Proc. 2011 International Conference on Intelligent Robotics and Applications (ICIRA 2011), Aachen, Germany, LNAI 7102, pp. 328-337, Springer 2011

Detection of Lounging People with a Mobile Robot Companion

Michael Volkhardt, Steffen Müller, Christof Schröter, and Horst-Michael Groß*

Neuroinformatics and Cognitive Robotics Lab,
Ilmenau University of Technology, 98684 Ilmenau, Germany
michael.volkhardt@tu-ilmenau.de
http://www.tu-ilmenau.de/neurob

Abstract. This paper deals with the task of searching for people in home environments with a mobile robot. The robust estimation of the user's position is an important prerequisite for human robot interaction. While detecting people in an upright pose is mainly solved, most of the user's various poses in living environments are hard to detect. We present a visual approach for the detection of people resting at previously known seating places in arbitrary poses, e.g. lying on a sofa. The method utilizes color and gradient models of the environment and a color model of the user's appearance. Evaluation is done on real-world experiments with the robot searching for the user at different places.

Keywords: people detection; various poses; home environment

1 Introduction

This work is part of the CompanionAble[1] project, which intends to develop a personal robot for assisting elderly people with mild cognitive impairments in their home. The goal of the project is to increase the independence of the user by means of a combination of a smart home and a mobile robot. Therefore, the system provides different services, like e.g. day-time management, and allows for video conferences with care-givers or friends. Furthermore, it recognizes emergency situations, like falls, and tries to prevent progression of the cognitive impairments by providing stimulation programs. To offer these service functionalities, the robot system provides several autonomous behaviors. First, observing the user in a non-intrusive way allows to facilitate services that require interaction or to react on critical situations. Second, the robot must seek for the user if a reminder has to be delivered or a video call comes in. A third behavior is following and approaching the user if interaction is desired. A prerequisite to these behaviors is the robust detection and tracking of the user in the apartment. In contrast to other interaction applications in public environments, people in

* This work has received funding from the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216487.

[1] www.companionable.net

home environments often do not face the robot in an up-right pose but sit on chairs or lie on sofas. Therefore, our system tries to detect the user independent of their pose at places, where he or she usually rests. In this work, we focus on a robot-only solution – not relying on any smart home sensors – to enable the robot to function autonomously in any home environment. The key idea is to learn the visual appearance of predefined resting places and the user beforehand and to compare the current visual impression to both of these models in the detection phase.

The remainder of this paper is organized as follows: Section 2 summarizes previous work carried out on the research topic. We present the innovation of detecting lounging people at places in detail in Sec. 3. Afterwards, Sect. 4 gives a description of the experiments carried out, while Sec. 5 summarizes our contribution and gives an outlook on future work.

2 Related Work

People detection and tracking are prominent and well-covered research areas, and impressive results have been accomplished in recent years. Considering the constrained hardware of mobile robots, two main fields for people detection have been established – range-finder-based and visual approaches. [1] employ AdaBoost on laser range scans to combine multiple weak classifiers to a final strong classifier that distinguishes human legs from the environment. Visual approaches mainly focus on the face or the human body shape. The most prominent up-to-date face detection method also utilizes AdaBoost, which learns and applies a cascade of simple, but very efficient image region classifiers to detect faces [2].

Histograms of Oriented Gradients (HOG) have been established as the state-of-the-art method for upright people detection. The basic idea is to compute block-wise histograms of gradient orientations, resulting in robustness to slight spatial variation of object shape, color, and image contrast. The histograms inside a detection window are concatenated into a high-dimensional feature vector and classified by a linear Support Vector Machine [3]. Further extensions to the original HOG method focus on upper body detection [4] or use deformable sub-parts, which increase detection performance given partial occlusion [5]. Detection, segmentation and pose estimation of people in images is addressed by [6] who combine HOG features with the voting scheme of the Implicit Shape Model [7]. [8] augment the HOG features with color and texture information achieving impressive results on outdoor datasets. Unfortunately, the latter two approaches are far beyond real-time capabilities.

Plenty of research has been done to develop methods for people tracking on mobile robots in real-world applications. Most of these approaches focus on pedestrian tracking and single poses [3, 7, 9]. Yet, few approaches handle the detection and tracking of people in home environments, especially on mobile robots. Often smart home technologies, like static cameras with background subtraction methods [10] or infrared presence sensors, are applied, which facilitate the problem of detection [11, 12]. On occasion, approaches working with

mobile robots process the data captured offline to apply computationally heavy detection methods [9]. The CompanionAble project aims to develop a mobile robot that is able to react on and interact with the user. Therefore, all those approaches employing background subtraction or a retro-perspective analysis are not applicable.

3 Detection of Lounging People at Places

Typical scenarios in a home environment include the user walking to another room, or the user sitting on a chair or lying on a sofa. The latter case occurs quite frequently, i.e. when the user is watching TV, reading newspaper, making phone calls, working or sleeping. Our previous system comprises a multi-modal, multi-cue tracking framework based on the Kalman Filter update regime similar to [13]. In this former work, we apply boosted laser-based leg detection [1], a face detector [2], motion detection [14], and two HOG people detectors [3, 4]. The system is able to detect and track people in upright pose (mainly through legs and HOG) and frontal-view (mainly through face and upper-body HOG) in the surroundings of the robot. In the following, we go beyond the state-of-the-art by presenting a trained method for detecting people in difficult poses, that solely runs on a mobile robot in real-time. The system at first learns the appearance of places in the apartment where the user usually rests. Afterwards, the deviation of occupied places from the model and the similarity to a user model are used for detection.

3.1 Definition of Places

We define places as positions in the apartment where the user is usually encountered, i.e. chairs, sofas, working desk. Each place P is represented by a 3D box $\mathbf{b} = (x, y, z, d_x, d_y, d_z)$ with x, y, z being the center coordinates of the box and d_x, d_y, d_z denoting the width in each dimension. Figure 1 shows an exemplary place position in the world centered occupancy map used for navigation and the 2D-projection of the place box into the current camera image of the robot. Naturally, the content of the place-boxes looks completely different in the camera image, if observed from different positions. Since the system is learning the appearance of different places in the apartment, we need to restrict the pose from which the robot is observing them. Therefore, each place is assigned n observation poses $\mathbf{O} = (\mathbf{o}_1, \ldots, \mathbf{o}_n)$, where $\mathbf{o} = (x, y, \phi)$ with x, y representing the world coordinates of the robot's position and ϕ denoting the heading of the robot. The restriction of the observation position ensures that the variance of the place appearance is limited. Additionally, some kind of feature description model $\mathcal{M}$ of the place is added, where the nature of the description is variable. In this work, we use a contextual color histogram (Sec. 3.2) and a HOG description (Sec. 3.3). Thus, the full description of a place is given by $P = (\mathbf{b}, \mathbf{O}, \mathcal{M})$.

(a) Place in Occupancy Map (b) Projection to Camera Image

Fig. 1. Place definition. (a) Bounding box of place in the occupancy map. The robot is in its observation position (red circle). (b) Place's bounding box projected into the camera image.

3.2 Color-based User Detection

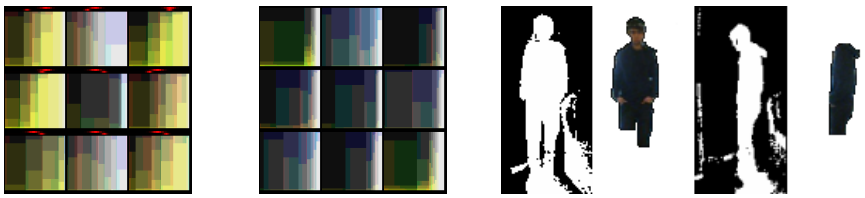
The color-based feature model comprises the appearance of each place in multi-modal histograms. Each place is observed from different, but predefined view-points given different illumination conditions, i.e. ambient day-light and electric lighting in the evening. Therefore, the color model must be learned for each day-time and observation angle, independently. Instead of storing the histograms for each context (day-time, view-point) in a vector, a more efficient way is using a set containing only the observed appearances and the corresponding context. The size of this set can be limited by merging similar entries and keeping only distinctive ones. Therefore, we use a multi-modal color model augmented by a discrete context distribution capturing the circumstances of the histogram's acquisition.

Multi-modal Contextual Color Model The model is defined by $\mathcal{M} = \{\kappa_1, \ldots, \kappa_n\}$, where $\kappa_i = (H_i, C_i)$ represents a component in the model with H_i denoting a color histogram and C_i being a multi-dimensional discrete context distribution. The histogram is 3 dimensional in RGB color space with 8 bins in each dimension (tests with other color spaces like HSV and Lab showed no significant difference in performance). The context distribution captures arbitrary aspects of the origin of the histogram, like viewpoint and day-time, in separate dimensions. We set the maximum number n of components in $\mathcal{M}$ to nine. At the start of training, the model comprises zero components. At first a histogram is extracted from the box of the non-occupied place in the camera image and added as a new component to the model. Once the number of components exceeds n, the model must be pruned by merging similar components. This is done by first calculating the pairwise similarity s of all components:

$$s = \mathcal{BC}\left([H_i, C_i], [H_j, C_j]\right) \ , \tag{1}$$

where $[H, C]$ is the concatenation of a histogram distribution and a context distribution and $\mathcal{BC}(p, q)$ denotes the Bhattacharyya coefficient of two distributions:

$$\mathcal{BC}(p, q) = \sum_{x \in X} \sqrt{p(x)q(x)} \ . \tag{2}$$



(a) Place Color Model (b) User Color Model (c) User Segmentation

Fig. 2. Color models. (a) Place model with 9 components. (b) User color model with 9 components. (c) Background subtraction output (binary image) and GrabCut refinement (color image) to learn color model of user. Left – shadow is removed very well. Right – parts of the person are removed.

The components with the highest similarity are merged by averaging the histogram and context distributions, respectively. The model $\mathcal{M}$ is learned for each place P in multiple teach runs including different day-times and illumination conditions. In the process of learning, the model maintains unique and distinctive representations of a place but merges similar descriptions. Figure 2(a) shows an exemplary color histogram of a place on the couch. Each bin in the 3 dimensional histogram is plotted as a 2D area with its corresponding mean color with the area size corresponding to the bin height. The state of the two dimensional context distribution, capturing point of view and day-time, is displayed in two small lines above each histogram. Red color indicates the probability of a state in the corresponding dimension. The histograms capture different lighting conditions, i.e. the couch normally appears in yellow-green (third column), bright given sunlight (second histogram in top row) or very dark at evening (histogram in the middle). Note that the model contains similar color histograms, but with different context distributions (first column).

Learning of the User Model The color model of the user is similar to the aforementioned color model of places, but without the context distribution. Model learning is done by first creating a Gaussian Mixture background model [10], when the robot is standing still and no hypothesis is in front of the robot's camera (given by the tracker output). This background model is used for background subtraction once a hypothesis is visible in the image. To remove shadows and to refine the segmentation we apply the GrabCut algorithm [15]. The algorithm is automatically initialized with foreground pixels of the segmentation and background pixels in the bounding box of the person. Figure 2(b) shows an exemplary learned color model of the user capturing mostly blue clothing which appear green under artificial light. One problem is the consistent segmentation of the user in the image. Although the GrabCut algorithm produces satisfying segmentation (Fig. 2(c) left), from time to time background pixels are misclassified in the segmentation, or parts belonging to the person are left out (Fig. 2(c) right). Therefore, at the moment and as a kind of interim solution, we trigger the learning of the user model once per day when the robot is standing in front of a white wall. The user is then asked to walk in front of the robot's camera.

Recognition of the User Once the place models and the user model have been trained, the system is able to detect the user in arbitrary poses at the learned places. For that purpose, the robot drives to the predefined observation positions and checks each place. By comparing the current appearance to the place and user model the system decides if the place is occupied by the user. Therefore, the robot first extracts the current color histogram H_c from the place's box in the camera image. Furthermore, a context distribution C_c is created including current day-time and view-angle. The system now calculates the similarity of the current observation histogram H_c to the color histogram H_l from the place model using the Bhattacharyya coefficient:

$$s = \mathcal{BC}(H_c, H_l) \ , \tag{3}$$

where H_l is the histogram of the best matching component κ_l in the place model with l selected by:

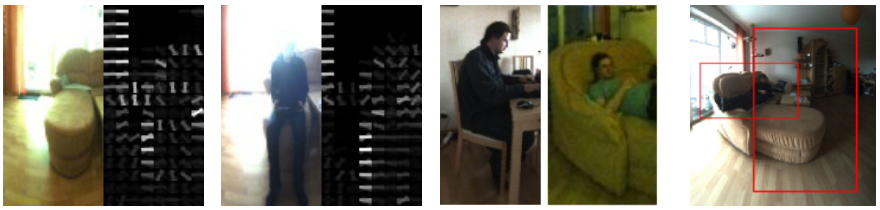
$$l = \underset{i=1,\ldots,n}{\arg\max} \{ \mathcal{BC} \left([H_c, C_c], [H_i, C_i] \right) \} \ . \tag{4}$$

Consequently, Eq. (3) is also used to calculate the similarity to the user model. Yet, a direct comparison of the complete histogram H_c to the user model's histograms would result in a very low match value, because the user usually only occupies a small region in the place's box and many background pixels are included in H_c. Therefore, the similarity to the user model is calculated by using a correlation window inside the place's bounding box and shifting it to find the highest similarity. To select the best matching component κ_l from the user model, Eq. 4 is applied again, but the context distribution is omitted and only the histograms are used.

If the user is present, this results in low similarity to the place model, because the appearance of the place is partially covered, and a high similarity to the user model, because the correlation window fits to the position of the user. If the user is not present, the results are vice versa. Proper decision criteria must be defined for both similarities to decide if a place is occupied. To this end, we trained a *single* linear Support Vector Machine (SVM) on data of multiple labeled runs with empty and occupied places [16]. The resulting SVM then decides for each place if the user is present given the similarities to both the place and user model. If the training data is diversified enough, the SVM is generally applicable to other scenarios without the need of retraining.

3.3 HOG-based User Detection

Besides the color-based approach described in Sec. 3.2, we developed an illumination invariant gradient-based feature description model $\mathcal{M} = (\mathbf{H})$, where $\mathbf{H} = (\mathbf{d}_1^T, \ldots, \mathbf{d}_n^T)$ with $\mathbf{d}_i$ being a HOG feature descriptor. Recall from Sec. 3.1 that the definition of a place is $P = (\mathbf{b}, \mathbf{O}, \mathcal{M})$. Each bounding box $\mathbf{b}$ of the place in the image viewed from different observation position of $\mathbf{O} = (\mathbf{o}_1, \ldots, \mathbf{o}_n)$ is scaled to a fixed size of 64×128 pixels allowing to extract a $3,780$-dimensional



(a) Empty Place (b) Occupied Place (c) Place Examples (d) Overlapping Places

Fig. 3. HOG descriptor and place examples. (a) HOG Descriptor of an empty place (b) HOG Descriptor of an occupied place. A linear SVM is trained on multiple views of both cases. (c) Examples of places given different illumination conditions. (c) Overlapping places induce confusions.

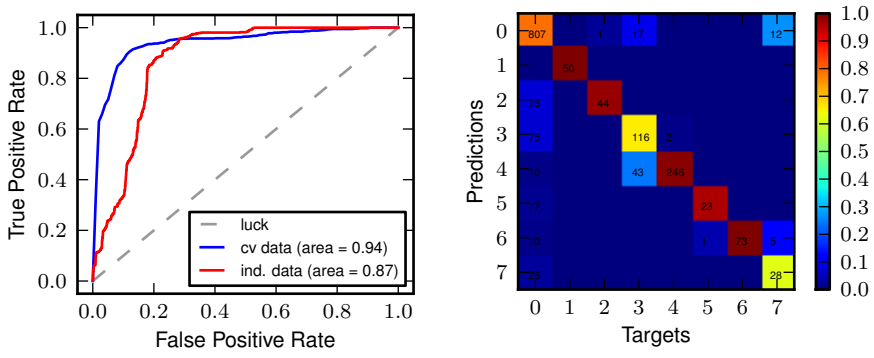
feature vector $\mathbf{d}_i$ using the standard HOG descriptor used for people detection [3]. We trained a separate linear SVM for *each* place on the same data as in Sec. 3.2. Therefore, the system learns to distinguish the HOG descriptors of an empty place, viewed from different positions, from the HOG descriptors of occupied places (Fig. 3). An explicit user model is not required in this case.

4 Experiments

In this section, the color-based and the HOG-based user detection are evaluated separately, because the training and estimation works differently for both approaches. The first method requires different illumination conditions to train robust empty place models and relies on a user model, while the latter needs to learn the gradient features of empty and occupied places without requiring an explicit user model.

4.1 Color-based User Detection

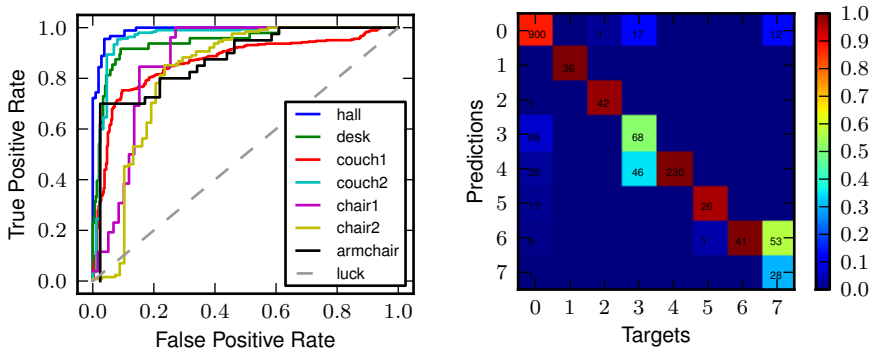
We first learned the appearance of seven predefined (empty) places of the apartment in multiple training runs including three different lighting conditions – ambient day light, bright sunlight, and artificial light at evening. Furthermore, a user model was trained with the user wearing two different clothes, which he also weared in the test runs. In the test scenario the robot was placed on a starting position and was searching for the user, who was either lounging at one of the places or not in the apartment. Figure 3(c) visualizes some exemplary places occupied by the user. The robot checked each place for the user's presence and logged the similarities to the place and user models. A linear SVM was trained and tested via 5-fold cross validation on the collected data of similarities of all places. The ground truth of the user's presence was labeled manually. For evaluation, we calculated the probability of the test examples belonging to the two classes of the SVM model (user present and not). By varying the probability threshold that is required to assign an example to one class, an ROC curve can



(a) ROC color-based (b) Confusion Matrix color-based

Fig. 4. Evaluation of the color-based detection approach (Sec. 3.2). (a) ROC curves for cross-validation and independent data (b) Confusion Matrix of multiple independent test runs. Classes $1-7$ represent different places, class 0 denotes user was not present.

be plotted (Fig. 4(a)). The blue curve shows the ROC of the cross-validated data used for training and validation. The red curve was generated on data from multiple independent test runs not seen by the system before. The high true positive rate and a low false positive rate of the red curve indicate that the system is actually able to robustly detect the user on the places. Furthermore, we evaluated the detection performance of the color-based detection system on the independent data sets used to generate the red ROC curve. The trained SVM is used to decide if the user is present or the place is empty. Each place is considered as a class and classification rates are calculated. Since the robot checks different places for the user's presence, it can wrongly detect the user at a place different from the ground truth. Additionally, sometimes more than one place is visible in the robot's camera image (Fig 3(d)). Hence the detection of places can be confused. Therefore, for evaluation a confusion matrix is chosen (Fig. 4(b)). Class 0 is used to denote that the user is not present (accumulated over all places). The classes $1-7$ correspond to places in the hall, a desk, 2 couch positions, 2 chair positions and an arm chair, respectively. Detection rates for each class are given in the main diagonal of the matrix. The average classification rate is above 85% with the biggest outliers in classes 3 (couch) and 7 (arm-chair). When the user is resting on the couch, he occasionally occupies two places (class 3 and 4) due to overlapping boxes (Fig. 3(d)). If the user is sitting on the edge of the couch like in Fig. 3(b), this results in high similarities to the user model in both couch places leading to the realtive high confusions in class 3. In the case of class 7, direct sunlight caused the camera to overexpose and proper color extraction was hardly possible. Additionally, some false positive detections occured (non-0-predictions of class 0).



(a) ROC HOG-based (b) Confusion Matrix HOG-based

Fig. 5. Evaluation of HOG-based detection (Sec. 3.3). (a) ROCs generated by SVMs for each individual place on the ind. data set. (b) Confusion Matrix generated by SVMs on ind. test runs. Classes $1-7$ represent places from (a), class 0 denotes user not present.

4.2 HOG-based User Detection

The HOG-based detection method was evaluated by first training a separate 5-fold cross validated linear SVM for each place. The SVMs learned to distinguish the HOG features of an empty and occupied place. In contrast to the color-based SVM, the resulting classifiers are only effective on data similar to the training set and must be retrained in new environments. The datasets were the same as for the evaluation of the color-based approach. For clarity, we omit the ROC curves of the cross-validated data and only present the ROCs for each place on the independent data set (Fig. 5(a)). The curves substantiate that the HOG-based approach is also applicable to detect the user resting on places. The worst ROCs are given by the chair and arm-chair examples (classes $5, 6, 7$). This happens because the chairs got moved and turned a little during the experiments, disturbing the HOG descriptor. Figure 5(b) presents the confusion matrix that was generated on the aforementioned test runs by using the trained HOG-SVMs to classify the user's presence at each corresponding place. Similar to the color-based approach, the system confuses some examples of the couch place (class 3). Furthermore, the overexposure of the camera in the case of class 7 results in misclassification and the aforementioned moving of the chairs causes confusions with class 6. Compared to the results of the color-based approach in Fig. 4(b), the HOG-based approach produces slightly fewer false positives.

5 Conclusion and Future Works

We presented a method to detect lounging people independent of their resting pose at predefined places. The system either learns color histograms of the user

and places in the apartment or builds up a gradient model for each place, beforehand. Then SVMs are trained to decide if a place is occupied by the user. Afterwards, the system is able to detect the user in the given environment. Experiments on multiple independent test runs substantiate that the approach actually improves recognition performance in living environments by detecting the user in situations not captured by common detection and tracking systems on mobile robots. In future work, we intend to combine the advantages of the color-based and gradient-based approach. With the color-based approach being invariant to moved furniture to some degree and the HOG-based approach's advantage of illumination invariance, a proper combination of both methods shall be developed. Furthermore, the segmentation of the user should be improved to facilitate the recording of the user model for the color-based approach. Last but not least, the manual definition of places should be replaced by an interactive training guided by the user.

References

1. Arras, K. O., Mozos Ó. M., Burgard, W.: Using Boosted Features for the Detection of People in 2D Range Data. In: Proc. IEEE ICRA. pp. 3402–3407, (2007)
2. Viola, P., Jones, M.: Robust Real-time Object Detection. In: International Journal of Computer Vision, (2001)
3. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: Conference on CVPR, IEEE Computer Society, pp. 886–893, (2005)
4. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: Conference on CVPR, pp. 1–8, (2008)
5. Felzenszwalb, P. F., et. al: Object Detection with Discriminatively Trained Part Based Models. In: PAMI, pp. 1–20, (2009)
6. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting People Using Mutually Consistent Poselet Activations. In: ECMR. pp. 168–181, (2010)
7. Leibe, B., et. al: Robust Object Detection with Interleaved Categorization and Segmentation. In: IJCV 1-3, vol. 77, pp. 259–289, (2008)
8. Schwartz, W. R., Kembhavi, A., Harwood, D., Davis, L. S.: Human detection using partial least squares analysis. In: 2009 IEEE 12th ICCV, pp. 24-31, (2009)
9. Ess, A., Leibe, B., Schindler K., Gool, L. van: A Mobile Vision System for Robust Multi-Person Tracking. In: IEEE Conf. on CVPR, pp. 30–37, (2008)
10. Stauffer, C., Grimson, W.E.L.: Adaptive Background Mixture Models for Real-Time Tracking. In: CVPR, pp. 2246–2253, (1999)
11. Han, T. X., Keller, J. M.: Activity Analysis, Summarization, and Visualization for Indoor Human Activity Monitoring. In: IEEE TCSVT, pp. 1489–1498, (2008)
12. Rusu, R. B., et. al: Human Action Recognition Using Global Point Feature Histograms and Action Shapes. In: Advanced Robotics, 23(14), pp. 1873–1908, (2009)
13. Müller, St., Schaffernicht, E., Scheidig, A., Böhme, H.-J., Gross, H.-M.: Are you still following me?. In: Proc. ECMR, pp. 211–216, (2007)
14. Martin, Chr., et. al.: Sensor Fusion using a Probabilistic Aggregation Scheme for People Detection and People Tracking. In: RAS, vol. 54, 9, pp. 721–728, (2006)
15. Rother, C., Vladimir, K., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In: ACM SIGGRAPH 2004 Papers, pp. 309–314, (2004)
16. Chih-Chung Chang and Chih-Jen Lin: LIBSVM: a library for support vector machines. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. (2001)