in: 2nd CompanionAble Workshop co-located with the 3rd European Conference on Ambient Intelligence November 18-21 2009, Salzburg, Austria

A Concept for Detection and Tracking of People in Smart Home Environments with a Mobile Robot

Michael Volkhardt, Christoph Weinrich, Christof Schröter, and Horst-Michael Groß

Neuroinformatics and Cognitive Robotics Lab,
Ilmenau University of Technology, Germany
{michael.volkhardt,christoph.weinrich,
christof.schroeter,horst-michael.gross}@tu-ilmenau.de
http://www.tu-ilmenau.de/neurob

Abstract. Ambient Assisted Living (AAL) describes concepts, products and services for integrating new technologies and social environment in order to support people in their daily routine and increase their quality of life, in particular in their homes. One aspect is the use of "intelligent" sensors and evaluation for situation awareness and automatic detection of critical conditions like falls or any distress and crisis situations. A preliminary requirement to enable these functionalities is the robust estimation of people's positions and poses. In this paper we present a survey of state-of-the-art methods for people detection and tracking, focussing on application on a mobile robot. Furthermore, we discuss problems arising when using these algorithms in the target environment and show a concept to increase the robustness of multi-cue people tracking.

Key words: people detection, people tracking, smart home environment, mobile robot, survey

1 Scenario Description

CompanionAble[1] is a research project aiming to develop assistive services for elderly people, in particular those with mild cognitive impairments (MCI), in their home environment. Key features are the support of cognitive stimulation and therapy management, day-time management and video-conferencing with relatives or care-givers. These services are presented to the user by means of a combination of smart home technology and a robot companion. In contrast to a pure smart home environment, the mobility of the robot increases the ability of the system to handle the naturally dynamic aspects of a person (being able to offer service where it is needed) and provides a much more natural human-machine interaction experience. Example monitoring functions of the mobile

[1] The CompanionAble project is funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216487. www.companionable.net

in: 2nd CompanionAble Workshop co-located with the 3rd European Conference on Ambient Intelligence November 18-21 2009, Salzburg, Austria

robot include pose analysis (in particular fall detection), activity recognition, and long-term behavior pattern analysis. Assistance aspects include reminders to take medication, video conferences, and detection of any distress and crisis situations.

A preliminary requirement to these functions is the robust detection and tracking of the care-recipient. Therefore the robot must be able to detect and track a person in its surrounding area at almost any time. Although the embedding in a smart home environment allows to enhance the perception by integrating additional external information cues, like infrared presence sensors or wall-mounted cameras, in general we prefer approaches that enable the robot to function autonomously in any home environment.

The focus of this paper lies on visual techniques for detecting people in a home environment. Visual people detection has received a lot of attention and shown impressive progress recently, as it yields the widest range of information and good detection performance, with the downside that it also requires more effort than for instance range-measurement-based detection. The remainder of this paper is organized as follows: the next section describes the particular challenges for people detection algorithms introduced by the home environment scenario. After that we give a survey of existing visual approaches for people detection and present their pros and cons in the focused setting. Section 4 develops a concept for robust visual people detection in a home environment. Finally we give a conlusion and discuss some future trends in Sec. 5.

2 Challenges and Requirements of People Detection

In contrast to controlled scenarios, a typical inhabited home is highly dynamic and unconstrained. This includes small rooms packed with furniture and other items[2] with dynamic configuration and a freely moving person. Therefore robust visual algorithms need to handle frequent occlusion by objects in the room and very high variances in the appearance of the care-recipient. The variance is primarily induced by dynamic positions, multiple articulations and poses of the person. The appearance of the person varies enormously when standing directly in front of the robot where only the upper body or face is visible and when standing several meters away where the person in the image is only a few pixels high. Furthermore, people should not only be detected in an upright position (as it is common in outdoor pedestrian detection), but also when sitting on a chair or even lying on the sofa. In order to reliably recognize a fall or state of human collapse from an upright position, the ability to detect people in different positions in the image and with uncommon poses is required. Additionally, various articulations and activities introduce self-occlusion or partial occlusion from objects the person is manipulating. Changing illumination conditions (day, night, artificial light sources, TV) further increase the variance of the appearance. While handling broad intra-class variability, algorithms also need to be

[2] This is the main reason why leg detection with laser range scanners leads to bad results, because tables and chairs are often mistaken for legs.

highly discriminative to distinguish persons from other objects and generalize adequately [1].

It is also worth noting that real-time operation of the employed methods is required: Since the robot needs to interact and therefore react on a person's behavior, a retroactive analysis (which is feasible in other applications like motion capturing or processing of recorded surveillance material) is not sufficient. Additionally cost factors of the senario constrain the use of expensive sensors like stereo-vision, thermal-imager or z-cameras. Therefore, we focus on visual approaches on monocular cameras.

3 Survey of Existing Visual Approaches

The field of visual people detection can be subdivided in implicit and explicit methods (Fig. 1). Implicit methods learn the background and detect foreground objects like persons or moving objects [2, 3] as a deviation from the background model. They require static cameras which makes their use very limited on a mobile robot. Furthermore, they require an additional classification of the detected objects, in order to distinguish people from other dynamic objects. In contrast to that, explicit methods detect people with any kind of person model. This model can consist of a representation of the complete body or only parts thereof, the best example being the face. The most prominent contemporary face detection

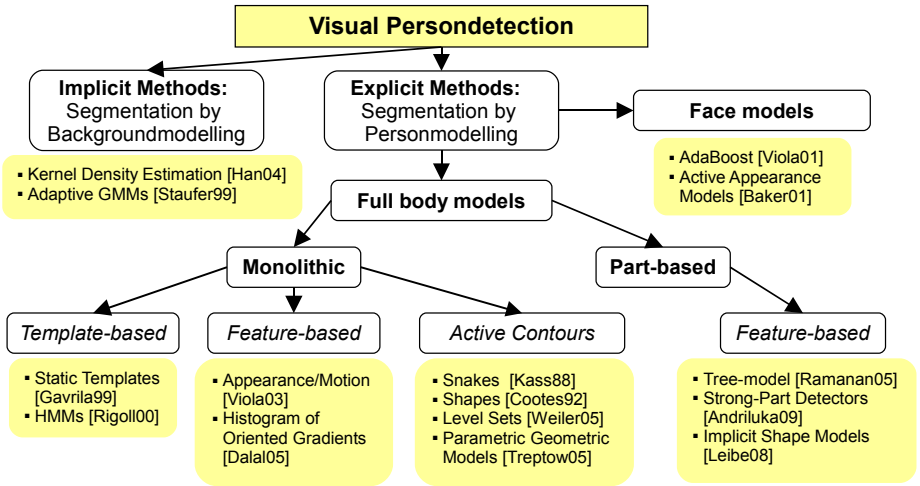
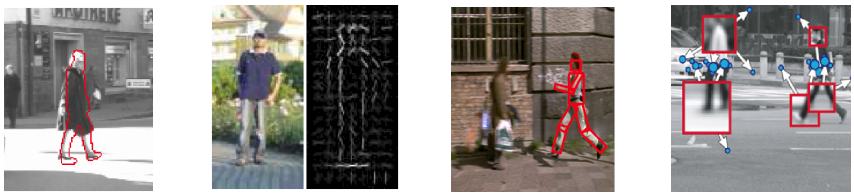


Fig. 1. Survey of existing visual approaches for people detection.

methods are AdaBoost [4] and Active Appearance Models (AAM) [5]. While the AdaBoost approach trains and applies a cascade of simple but very efficient image region classifiers to detect faces, Actice Appearance Models try to build

a parametric model of general face geometry and texture and iteratively fit this model to the current image, this process includes determining the specific parameters of the current face articulation. While AAM yields more information, it also requires significantly higher resolution of the face region than AdaBoost.

Full body models can be distinguished into monolithic methods, which try to detect the full body at once and part-based models, which represent the body by its individual parts and (often) their spatial relationships. Monolithic body representations (almost always discriminative models) use templates, active countours or features. Templates cannot efficiently capture high variances in people's articulations and poses [6, 7]. Therefore, huge numbers of different templates need to be processed over the image in a sliding window fashion to detect people (Fig. 2(a)). Active contours try to overcome the rigidness of templates by fitting flexible structures to the human's body [8–11]. They usually require high computational effort and are often distracted by background structures. Popular representatives of feature-based monolithic methods are approaches using features of appearance and motion [12] or approaches using Histograms of Oriented Gradients (HOG) based on [13].



(a) Template-tree (b) HOG (c) Part-tree-model (d) ISM

Fig. 2. Full-body models. Monolithic methods: (a) template-based method from [7], (b) feature-based method [13]. Part-based methods: (c) Tree-model from [15], (d) Implicit shape model from [18].

The basic idea of HOG is to compute block-wise histograms of gradient orientations, resulting in robustness to slight spatial variation of object shape (Fig 2(b)). Furthermore it is invariant to color properties, as caused e.g. by varying clothes. Additionally, a normalization of the resulting histograms supports invariance to the illumination conditions (image contrast). The gradients within images showing people are mostly caused by texture and contour edges. Both are highly variable due to different clothes and the flexibility of the human body. Therefore, commonly Support Vector Machines (SVM) are used to classify the high-dimensional feature space [14]. However the classification of people with different poses remains a particularly challenging task (see Sect. 4).

Existing part-based body models are mostly generative and feature-based. These models consist of different body parts (head, torso, limbs) and a representation of their spatial relationship [15]. Many methods use tree-based models, which represent the body by rectangles starting from the torso as root node and

linking the head and limbs to it (Fig 2(c)). The limbs are sometimes subdivided in further parts like upperarm, forearm and hands to detect complex poses. The stiffness of the links constrains possible articulations. Generation of the model is done by learning features for each part, which range from simple color histograms [16] to complex boosted shape context features [17]. The biggest disadvantage of these approaches is the high computational costs for the detection of each body part. Therefore many heuristics like initialization of tracking algorithms on predefined poses and assumptions about self-occlusion and illumination conditions. Hence, they are not suitable for unconstrained real-world applications, where people appear in arbitrary poses.

Other feature-based approaches implicitly learn the appearance of a person by features extracted at interest point locations and model their spatial occurrences. A popular approach, which can handle partial occlusion, is known as the Implicit Shape Model [18]. This approach was originally proposed for pedestrian detection (mainly upright poses and lateral viewpoint). In a training stage a codebook of the appearance of people is learned by extracting local features at scale-invariant interest points and clustering these features based on their similarity. Additionally the local occurrence of each feature relative to the object center is retained. Detection of people is done by searching interest points in the image and matching the corresponding features to the codebook. The occurrences of matching codebook entries are projected into the image (Fig. 2(d)). Then, a maximum search can find possible people locations. To improve the hypotheses a top-down segmentation step can be applied afterwards. Main disadvantages of the approach include processing time and false positives in structured backgrounds. In Sect. 4 we present enhancements to improve the performance and detection results of the approach and try to extend it to detection of multiple poses.

4 Concept for Multi-Cue People Tracking

As shown in the last section people detection algorithms used in this specific scenario either lack of robust detection or require high computational effort. To increase the robustness we combine different aforementioned methods in a probabilistic framework. Furthermore we adapt two methods, namely ISM and HOG, to recognize multiple poses.

Previous work [19] showed that probabilistic modeling techniques can enormously increase detection and tracking quality. Merging detections from the different visual sensor cues in a Bayesian filter framework has several advantages (Fig 3). First the insertion of new hypotheses can be delayed until there is enough evidence from multiple consistant observations from different sensor cues. As a consequence single weak (and often false positive) hypotheses from a single sensor cue can be discarded. Second the tracking component includes a smoothing of the detections and handles short occlusions by preserving hypotheses for a short period of time if there are no supporting observations. Last but not least the tracking aspect can lower the computational costs by mostly

applying the detection algorithms in areas, where there is already support from existing hypotheses [20]. Only once in while one has to process the full image to detect new persons.

We propose a system that uses a face detector for people standing close to the robot [4]. When people move away from the robot we use a combination of HOG and ISM people detection. At all time these detection methods are support by further detection and verification modules. These modules include: edge detection, motion detection, and a skin color tracker [21]. The Motion detection cue is only active when the robot is standing still. A noise and complexity reduction is achieved by working only on sums of image columns. The skin color tracker uses multiple instances of particle filters for clustering color regions in the image. Therefore the camera is dynamically calibrated by a white reference [22]. The skin color model is adapted online with pixels from face detections to manage illumination changes [23]. Note that none of these modules can reliably detect person on its own, but the combination in the Bayesian filtering framework allows for robust people detection.

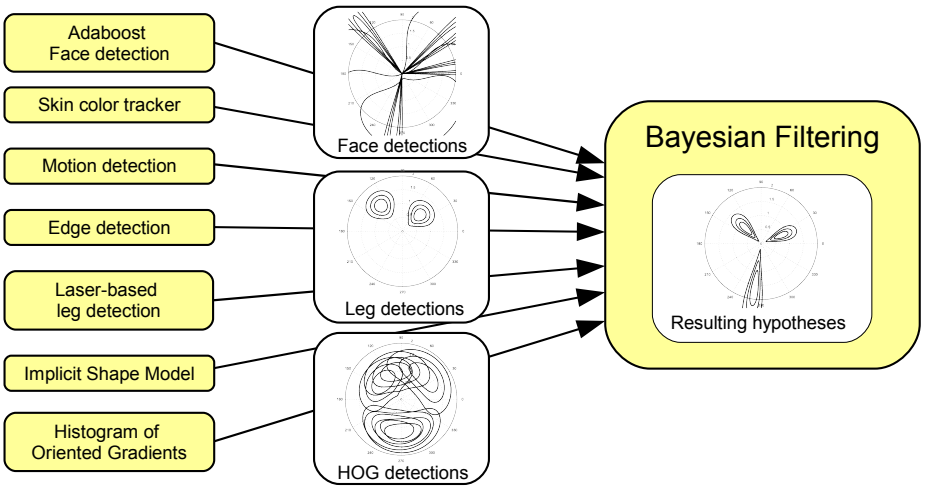


Fig. 3. Concept for probabilistic multi-cue tracking. The sensor cues create hypotheses of people (shown for face, leg and HOG detections). Each senor cue is incomplete and noisy. A Bayesian filtering approach aggregates sensor hypotheses from the current and previous timesteps to correct and sharpen the representation of the current situation.

As a step beyond the state of the art we plan to extend the Implicit Shape Model (ISM) to work in realtime for multiple poses like standing, sitting and lying. [24] proposed several new algorithms like determining very stable features and approximative NN search to increase robustness and performance of the original ISM. With these extensions and the help of laser range scanners to constrain the search space of feature detection, the approaches is real-time capable.

We are currently working on more lightweight representations for the modeling of features and especially their occurrences. We hope to further improve the processing speed and allow real time processing for multiple trained poses without the need for range information. Multiple poses can either be detect by one big Implicit Shape Model, multiple specialized models for each pose or a 4D-ISM coding the pose in the 4th dimension [25].

In the field of HOG some extensions are investigated. Through boosting it is expected to enable detection by considering only significant blocks allowing a more efficient classification [26]. To solve the complex problem of varying poses a cascade is used for problem decomposition. At each cascade level the feature space is separated by a hyper-plane specified by a corresponding linear SVM. Each SVM is parameterized in a way, that one preferably big sub-space contains only non-person features. Due to the complexity of the separation problem, the remaining sub-space contains person but still non-person features. Therefore detection windows that generated features within this sub-space are processed by the next cascade level. This level allows a finer classification, particularly due to the use of a feature space clamped by blocks, especially significant for the according classification problem. The last cascade level is reached once the problem simplifies to a linear separable problem and a final positive classification result is possible.

5 Conclusion and Outlook

This paper presented a concept for multi-cue people tracking on a mobile robot in a target environment of a smart home. These environments create new challenges for people detection algorithms like frequent occlusion, multiple poses, and dynamic illumination. In a survey we showed the advantages and disadvantages of current visual people detection methods in home environments. Because none of the presented methods is able to solve the problem of robust, real-time people detection on its own, we propose a concept to combine different sensor-cues in a Bayesian filtering framework. We hope to robustly detect people in large variability of appearances, caused by dynamic view-points, articulations and poses.

References

1. Schiele, B., Andriluka M., Majer, N. Roth, S., Wojek, C.: Visual People Detection: Different Models, Comparison and Discussion. In: ICRA, Workshop on People Detection and Tracking. pp 1–8. (2009)
2. Stauffer, C., Grimson, W.E.L.: Adaptive Background Mixture Models for Real-Time Tracking. In: Conference on CVPR, Vol. 2, pp. 2246. (1999)
3. Han, B., Comaniciu, D., Davis, L.: Sequential kernel density approximation through mode propagation: applications to background modelling. In: Asian Conference on Computer Vision, (2004)
4. Viola, P., Jones, M.: Robust Real-time Object Detection. In: International Journal of Computer Vision. (2001)

5. Baker, S., Matthews, I.: Equivalence and efficiency of image alignment algorithms. In: Conference on CVPR, pp. 1090–1097. (2001)
6. Rigoll, G., Eickeler, S., Muller, S.: Person tracking in real-world scenarios using statistical methods. In: International Conference on Automatic Face and Gesture Recognition, pp. 342–347. IEEE Press, (2000)
7. Gavrila, D.M.: Real-time object detection for "smart" vehicles. In: International Conference on Computer Vision, pp. 87–93. IEEE Press, (1999)
8. Kass, M., Witkin, A. P., Terzopoulos, D.: Snakes: Active contour models. In: International Journal of Computer Vision 4, Vol. 1, pp. 321–331. Springer, (1988)
9. Treptow, A., Cielniak, G., Duckett, T.: Comparing Measurement Models for Tracking People in Thermal Images on a Mobile Robot. In: ECMR, (2005)
10. Cootes, T. F., Taylor, C. J., Cooper, D. H., Graham, J.: Training models of shape from sets of examples. In: Brit. Machine Vis. Conf., pp. 9–18. Springer Verlag, (1992)
11. Weiler, D., Eggert, J.: Level-Set Segmentation with Contour based Object Representation. In: International Joint Conference on Neural Networks. (2009)
12. Viola, P., Jones, M., Snow, D.: Detecting Pedestrians Using Patterns of Motion and Appearance. In: Int. Conference on Computer Vision, Vol. 2, pp. 734–741. (2003)
13. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: Conference on CVPR, pp. 886–893. IEEE Computer Society (2005)
14. Wojek, C., Schiele, B.: A Performance Evaluation of Single and Multi-feature People Detection. In: DAGM-Symposium. pp 82–91. (2008)
15. Ramanan, D., Forsyth, D. A., Zisserman, A.: Strike a Pose: Tracking People by Finding Stylized Poses. In: Conference on CVPR, San Diego, pp. 271–278. (2005)
16. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: Conference on CVPR, pp. 1–8. (2008)
17. Andriluka, M., Roth, S., Schiele, B.: Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. In: Conference CVPR, pp. 1014–1021. (2009)
18. Leibe, B., Leonardis, A., Schiele, B.: Robust Object Detection with Interleaved Categorization and Segmentation. In: International Journal of Computer Vision 1-3, Vol. 77, pp. 259–289. (2008)
19. Gross, H.-M., Böhme, H.-J., Schröter, Ch., Müller, St., König, A., Einhorn, E., Martin, Ch., Merten, M., Bley, A. TOOMAS: Interactive Shopping Guide Robots in Everyday Use - Final Implementation and Experiences from Long-term Field Trials. In: IROS, pp. 424–429. IEEE Press (2009)
20. Küblbeck, C., Ernst, A.: Face detection and tracking in video sequences using the modified census transformation. In: Image Vision Comput. 24, pp. 564–572. (2006)
21. Martin, Chr., Schaffernicht, E., Scheidig, A., Gross, H.-M.: Sensor Fusion using a Probabilistic Aggregation Scheme for People Detection and People Tracking. Robotics and Autonomous Systems 9, Vol. 54, pp. 721–728. (2006)
22. Gross, H.-M., König, A., Schröter, Ch., Böhme, H.-J.: Omnivision-based Probalistic Self-Localization for a Mobile Shopping Assistant Continued. In: IROS, pp. 1505-1511, IEEE Omnipress. (2003)
23. Martin, Ch., Bhme, H.-J., Gross, H.-M.: Conception and Realization of a Multisensory, Interactive Mobile Office Guide. In: Proc. 2004 IEEE Int. Conf. on Systems, Man and Cybernetics, pp. 5368-5373, IEEE omnipress. (2004)
24. Spinello, L., Triebel, R., Siegwart, R.: Multiclass Multimodal Detection and Tracking in Urban Environments. In: Int. Conf. on Field and Service Robotics. (2009)
25. Seemann, E., Leibe, B., Schiele, B.: Multi-Aspect Detection of Articulated Objects. In: IEEE Conference on CVPR, pp. 1582–1588. IEEE Computer Society, (2006)
26. Zhu, Q., Avidan, S., Yeh, M.C., Cheng, K.T.: Fast Human Detection Using a Cascade of Histograms of Oriented Gradients, In: CVPR, pp. 1491–1498. (2006)