Multimedia Tools and Applications 78 (2019): 14109–14127

DOI: https://doi.org/10.1007/s11042-018-6822-7

Incorporating Wheelchair Users in People Detection

Rafael Martín-Nieto · Alvaro García-Martín · José M. Martínez

Abstract A wheelchair users detector is presented to extend people detection, providing a more general solution to detect people in environments such as houses adapted for independent and assisted living, hospitals, healthcare centers and senior residences. A wheelchair user model is incorporated in a detector whose detections are afterwards combined with the ones obtained using traditional people detectors (we define these as standing people detectors). We have trained a model for classical (DPM) and for modern (Faster-RCNN) detection algorithms, to compare their performance. Besides the extensibility proposed with respect to people detection, a dataset of video sequences has been recorded in a real in-door senior residence environment containing wheelchairs users and standing people and it has been released together with the associated ground-truth.

Keywords People detection · Wheelchair users · Assisted living · Independent living · Healthcare system

1 Introduction

In health care centers, senior residences, hospitals, etc., it is usual to see people who need wheelchairs and their detection is useful to monitor them and to provide them assistance in case they need. Knowing the location of a wheelchair user can be useful for some healthcare applications (e.g. monitoring), and it can be used to analyze the behaviour and actions of such users in different environments. The automatic detection of mobility impaired people, including wheelchair users, is also an important problem for Intelligent Transportation Systems (ITS) in public traffic areas [1]. Many assistance

This work has been partially supported by the Spanish government under the project TEC2014-53176-R (HAVideo) and by the Spanish Government FPU grant programme (Ministerio de Educación, Cultura y Deporte).

Rafael Martín-Nieto
Video Processing and Understanding Lab (VPULab), Universidad Autónoma de Madrid, Spain, 28049
Tel.: +34-914972260
E-mail: rafael.martinn@uam.es

applications that can be derived from the automatic wheelchair users detection (e.g., doors, elevators, escalators), can automatically activate a special operation mode for such people after detecting them, the green-light time can be increased in pedestrian crossing with traffic lights when a wheelchair user is detected, etc. All these events could be activated manually by one person, but, if automatic activation is achieved, people in wheelchairs will feel more comfortable and these events would become something natural, and the operation would not need human agents for correct functioning.

Another application for which the presented detector is useful is independent living. According to the definition given by the World Institute on Disability (`http://www.wid.org/`), independent living is defined as allowing people with disabilities to have the same level of choice, control and freedom in their daily lives as anyone else. In the context of caring for the elderly, independent living is seen as a continuum care, whose next step would be the incorporation to a nursing home. The proposed model detector is useful for both stages, first to monitor the wheelchair user in their domestic environment ensuring that everything runs properly, and then to video monitor people in a nursing home, allowing to detect interesting events such as fall detections [2,3].

2 State of the art

In this section we present an overview of works related with the presented detector. First, some works related to the standing people detection problem are presented, and how it has been solved from different viewpoints. After that, the different previous wheelchair users detections approaches are classified and described. Finally, the selected solutions are presented.

2.1 Standing people detection

In computer vision, standing people detection can be considered as a two steps process [4,5]. First, it is necessary to localize the initial objects candidates to be standing person in the scene. The two most common approaches to localize those objects are those based on some kind of segmentation of the scene in foreground (objects) and background [6] and those based on a scanning approach [7,8]. In general, those algorithms based on a scanning approach have been proved to be more robust to real and more complex sequences where there are several background and people variabilities [7,9,10, 11]. There are also some approaches that try to combine both approaches together [12, 13].

The second step in any standing people detection can be considered as a standard pattern recognition issue. In this case, it is necessary to previously define a standing person model and then classify any new candidate selected during the previous step as a standing person or not. The classification process will be characterized according to the chosen standing person model. Therefore, standing people detection approaches can be classified into two groups, namely, holistic and part-based detectors, depending on the model properties. The holistic detectors define the person as a region or shape [14, 15,16,17], whilst the part based detectors define the person as combination of multiple regions or shapes [8,18]. In general, those algorithms based on part-based models are able to deal with partial occlusions better than those based on a holistic model, but significantly increasing the model complexity.

In recent years the object detection results (and therefore people detection results) have been greatly improved thanks to the use of deep learning algorithms. Some examples of these algorithm are [19], [20] or [21].

2.2 Wheelchair users detection

There are some works in the state of the art trying to address the wheelchair users detection problem. These works can be classified into two main groups. The first group focuses on detecting ellipses which correspond to the wheelchair wheels. The second group is based on detecting the wheelchair users using discriminative features, usually color and Histogram of Oriented Gradients (HOG).

The first approach of the works that try to find the wheel ellipses is presented in [22]. The model considered here is based on two wheels with a head over them. The wheels are detected using the Hough transform to detect ellipses in an edge image obtained via the Canny detector. The head is found using a skin detector. All these stages are performed after a background subtraction. In [23], the detection is based only in determining the location and orientation of the wheels, proposing a mathematical method of ellipse-circle geometry. [24] follows the work presented in [22] and includes tracking and event detection. In this case, Zimmer frames are also detected. The location of doors is also used for the detections. The wheelchair users detector presented in [25] starts from a background subtraction stage, similar to [22]. After obtaining the foreground, the resulting bounding boxes are analyzed locating the wheel, and then the user and the assistant (if any). A novel idea is presented in this work, which is to recognize whether an assistant is pushing the wheelchair.

On the other hand, the second group aims to find discriminative features to detect the wheelchair user. Similar to other studies, [26] starts with a background subtraction. This solution is based on detecting wheelchair user parts (e.g., head, chest, legs) and wheelchair parts. Finding each part is based on color, which is previously defined. After that, the object position is obtained using a stereo vision camera. The justification for not trying to locate the wheels is that there are orientations (front and rear) in which they are unobservable. Besides, there are different wheelchair models, especially electrical, which do not have large wheels as conventional wheelchairs. The recognition proposed in [1] also uses stereo vision cameras. The feature used is HOG, allowing discrimination between standing people and wheelchair users thanks to a previously trained Support Vector Machine (SVM). The detector proposed in [27] considers two descriptors, HOG and Contrast Context Histogram (CCH), which are adopted to model, respectively, the shape and appearance of the wheelchair. An AdaBoost learning stage selects the features which better discriminate the object. All the possible wheelchair orientations are classified in 8 different models composing a state graph whose elements can change to adjacent orientation models. A Gaussian pyramid is constructed to overcome the scale problem by downsampling the image from the original resolution. The approach proposed by [28] focuses on a dimensionality reduction using sparse representation to improve the generalization capability. To characterize the wheelchair users, directional maps are defined by determining the dominant direction of motion in each local spatiotemporal region.

2.3 Discussion

The proposed wheelchair users detectors approach (see section 3.1) has advantages over the previously existing solutions: it does not need a background model for background subtraction, it can detect wheelchairs in any orientation, it does not need to know the dimension of some parts of the wheelchair, it does not need stereo vision cameras, it does not need to know the wheelchair colors in advance, and it does not consider the wheelchair user as a rigid object, allowing deformations.

We have chosen a scanning approach with a part-based model (DPM), and a deep learning approach (Faster-RCNN) for the object detection algorithms. We have chosen these two detection algorithms as the first one, DPM, is a classic algorithm, based on HOG filters, that offers good detection after the great improvement of the detection algorithms in the last years, and the second one, Faster-RCNN, to observe the operation of the proposed technique using one of the most modern and effective algorithms of the state of the art, based on neural networks.

3 Detection approach

This section describes the original detection algorithms (see section 3.1), whose training method is used to generate the wheelchair users detection models (see section 3.2). The detections from the different models (standing people and wheelchair users) are combined for an integrated detection (see section 3.3).

3.1 Detection algorithms

The first considered detection algorithm is the Deformable Parts Model (DPM) detector [8]. The DPM detector is based on exhaustive search and a part-based person model. It is a part-based adaptation of the original Histogram of Oriented Gradients detector (HOG) [14]. It proposes an object detection system based on mixtures of multiscale deformable part models where each deformable body part is modeled as the original HOG detector [14]. The algorithm model also contains the flip (horizontally mirrored) of the model.

The second detection algorithm chosen for the experiments of the proposed system is the Faster RCNN (Regions with Convolutional Neural Network Features) [21] detector, which consist in a more efficient variation, mainly in terms of computational cost but also in performance, of the previous versions R-CNN [29] and Fast R-CNN [30] detectors. The three variations have in common the combination of bottom-up region proposals with rich features computed by a convolutional neural network. The main difference of the Faster-RCNN is the use of a Region Proposal Network (RPN) that enables nearly cost-free region proposals.

The computational cost of the detections is not treated in this paper as this aspect is analyzed by the authors of DPM ([8]) and Faster-RCNN ([21]). The used DPM approach is implemented with MATLAB and the computational cost is about 2 seconds per frame, considering an image of 352 $\times$ 288 pixels. Note that there is also a faster implementation in OpenCV that increases the detection time to about 1 second per frame. The used Faster RCNN approach is implemented with MATLAB and Caffe,

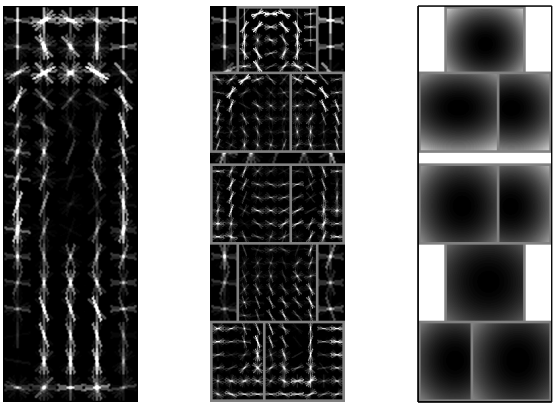

Fig. 1 DPM standing people model. The three columns are, from left to right, root model, parts model and parts deformation.

and the computational cost is about 150-200 milliseconds per frame (Faster RCNN, VGG-16 with GPU), considering an image of 500x375 pixels.

3.2 Detection models

This subsection adds some details about the two different trained algorithm models. The standing people has not been trained in this work, but it is presented here for comparing it with the wheelchair users model.

Figure 1 shows a visual example of the DPM person model, namely the INRIA person model, extracted from [31]. The model also contains the flip of the model, but it has not been included in the figure as it does not provide additional information different from the data already shown.

Following the original people detection algorithm, we train a wheelchair users detector model.

To generate the wheelchair users model, we used the annotations of the training set from the Smile Lab training dataset (see subsection 4.1.1), containing 3674 positive examples. For the negative examples set, we used the standing people model negative examples from [31]. For this purpose, we ensure that this image set does not contain any pictures with a wheelchair nor a wheelchair user.

For the DPM standing people detector model, there is just one model variation as the appearance from the different points of view are similar. Unlike the standing people model, a model with two variations is trained for the wheelchair users, as it is considered that the appearance of the front and side wheelchair users are different enough to be independent in their appearance classification. We have also performed experiments testing from 2 to 8 model variations, obtaining very similar or worse results, due to the overfitting of the model to the training data. Figure 2 shows the resulting wheelchair user model. The trained model also contains the flip of each model (as the original

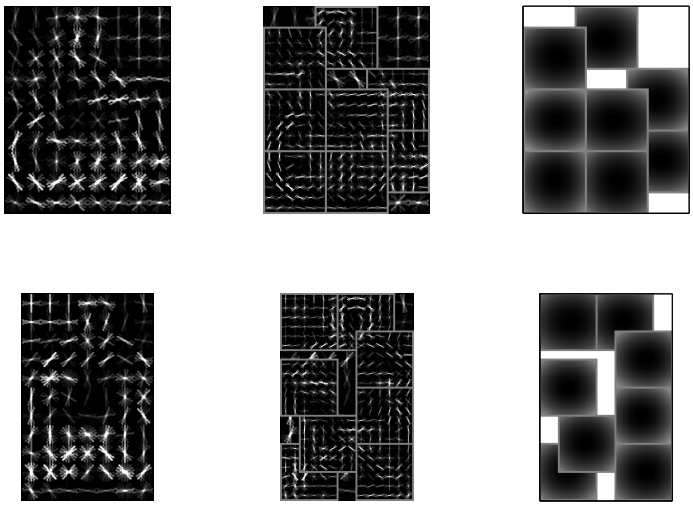

Fig. 2 DPM wheelchair user model. Each row represents a model variation. The three columns are, from left to right, root model, parts model and parts deformation.

people detector model), but it has not been included in the figure as again it does not provide additional information different from the data already shown.

For the Faster-RCNN detector, and according to the author's results [21], we have chosen the pre-trained network VGG-16 model [32] that has 13 convolutional layers and 3 fully-connected layers. We have retrained the network using the PASCAL VOC 2007 and 2012 datasets, and we have added a new object class, the wheelchair user object, using the same positive and negative examples than for the DPM model training. The Faster-RCNN model does not have a graphic representation as in the case of the DPM model.

This wheelchair users models are available for research purposes in the Wheelchair users dataset webpage (`http://www-vpu.eps.uam.es/DS/WUds/`).

3.3 Detectors combination

The DPM wheelchair user detections and the standing people detections are combined to obtain the general people detections. All the detections from each detector are maintained as we consider that each detector works for disjoint people models. As each detector has a different Standing People (SP) / Wheelchair User (WU) Detection Confidence output space or range $C_{SP/WU}$ (see Figure 4), in order to add the outputs from both detectors (each output is a set of bounding boxes, each of them with an associated confidence), it is necessary to normalize both confidence outputs. Therefore, we normalize both detectors, C_{SP} ($0 \leq C_{SP} \leq 1$) and C_{WU} ($0 \leq C_{WU} \leq 1$). The normalization is performed according to the probability density function (pdf) of each Detection Confidence. In particular, the Standing People Detection Confidence

distribution has been estimated using the detector output over the INRIA dataset [14], whilst the Wheelchair Users Detection Confidence distribution is obtained detecting the wheelchair users from the training images set. Using the score histogram, the pdf is estimated trying to adjust properly to the obtained scores. The score histogram and the estimated pdf are shown in Figure 3.

In order to facilitate comparison between models, pdf and cdf (cumulative distribution function) are represented in Figure 4 for both standing people (from [33]) and wheelchair users models. As the considered detection algorithm is the same for both models, the density functions obtained are relatively close, but this conversion should be performed to join the detectors results rigorously. After normalizing the detections of the different models, both sets of detections are joined together to obtain the complete set that considers the different people appearances.

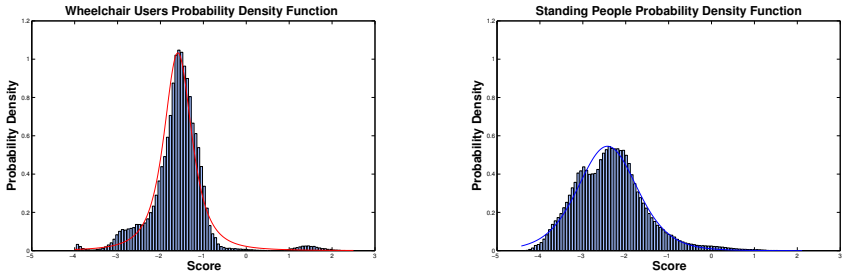

Fig. 3 DPM wheelchair user (left) and standing people (right, extracted from [33]) models score histograms with the fitted pdfs.

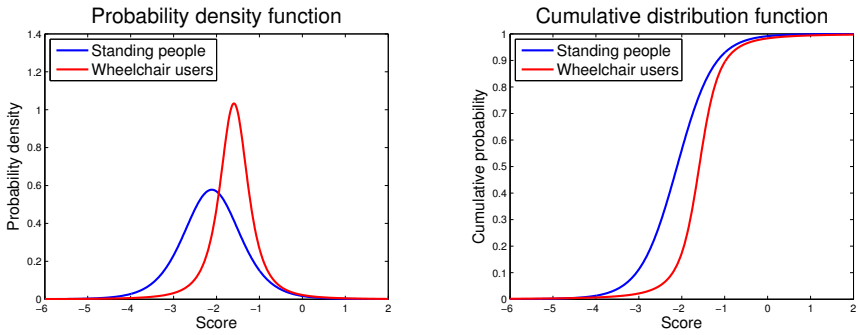

Fig. 4 DPM Standing people and wheelchair user detectors pdf (left) and cdf (right).

The Faster-RCNN output detections are by default normalized between 0 and 1 in the algorithm, so this step does not apply to its results as the normalization is internally included in the algorithm.

Sequence number	#Frames	#Wheelchair users	#Standing people
1	449	1	From 3 to 5
2	351	1	From 2 to 5
3	239	1	From 4 to 5
4	287	1	From 3 to 7

Table 1 Properties of each of the sequences from the Smile dataset.

4 Experimental setup

4.1 Datasets description

This section contains the two different datasets used for the experimental validation and the evaluation metrics. The SmileLab wheelchair dataset [27] was used for the models generation (see section 3.2) and validation (see section 5.1). The Wheelchair Users dataset was used to check the generated model in a different and independent scenario.

4.1.1 Smile Lab wheelchair dataset

This dataset was created by the Smile Lab (`http://smile.ee.ncku.edu.tw/`) at the Department of Electrical Engineering, National Cheng Kung University, Taiwan. The dataset is divided into two main image sets: the train sequences and the test sequences. Each of the frames has a resolution of 720x480 pixels.

The training sequences are composed of 8 image subsets and a total of 3674 images, each one of them contains a set of images of wheelchairs with a defined orientation relative to the camera. The different orientations and models are shown and defined in [27].

The test sequences are composed of 4 image subsets, each one of them containing a sequence with a wheelchair and some standing people walking around. Unlike the training set, each of these frame subsets contains a continuous recording, allowing to use tracking techniques to improve detection, as shown in [27]. The test set contains a total of 1314 frames divided in 4 folders. Table 1 shows the properties of each sequence.

The ground truth of this dataset was not available, so we created it annotating manually each of the frames from both sets. This ground truth is available for downloading as additional content in the Wheelchair users dataset webpage (`http://www-vpu.eps.uam.es/DS/WUds/`).

4.1.2 Wheelchair Users dataset

This dataset was recorded by the Video Processing and Understanding Lab due to the lack of public wheelchair datasets. We used it to test the trained wheelchair users detector, as it contains sequences with a higher number of wheelchairs (up to four) and some more complex situations and scenarios (illumination changes, occlusions, etc.). The sequences were recorded in a real environment of a senior residence, in order to work with an environment as realistic as possible (due to privacy issues, real recording with actual residents was not possible). Each of the frames has a resolution of 768x432 pixels and the sequences are recorded at 25 fps. Compared to the other dataset, this

one contains a new environment with a larger number of sequences, a greater number of frames per sequence, and more wheelchair types (three different wheelchairs).

The dataset consists of 11 sequences (S1 to S11), each of them recorded from two points of views (V1 and V2), resulting in a total of 22 sequences. Table 2 shows the properties of each recorded sequence.

All sequences were recorded in the same room, using two GoPro cameras (HERO3 White edition). The fisheye effect was corrected using the GoPro Studio software tool. Each camera views are shown in Figure 5 and a room top view map is shown in Figure 6.

This dataset and its annotated ground truth are publicly available for research purposes. The ground truth of this dataset was manually annotated for each frame of each sequence. The annotated ground truth considers the wheelchair users and the standing people present in every frame, even if they are highly occluded.

Fig. 5 Camera views of the Wheelchair Users dataset. Left: viewpoint 1. Right: viewpoint 2.

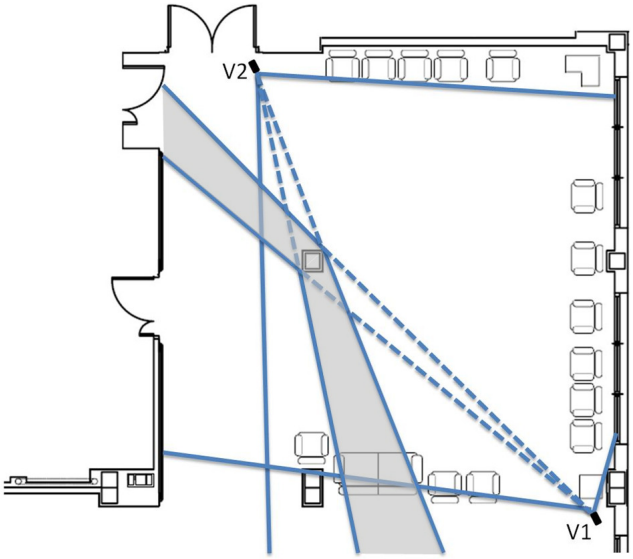

Fig. 6 Top view map of the Wheelchair Users dataset. V1 and V2 represent camera 1 and camera 2 locations and fields of view.

Sequence number	#Frames	#Wheelchair users	#Standing people
1	1318	1	0
2	916	1	0
3	860	1	1
4	1167	1	1
5	1638	2	0
6	723	2	0
7	1082	2	2
8	743	2	2
9	2102	2	2
10	2460	2	2
11	1855	4	0

Table 2 Properties of each of the recorded sequences from the Wheelchair Users dataset.

4.2 Evaluation metrics

In order to evaluate the proposed approach, we quantify the performance results. Global sequence performance is usually measured in terms of Precision-Recall (PR) curves [16, 34, 35]. These curves compare the similarities between the output and ground truth bounding boxes. For each value of the detection confidence or score, Precision-Recall curves are computed:

$$Precision = \frac{\#TPPD}{\#TPPD + \#FPPD} \tag{1}$$

$$Recall = \frac{\#TPPD}{\#TPPD + \#FNPD} \tag{2}$$

Where TPPD are True Positive People Detections, FPPD are False Positive People Detections, and FNPD are False Negative People Detections.

In addition, in order to evaluate not only the yes/no detection decision but also the precise people locations, we take into account the three evaluation criteria defined in [36], that allow to compare hypotheses at different scales: relative distance (dr), cover and overlap. A detection is considered true if $dr \leq 0.5$ (corresponding to a deviation up to 25% of the true object size) and cover and overlap are both above 50%.

The integrated Average Precision (AP) is generally used to summarize the algorithm performance in a single value, represented geometrically as the area under the PR curve (AUC-PR). In order to approximate the area correctly, we use the approximation described by [37].

5 Experimental validation

The detectors are run on the evaluation datasets in order to analyze their performance. As the wheelchair users models were trained using the dataset presented in subsection 4.1.1, the results obtained on its test images are expected to be better than the results obtained on the images of the dataset presented in subsection 4.1.2, as it is a completely independent scenario with different wheelchairs of those used to train the model.

This section contains results of the detector over the Smile Lab dataset (see subsection 5.1) and over the Wheelchair Users dataset (see subsection 5.2).

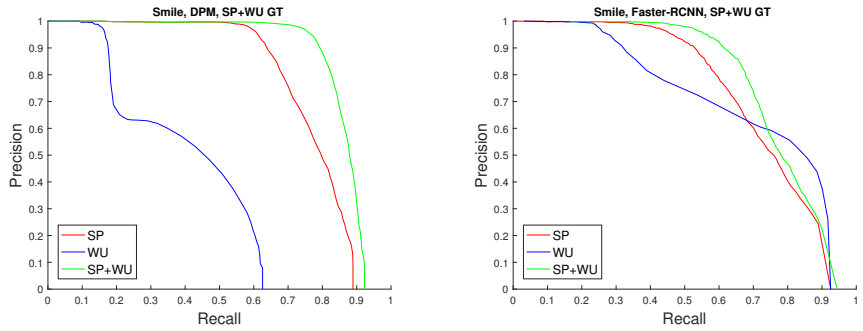

Fig. 7 Precision vs Recall detection curves for the Smile Lab dataset test sequences using complete (standing people, SP, and, wheelchair users, WU) ground truth.

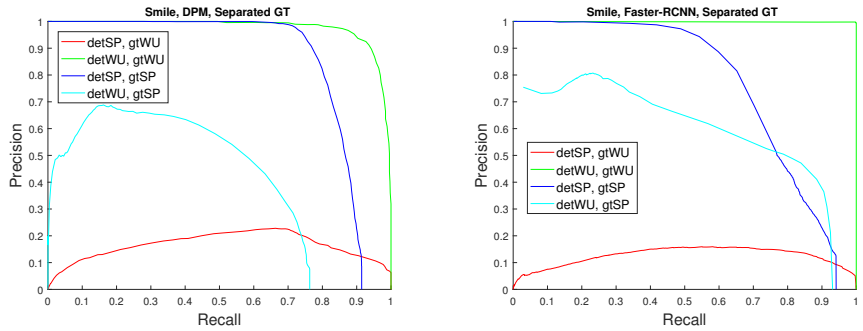

Fig. 8 Precision vs Recall detection curves for the Smile Lab dataset test sequences using separated detection results and separated ground truth. detSP corresponds to the Standing Person model detections, detWU corresponds to the Wheelchair Users model detections, gtSP corresponds to the Standing Person ground truth, and gtWU corresponds to the Wheelchair Users ground truth.

		Ground Truth All (SP+WU)			
		Smile		WUds	
		DPM	Faster-RCNN	DPM	Faster-RCNN
Detector	SP	0,777	0,734	0,577	0,688
	WU	0,405	0,712	0,637	0,735
	SP+WU	**0,864**	**0,777**	**0,733**	**0,811**

Table 3 Detectors AUC using complete (standing people, SP, and, wheelchair users, WU) ground truth.

5.1 SmileLab dataset results

Figures 7 and 8 show the resulting precision-recall detection curves obtained for the detection on the Smile Lab dataset test sequences. Table 3 presents the numerical AUC values of the precision-recall detection curves. All these curves are also available for downloading in the publication webpage (`http://www-vpu.eps.uam.es/publications/IntegratingWheelchairUsersInPeopleDetection/`).

The combination of the detection results of both models (standing person and wheelchair user models) improves the results of each model separately, for both detec-

			Ground Truth			
			Smile		WUds	
			SP	WU	SP	WU
Detector	DPM	SP	0,852	0,163	0,883	0,283
		WU	0,415	0,977	0,265	0,833
	Faster-RCNN	SP	0,767	0,124	0,728	0,391
		WU	0,599	0,999	0,268	0,912

Table 4 Detectors AUC using separated detection results and separated ground truth.

	H	M	F	R	P
[27]	1086	83	73	0.929	0.937
DPM	1218	96	99	0.927	0.925
Faster-RCNN	1314	0	7	1	0.995

Table 5 Comparative results for the wheelchair users detections between [27] and our approaches. H, M, F, R and P are, respectively, hits, miss detects, false detects, recall and precision.

tion algorithms (DPM and Faster-RCNN), as seen in Table 3, in which the area under the curve of the combination of models is better than the detection of each model separately, for both detection algorithms. The final results obtained by the DPM detector are better than those obtained by the Faster-RCNN, but it is probably due to overfitting, as the result of the Faster-RCNN is better when using a different dataset for evaluation (see the following section). With respect to the results using separated ground truth (Table 4), the DPM detector is able to better detect standing people, but wheelchair users are better detected by the Faster-RCNN model. The proposed detection improves the initial performance 11,3% and 5,8% on average for this dataset. Note that the training images and the test images are different but contain the same (people and wheelchair model) standing people and wheelchair users.

The obtained results can not be directly compared with the results presented in [27] for several reasons. Only the wheelchair users are detected in [27], while we detect both wheelchair users and standing people, but for this comparative we will use only the wheelchair users model detections. Also they consider a detection error when a wheelchair is detected with an orientation different from the one (between eight possible orientations) annotated in the ground truth. The work presented in this paper does not consider the wheelchair orientation, as defined in previous sections. The authors of this dataset did not provided us the ground truth that they had used, so we had to generate a new one, as commented in subsection 4.1.1. Table 5 shows the results given by [27] and our wheelchair users detection results. We have selected the closest point between our precision-recall curve and the point given by the authors of the dataset. The results obtained by the DPM detector are very close to those presented in [27] but slightly worse, and the results obtained by the Faster-RCNN are significantly better, especially highlighting the null value of miss detections in all sequences. It is noteworthy that our ground truth has more frames annotated than the results given by [27] (1314 vs 1169 frames). Our ground truth has annotations of every sequences frames, regardless of it complexity, the existence of occlusions, etc. Our work does not tries to improve the detections results of [27], as we present a different approach integrated into a complete system, but the result is greatly improved in the case of the Faster-RCNN detector.

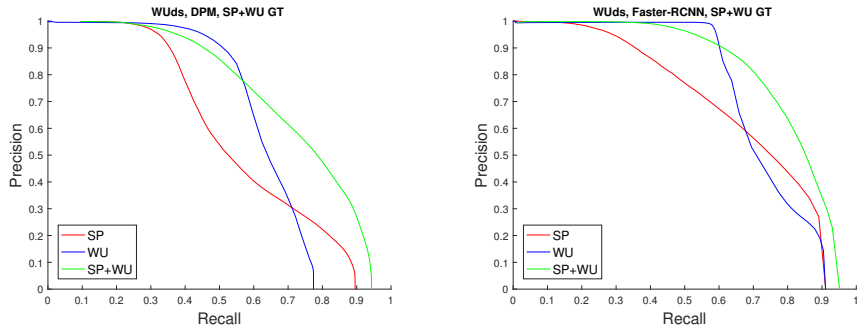

Fig. 9 Precision vs Recall detection curves for the Wheelchair Users dataset using complete (standing people, SP, and, wheelchair users, WU) ground truth.

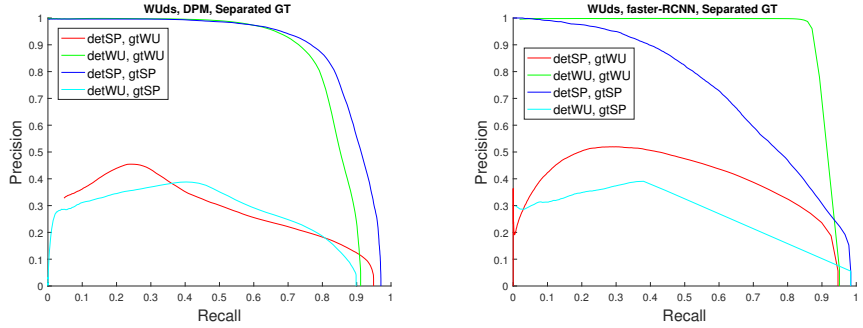

Fig. 10 Precision vs Recall detection curves for the Wheelchair Users dataset sequences using separated detection results and separated ground truth. detSP corresponds to the Standing Person model detections, detWU corresponds to the Wheelchair Users model detections, gtSP corresponds to the Standing Person ground truth, and gtWU corresponds to the Wheelchair Users ground truth.

5.2 Wheelchair Users datasets results

Figures 9 and 10 show the resulting precision-recall detection curves obtained for the detection on Wheelchair Users dataset sequences. Table 3 presents the numerical AUC values of the precision-recall detection curves. All the obtained curves are available in the publication webpage (`http://www-vpu.eps.uam.es/publications/IntegratingWheelchairUsersInPeopleDetection/`).

In the Wheelchair Users dataset sequences there is a greater number of wheelchair users, both in absolute value (greater number of wheelchair users in the sequences) and relative value (wheelchair users vs standing people ratio), so it is expected to get a greater improvement with respect to the original standing people detector. In this case the percentage increase of the AUCs, compared to the initial detector, is 27% and 17,9% on average, much higher than the 11,3% and 5,8% obtained in the previous dataset. The Faster-RCNN detector performance is better in this dataset than in the one discussed in the previous section. With respect to the evaluation with partial ground truths (Table 4), the results obtained with the WUds are The combination of the detection results of both models (standing person and wheelchair user models) improve the results of each

model separately, for both detection algorithms (DPM and Faster-RCNN), as seen in Table 3, in which the area under the curve of the combination of models is better than the detection of each model separately, for both detection algorithmssimilar to those observed with the Smile dataset.

The transfer learning to the new sequences is generic enough to improve the results, reaching in fact a higher percentage increase in the recorded sequences than for the Smile dataset sequences when using the Faster-RCNN algorithm. . The new recorded scenario dataset presents a realistic scenario for the detectors, where not all the wheelchair types can be considered in the model, in the same way as in the standing people detector not every person, orientation and pose are present. The recorded sequences also contains severe illumination changes and occlusions.

6 Conclusion

In this paper, we treat the problem of different appearances for the same semantic object class detection. Typical senior residences scenarios are an example of this problematic situation. In particular, our main objective is to detect both standing people and wheelchair users simultaneously. For this reason, an extension of people detection that allows to detect people with the need of using a wheelchair has been presented. We have trained two additional wheelchair users detectors models whose detections can be combined with the detections obtained using the traditional standing people detectors models, providing generality and supplementary detection capacity. This approach can not only be applied to the case of wheelchairs but the ideas exposed here can be extrapolated to other scenarios where there are individuals with an appearance different from the standard, as Zimmer frames users or people using walking sticks.

Due to the appearance of wheelchairs, we have trained a model with two different variations (front/rear and side point of views), allowing to detect different orientations. The proposed detector does not consider the wheelchair orientation in the output, but we consider that this does not provide much information for the different applications derived from the detection. In any case, if the orientation estimation were interesting or necessary, the wheel ellipse can be located in the detection bounding box after detecting the wheelchair, following one of the existing methods.

Due to the absence of public datasets with this type of content, new sequences with greater complexity have been recorded in order to test the designed approach and to provide future researchers with images and sequences for their experiments. We have made publicly available the generated wheelchair user model, the recorded sequences and ground truth files. We have proven the capacity of transfer learning from a training dataset to a new one completely independent.

There are multiple future work lines to improve the different proposals. About the wheelchair users detector, more complex models can be studied, for example considering more model variations. About the combination, we have chosen a simple technique, therefore it could be improved in order to optimize the combination of the different information sources. Also a new model can be trained using both the Smile Lab dataset and the recorded sequences to achieve greater generality. A tracker can be added to the sequence detection to combine the information extracted during the sequence frames giving temporal continuity to the detections. As the recorded dataset uses a multi-camera deployment, the detections obtained for each viewpoint can help to reinforce the detections from the other one. Apart from this, the typical lines of future work for

object detection can be applied here. Finally, the improved people detection can be used as a starting point for multiple event detection systems, in scenarios where the presence of wheelchair users is very common, such as hospitals, healthcare centers or senior residences.

References

1. D. Hosotani, I. Yoda, and K. Sakaue, "Wheelchair recognition by using stereo vision and histogram of oriented gradients (hog) in real environments," in *Workshop on Applications of Computer Vision*, 2009, pp. 1–6.
2. E. Auvinet, F. Multon, A. Saint-Arnaud, J. Rousseau, and J. Meunier, "Fall detection with multiple cameras: An occlusion-resistant method based on 3-d silhouette vertical distribution," *Information Technology in Biomedicine*, vol. 15, no. 2, pp. 290–300, 2011.
3. Z.-P. Bian, J. Hou, L.-P. Chau, and N. Magnenat-Thalmann, "Fall detection based on body part tracking using a depth camera," *Journal of Biomedical and Health Informatics*, vol. 19, no. 2, pp. 430–439, 2015.
4. W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 34(3), pp. 334–352, 2004.
5. M. Valera and S. A. Velastin, "Intelligent distributed surveillance systems: a review," *Visual Image Signal Processing*, vol. 152(2), pp. 192–204, 2005.
6. P. Kilambi, E. Ribnick, A. J. Joshi, O. Masoud, and N. Papanikolopoulos, "Estimating pedestrian counts in groups," *Computer Vision and Image Understanding*, vol. 110(1), pp. 43–59, 2008.
7. M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 31(12), pp. 2179–2195, 2009.
8. P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32(9), pp. 1627–1645, Sept 2010.
9. P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34(4), pp. 743–761, 2012.
10. D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32(7), pp. 1239–1258, 2010.
11. D. Simonnet, S. Velastin, E. Turkbeyler, and J. Orwell, "Backgroundless detection of pedestrians in cluttered conditions based on monocular images: a review," *IET Computer Vision*, vol. 6(6), pp. 540–550, 2012.
12. I. P. Alonso, D. F. Llorca, M. A. Sotelo, L. M. Bergasa, P. R. de Toro, J. Nuevo, M. Ocaña, and M. A. G. Garrido, "Combination of feature extraction methods for svm pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8(2), pp. 292–307, 2007.
13. A. Garcia-Martin and J. M. Martinez, "Robust real time moving people detection in surveillance scenarios," in *International Conference on Advanced Video and Signal based Surveillance*, 2010, pp. 241–247.
14. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *In proc. of Computer Vision and Pattern Recognition*, pp. 886–893, 2005.
15. P. Dollár, R. Appel, and W. Kienzle, "Crosstalk cascades for frame-rate pedestrian detection," in *European Conference on Computer Vision*, 2012, pp. 645–659.
16. B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *In proc. of International Journal of Computer Vision*, vol. 77(1-3), pp. 259–289, May 2008.
17. P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57(2), pp. 137–154, 2004.
18. M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Computer Vision and Pattern Recognition*, 2009, pp. 1014–1021.

19. W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, Y. Xiong, C. Qian, Z. Zhu, R. Wang, C. C. Loy, X. Wang, and X. Tang, "Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection," *In proc. of Computer Vision and Pattern Recognition*, 2014. [Online]. Available: http://arxiv.org/abs/1409.3505
20. R. B. Girshick, F. N. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," *In proc. of Computer Vision and Pattern Recognition*, vol. abs/1409.5403, 2014. [Online]. Available: http://arxiv.org/abs/1409.5403
21. S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, pp. 91–99, June 2015.
22. A. Myles, N. da Vitoria Lobo, and M. Shah, "Wheelchair detection in a calibrated environment," in *Asian Conference on Computer Vision*, 2002, pp. 1–7.
23. C.-A. Yang and P.-C. Chung, "Recovery of 3-d location and orientation of a wheelchair in a calibrated environment by using single perspective geometry," in *Region 10 Conference*, 2007, pp. 1–4.
24. C.-W. Wu, C.-D. Liu, and P.-C. Chung, "Assistance instruments detection using geometry constrained knowledge for health care centers," in *International Conference on Future Information Technology*, 2010, pp. 1–5.
25. Y.-X. Huang, S.-P. Hsu, C.-C. Yu, Y.-N. Chung, and C.-T. Lin, "Applying image technology to detect and track the wheelchair patient safety," in *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, 2013, pp. 2333–2415.
26. F. de Chaumont, B. Marhic, L. Delahoche, and C. Cauchois, "Generic method for recognition of a wheelchair, even with a low resolution-effective sensor," in *International Conference on Industrial Technology*, 2004, pp. 56–60.
27. C.-R. Huang, P.-C. Chung, K.-W. Lin, and S.-C. Tseng, "Wheelchair detection using cascaded decision tree," *Information Technology in Biomedicine*, vol. 14(2), pp. 292–300, 2010.
28. P.-J. Huang and D. yu Chen, "Robust wheelchair pedestrian detection using sparse representation," in *Visual Communications and Image Processing*, 2012, pp. 1–5.
29. R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2013. [Online]. Available: http://arxiv.org/abs/1311.2524
30. R. B. Girshick, "Fast R-CNN," *International Conference on Computer Vision*, pp. 1440–1448, 2015. [Online]. Available: http://arxiv.org/abs/1504.08083
31. P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Discriminatively trained deformable part models, release 4," http://people.cs.uchicago.edu/ pff/latent-release4/, 2010.
32. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *In proc. of Computer Vision and Pattern Recognition*, Sep 2014.
33. A. Garcia-Martin and J. M. Martinez, "Post-processing approaches for improving people detection performance," *Computer Vision and Image Understanding*, vol. 133, pp. 76 – 89, 2015.
34. M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," *Proc. of Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.
35. C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," *In proc. of Computer Vision and Pattern Recognition*, pp. 794–801, June 2009.
36. B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," *In proc. of Computer Vision and Pattern Recognition*, pp. 878–885, June 2005.
37. J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," *International Conference on Machine Learning*, pp. 233–240, June 2006.

Rafael Martín Nieto

Rafael Martín Nieto received the M.S. degree in Electrical Engineering ("Ingeniero de Telecomunicación" degree) in 2012 (2007-2012) at Universidad Autónoma de Madrid (Spain) and the MPhil degree in Research and Innovation in Information and Communication Technologies (postgraduate Master) in 2013. In 2007 he joined Video Processing and Understanding Lab at Universidad Autónoma de Madrid. His research interests are focused in the analysis of video sequences for video surveillance (object tracking, object detection, multicamera systems,…).

Álvaro García Martín

Alvaro García Martín received the M.S. degree in Electrical Engineering ("Ingeniero de Telecomunicación" degree) in 2007 (2002-2007) and the MPhil degree in Electrical Engineering and Computer Science (postgraduate Master) in 2009 and PhD in Computer Science in 2013 at Universidad Autónoma de Madrid (Spain). From 2006 to 2014, he has been with the Video Processing and Understanding Lab (VPU-Lab) at Universidad Autónoma of Madrid as a researcher and teaching assistant. In 2008 he received a FPI research fellowship from Universidad Autónoma de Madrid. He is currently an assistant professor at Universidad Autónoma of Madrid. His research interests are focused in the analysis of video sequences for the video surveillance (moving object extraction, object tracking and recognition, event detection…).

José M. Martínez

José M. Martínez was born in Madrid, Spain, in 1967. He received the Ingeniero de Telecomunicación degree (six years engineering program) in 1991 and the Doctor Ingeniero de Telecomunicación degree (PhD in Communications) in 1998 (Retevisión Award of the Official Professional College for Telecommunication Engineers to the best Ph.D. thesis of the academic year 1997-98), both from the E.T.S. Ingenieros de Telecomunicación of the Universidad Politécnica de Madrid. Since 2002 he is Associate Professor at the Escuela Politécnica Superior of the Universidad Autónoma de Madrid. His professional interests cover different aspects of multimedia information systems, focusing on content analysis, understanding and description, content adaptation and personalization for Universal Multimedia Access, and video summarization.