector, $\mathbf{y}_{p,t} = (x_{p,t}, y_{p,t})$.

$$\mathbf{x}_{p,t+1} = \begin{bmatrix} x_{p,t+1} \\ y_{p,t+1} \\ v_{x_{p,t+1}} \\ v_{y_{p,t+1}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{p,t} \\ y_{p,t} \\ v_{x_{p,t}} \\ v_{y_{p,t}} \end{bmatrix} + \mathbf{w}_{p,t} \tag{2}$$

$$\mathbf{y}_{p,t} = \begin{bmatrix} x_{p,t} \\ y_{p,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_{p,t} \\ y_{p,t} \\ v_{x_{p,t}} \\ v_{y_{p,t}} \end{bmatrix} + \mathbf{v}_{p,t} \tag{3}$$

More specifically, the tracking system used here is based on the eXtended Particle Filter with Clustering Process (XPFCP (Marron et al., 2005)), with a set of $n = 1..N_{Total}$ particles (people hypotheses), from which $N_{New}$ are renewed at each iteration of the filter, and $N_{Save}$ are kept in order to ensure the estimation multimodality and skip the impoverish problem that this proposal may suffer from (A. Doucet, 2001), thus $N_{Total} = N_{Save} + N_{New}$.

The filter is therefore conformed by five stages, as shown in Figure 10, whose functionality is described below:

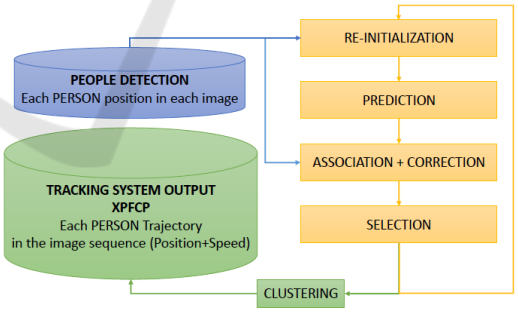



Figure 10: XPFCP functional diagram.

- **Prediction.** Using the state model in equation 2, a propagation from time $t-1$ to time $t$ of all $n = 1..N_{Total}$ hypotheses (particles) state vector $\mathbf{x}_{n,t} = (x_{n,t}, y_{n,t}, v_{y_{n,t}}, v_{y_{n,t}})$ is performed.
- **Association+Correction.** The reliability of each $n = 1..N_{Total}$ particle (person position and speed hypothesis, represented by the state vector) in the tracking is obtained through the particle weight $w_{n,t}$. This is computed with the Mahalanobis distance in the image of the position represented

by the particle output vector (from the predicted state vector $\mathbf{x}_{n,t} = (x_{n,t}, y_{n,t}, v_{y_{n,t}}, v_{y_{n,t}})$ with the output model 3) and the nearest person detection $\mathbf{y}_{p,t} = (x_{p,t}, y_{p,t})$: the smallest distance will give the biggest weight through a Gaussian model of the noise $v_{p,t}$ in equation 3, and thus, the biggest reliability of the hypothesis represented by the corresponding particle $\mathbf{x}_{n,t}$, using a Nearest Neighbor association strategy (Ekman, 2008).

- **Selection.** Using their normalized weights $w_{n,t}$ (with $n = 1..N_{Total}$), the most reliable particles are selected with a residual resampling algorithm (Liu and Chen, 1998), giving a final set of $N_{Save}$ particles, and taking out the least $N_{New}$ reliable ones from the set, that will be substituted by new ones at the next re-initialization filter step.
- **Clustering.** A K-means clustering is then performed over the final $N_{Save}$ set of particles, thus obtaining the result of the global people detection and tracking system: a list of the filtered trajectory estimations for all the persons in the input sequence, represented by the clusters' centroids of the particles' set, i.e., a state vector containing the position and speed of their represented persons in the image $\mathbf{x}_{p,t} = (x_{p,t}, y_{p,t}, v_{y_{p,t}}, v_{y_{p,t}})$. Another NN association process is finally carried out between this global output in $I_t$, and its previous result in $I_{t-1}$, allowing the identification of each person track, with a certainty value.
- **Re-initialization.** Before finishing, the filter prepares the set of particles for its next iteration, recruiting the needed $N_{New}$ hypotheses to complete the $N_{New}$ set. These particles are generated from the people detection output $\mathbf{y}_{p,t} = (x_{p,t}, y_{p,t})$ with $p_t = 1..P_t$ (a set of points next of each of its classification output), increasing its robustness and avoiding the impoverish problem of the multimodal particle filter thanks to the re-initialization strategy in the XPFCP that reinforces the weakest modes in the probabilistic people location density function that the global set represents.

# 3 EXPERIMENTAL WORK

## 3.1 Experimental Setup

In order to provide data for training and evaluating the proposal, we used a preliminary subset of a depth database that is being recorded with a Kinect® v2 device located at a height of $3.4m$[1] The recordings tried to cover a broad variety of conditions, with scenarios comprising:

- Single and multiple people
- Single and multiple non-people (such as chairs)
- People with and without accessories (hats, caps)
- People with different complexity, height, hair color, and hair configuration
- People actively moving and performing additional actions (such as using their mobile phones, moving their fists up and down, etc.).

The data used was split in two subsets, one for training and the other for testing. The subsets are fully independent, so that no person present in the training database was present in the testing subset.

Table 1 and Table 2 show the details of the training and testing subsets, respectively. *#Samples* refers to the number of all the heads over all the frames in the recorded sequences (in our recordings we used 39 different people). The database contains sequences in which the users were instructed on how to move under the camera (to allow for proper coverage of the recording area), and sequences where people moved freely (to allow for a more natural behavior).

The testing subset only included sequences with two or more people (up to eight), and it was further divided in two subsets (C1 and C2), to evaluate the developed systems.

## 3.2 Results and Discussion

Our baseline system was the one based on the PCA classifier and with no tracking stage. In the tables below, **FP** and **FN** are the number of false positives and false negatives respectively, and %**ERR** is the system error rate ($ERR = 100 \cdot [(FP+FN)/\#Samples]$). The tables also include confidence intervals calculated on the *ERR* metric, for a confidence level of 95%.

Table 3 shows the results of our first experiment comparing the performance of the PCA classifier (row *PCA*) to that of the SVM one (row *SVM*), using testing subset C1. From the table, it can be clearly seen that the SVM classifier is much better at accurately modeling people: The simple linear approach by the PCA strategy is not able to cope with the variability of people characteristics and varying positions along the recorded space.

Table 4 shows the effect of using the tracker (row *PCA+XPF*) as compared to only using the PCA classifier (row *PCA*), using testing subset C2. In this case, it's also clear that the tracker provides an improvement as compared to the baseline system, although

[1]The GOTPD1 database ( (Macias-Guarasa et al., 2016), that is available to the academic community for research purposes .

Table 1: Details of the training subset.

| **Sequence ID** | *#Samples* | **Description** | **Class** |
|---|---|---|---|
| S0041→S046 | 1682 | Single person | Class 1: Person without accessories |
| S0047→S0048 | 373 | Multiple people with accessories (hats, caps) | Class 2: Person with accessories (hats, caps) |

Table 2: Details of the testing subsets.

| **Testing subset** | *#Samples* | *#Class*1 | *#Class*2 |
|---|---|---|---|
| C1: Sequences with two or more people | 8592 | 7013 | 1579 |
| C2: Sequences with two or more people | 9510 | 7762 | 1748 |

Table 3: Comparison betwen the PCA and SVM classifiers (using testing subset C1).

| **Classifier** | **FN** | **FP** | **%ERR** |
|---|---|---|---|
| *PCA* | 756 | 25 | $9.09 \pm 0.61$ |
| *SVM* | 335 | 27 | $4.21 \pm 0.42$ |

Table 4: Comparison betwen the use of not of the tracker (using testing subset C2).

| **Tracking use** | **FN** | **FP** | **%ERR** |
|---|---|---|---|
| *PCA* without tracking | 355 | 4 | $3.77 \pm 0.38$ |
| *PCA* plus tracking | 335 | 9 | $3.60 \pm 0.37$ |

the differences are not statistically significant. This result is consistent with the idea that the people detection process is very accurate, so that the tracking stage can only achieve minor improvements, specially in the reduction of false negatives, at the expense of a very slight increase in the number of false positives. Both effects are due to the ability of the tracker to provide additional hypothesis that the people detector could not generate (due to either misclassifications or occlusions in the depth image).

# 4 CONCLUSIONS

In this work, we proposed a system for the robust detection of people in depth images, captured by an overhead ToF camera. The proposal comprises several stages, and it allows achieving the detection of multiple people in the scene in a robust way.

First, the isolated maximums in the scenes are detected. Then, a Region of Interest (ROI) is precisely defined around each maximum, and from the pixels included in the ROI, a 6-component feature vector is extracted, with their component values related to the number of pixels in given areas of the ROI. The selected feature vector has proved its efficacy for properly characterizing the people upper body geometry.

For the feature vector classification, we have presented two alternatives, based on PCA and SVM respectively. The obtained results show that the SVM classifier exhibits a higher performance, as the PCA based strategy is not able to cope with the high variability of people and scene characteristics.

The proposal also includes a particle filter tracker to allow people identification and tracking. The performance impact of the tracker have been analyzed by comparing the results with and without this stage. The results demonstrate that the incorporation of the tracking stage not only allows to have information about the trajectory and velocity of each person, but also improve the detection results, reducing the error rate.

Future work will include a more exhaustive experimental work, exploiting more sophisticated classification strategies, and applying the system to actual people counting solutions in realistic scenarios.

# ACKNOWLEDGEMENTS

This work has been supported by the Spanish Ministry of Economy and Competitiveness under projects SPACES-UAH (TIN2013-47630-C2-1-R) and HEIMDAL (TIN2016-75982-C2-1-R), and by the University of Alcalá under projects SCALA (CCG2016/EXP-010), DETECTOR (CCG2015/ EXP-019) and ARMIS (CCG2015/EXP-054).

# REFERENCES

A. Doucet, N. de Freitas, N. G. (2001). *Sequential Monte-Carlo Methods in Practice*. Springer Verlag.

Antic, B., Letic, D., Culibrk, D., and Crnojevic, V. (2009). K-means based segmentation for real-time zenithal people counting. In *Proc. of the 16th IEEE International Conference on Image Processing*, ICIP'09, pages 2537–2540.

Bar-Shalom, Y., Willett, P. K., and Tian, X. (2011). *Tracking and data fusion*. YBS publishing.

Bevilacqua, A., Di Stefano, L., and Azzari, P. (2006). People tracking using a time-of-flight depth sensor. In *IEEE International Conf. on Video and Signal Based Surveillance. AVSS '06.*, pages 89–89.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Cai, Z., Yu, Z. L., Liu, H., and Zhang, K. (2014). Counting people in crowded scenes by video analyzing. In *Industrial Electronics and Applications (ICIEA), 2014 IEEE 9th Conference on*, pages 1841–1845.

Dan, B.-K., Kim, Y.-S., Suryanto, Jung, J.-Y., and Ko, S.-J. (2012). Robust people counting system based on sensor fusion. *IEEE Trans. on Consumer Electronics*, 58(3):1013–1021.

Del Pizzo, L., Foggia, P., Greco, A., Percannella, G., and Vento, M. (2016). Counting people by rgb or depth overhead cameras. *Pattern Recognition Letters*.

Ekman, M. (2008). Particle filters and data association for multi-target tracking. In *Information Fusion, 2008 11th International Conference on*, pages 1–8.

Galčík, F. and Gargalík, R. (2013). Real-time depth map based people counting. In *International Conf. on Advanced Concepts for Intelligent Vision Systems*, pages 330–341. Springer.

Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE trans. on Neural Networks*, 13(2):415–425.

Isard, M. and Blake, A. (1998). Condensation - conditional density propagation forvisual tracking. *International Journal of Computer Vision*, 29(1):5–28.

Jeong, C. Y., Choi, S., and Han, S. W. (2013). A method for counting moving and stationary people by interest point classification. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 4545–4548.

Jia, L. and Radke, R. (2014). Using time-of-flight measurements for privacy-preserving tracking in a smart room. *IEEE Trans. on Industrial Informatics*, 10(1):689–696.

Jia, Z., Balasuriya, A., and Challa, S. (2008). Autonomous vehicles navigation with visual target tracking: Technical approaches. *Algorithms*, 1(2):153–182.

Jiménez, J. A., Mazo, M., Ureña, J., Hernández, A., Alvarez, F., García, J. J., and Santiso, E. (2005). Using PCA in time-of-flight vectors for reflector recognition and 3-D localization. *IEEE Trans. on Robotics*, 21(5):909–924.

Liu, J. S. and Chen, R. (1998). Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93:1032–1044.

Luna, C. A., Losada-Gutierrez, C., Fuentes-Jimenez, D., Fernandez-Rincon, A., Mazo, M., and Macias-Guarasa, J. (2016). Robust people detection using depth information from an overhead time-of-flight camera. *Expert Systems with Applications*, pages –.

MacCormick, J. and Blake, A. (2000). A probabilistic exclusion principle for tracking multiple objects. *International Journal of Computer Vision*, 39(1):57–71.

Macias-Guarasa, J., Losada-Gutierrez, C., Fuentes-Jimenez, D., Fernandez, R., Luna, C. A., Fernandez-Rincon, A., and Mazo, M. (2016). The GEINTRA Overhead ToF People Detection (GOTPD1) database. http://www.geintra-uah.org/datasets/gotpd1. (accessed June 2016).

Marron, M., Garcia, J. C., Sotelo, M. A., Fernandez, D., and Pizarro, D. (2005). "xpfcp": an extended particle filter for tracking multiple and dynamic objects in complex environments. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2474–2479.

Marron, M., Garcia, J. C., Sotelo, M. A., Pizarro, D., Mazo, M., Canas, J. M., Losada, C., and Marcos, A. (2010). Stereo vision tracking of multiple objects in complex indoor environments. *Sensors*, 10(10):8865.

Matzner, S., Heredia-Langner, A., Amidan, B., Boettcher, E., Lochtefeld, D., and Webb, T. (2015). Standoff human identification using body shape. In *Technologies for Homeland Security (HST), 2015 IEEE International Symposium on*, pages 1–6.

Ramanan, D., Forsyth, D. A., and Zisserman, A. (2006). Tracking People by Learning Their Appearance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(1):65–81.

Rauter, M. (2013). Reliable human detection and tracking in top-view depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 529–534.

Sell, J. and O'Connor, P. (2014). The Xbox one system on a chip and Kinect sensor. *Micro, IEEE*, 34(2):44–53.

Shlens, J. (2014). A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100. (accessed June 2016).

Stahlschmidt, C., Gavriilidis, A., Velten, J., and Kummert, A. (2014). Applications for a people detection and tracking algorithm using a time-of-flight camera. *Multimedia Tools and Applications*, pages 1–18.

Zhang, X., Yan, J., Feng, S., Lei, Z., Yi, D., and Li, S. Z. (2012). Water filling: Unsupervised people counting via vertical kinect sensor. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 215–220. IEEE.

Zhu, L. and Wong, K.-H. (2013). Human tracking and counting using the kinect range sensor based on adaboost and kalman filter. In *International Symposium on Visual Computing*, pages 582–591. Springer.