# Human Body Orientation Estimation in Multiview Scenarios

Lili Chen[1], Giorgio Panin[2], and Alois Knoll[1]

[1] Department of Informatics,
Technische Universität München,
85748 Garching bei München, Germany
[2] German Aerospace Center (DLR),
Institute of Robotics and Mechatronics,
82230 Wessling, Germany

**Abstract.** Estimation of human body orientation is an important cue to study and understand human behaviour, for different tasks such as video surveillance or human-robot interaction. In this paper, we propose an approach to simultaneously estimate the body orientation of multiple people in multi-view scenarios, which combines a 3D human body shape and appearance model with a 2D template matching approach. In particular, the 3D model is composed of a generic shape made up of elliptic cylinders, and a 3D colored point cloud (appearance model), obtained by back-projecting pixels from foreground images onto the geometric surfaces. In order to match the reconstructed appearance to target images in arbitrary poses, the appearance is re-projected onto each of the different views, by generating multiple templates that are pixel-wise, robustly matched to the respective foreground images. The effectiveness of the proposed approach is demonstrated through experiments in indoor sequences with manually-labeled ground truth, using a calibrated multi-camera setup.

## 1 Introduction

In the context of automatic video surveillance, people detection and tracking [1–5] are probably the most important tasks, which received a significant amount of attention in the area of research and development. However, an intelligent surveillance system should not only be able to locate and track people, but also understand and recognize their behaviors. To this aim, a representative cue that is often neglected is body orientation, which provides hints about what the person is going to do, and where the person is probably looking at. The aim of this paper is an approach to robustly estimate body orientation of multiple people, through multiple calibrated views mounted on the ceiling from different viewing angles.

To our knowledge, very few works have been conducted to visually estimate the orientation of human body. Some have addressed this issue by considering the person dynamics, assuming that the orientation is simply given by the walking direction [6, 7]. In [6], estimation is based on the motion of a tracked person and the

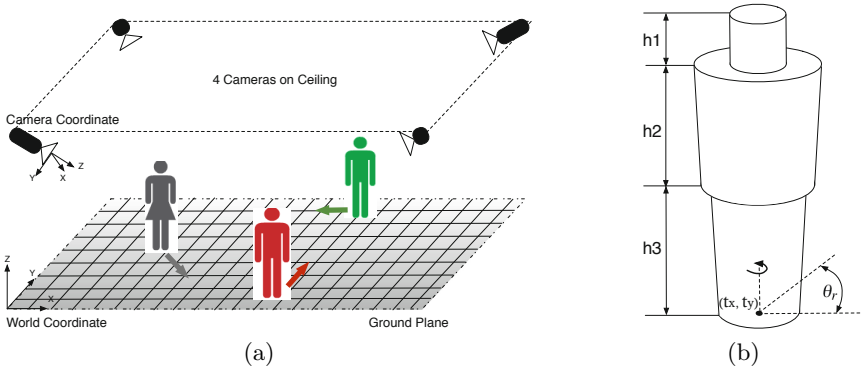
G. Bebis et al. (Eds.): ISVC 2012, Part II, LNCS 7432, pp. 499–508, 2012.

size of its bounding ellipse, but it fails in the case of people who are not moving, or slowly walking. [7] also couples human body orientation with motion direction, where the coupling is given by a dynamical model. This work assumes a loose coupling at low speed, whereby the absence of an explicit observation model for the orientation results in similar problems. Some other works [8, 9] consider this task as a classification problem. [8] use HoG descriptors to classify the orientation of pedestrians, recovering an estimate based on 2D low-resolution images. However, they group the orientation of a person very roughly, covering a 45-degrees range for in-plane rotations only. In [9], body pose classification is performed by using multi-level HoG features, and a sparse representation technique, at each frame of the sequence. However, for each human region they end up with a very high dimensional feature vector, which has the drawback of computational complexity, and their experimental scenarios are limited to a single view.

In order to address the shortcomings of motion-based or classification-based techniques, in this paper we propose a methodology combining a 3D human body/appearance model with 2D template-based matching. The primary goal is to robustly and accurately determine the body orientation of a variable number of people, randomly walking in an overlapping, multi-camera environment, which is able to deal with still-standing or slowly moving people, while covering a full 360 degrees range. More precisely, the 3D human appearance model is represented by a colored point cloud, obtained by back-projecting pixels from foreground images onto the surface of a simple geometric model, composed of three cylinders with elliptical section. In order to match the reconstructed model to new images under arbitrary poses, we reproject the 3D colored point cloud onto each view at different locations and orientations, thereby generating multiple 2D templates, which are robustly matched to the underlying foreground image through pixel-level measurements. For this purpose, a recently developed multi-camera, multi-people tracking system based on silhouette edges [10] yields ground plane locations, that are used as a position reference when estimating the body orientation. Subsequently, orientation is estimated by a full search, only restricted to a neighborhood of the previous orientation estimate.

By resuming, the effectiveness of our system is due to three main aspects. One is the reconstruction of a detailed 3D appearance, that makes a pixel-level matching more precise with respect to global or local statistics, such as color histograms. The second is the use of a robust multi-target detector and tracker, since a good location reference is necessary for an accurate orientation estimation. And the third is the integration of calibrated multi-camera views, both for position and orientation estimation.

The remainder of the paper is organized as follows: Sec. 2 describes the overall system architecture, from hardware setup to software framework. Our 3D appearance modeling is then followed by in Sec. 3. The proposed orientation estimation approach is described in Sec. 4. Sec. 5 describes and discusses the experimental results, and Sec. 6 concludes the paper, proposing future development roads.




(a) (b)

**Fig. 1.** (a) Hardware setup. (b) 3D human body model.

## 2 System Overview

In this section, our system for human body orientation estimation is described, starting from the hardware and software setup, followed by details on the specific components that are involved in the implementation.

Our hardware setup is depicted in Fig. 1(a). Four *uEye* USB cameras, with a resolution of $752 \times 480$, are mounted overhead on the corners of the ceiling, each of them observing the same 3D scene synchronously from different viewpoints, thus providing an informative measurement set. Furthermore, all cameras are connected to one multi-core PC. A necessary step to get accurate 3D information, is the recovery of intrinsic and extrinsic camera parameters, that we perform with the Matlab Calibration Toolbox [1], with respect to a common *world* coordinate system located on the floor.

Furthermore, Fig. 2 outlines the flow chart of our approach. After acquiring a frame from the four cameras, an edge-based detection using hierarchical grids [10] is performed, to obtain the location of each person in the scene. Meanwhile, in order to avoid collecting background pixels into the appearance model, a fast GPU-based foreground segmentation is utilized. Subsequently, the 3D appearance is reconstructed by applying pixel colors to the predefined body geometry, which makes our approach independent of the camera viewpoint. Also, the appearance model can be rendered onto 2D image plane in arbitrary poses, thereby generating multiple 2D templates, which then can be matched to the segmented foreground image through pixel-level measurements, leading to the orientation estimation result.

## 3 Appearance Model

If we take into account the flattened shape of a human body along the depth dimension, we can approximate the overall geometry by three cylinders with

[1] `http://www.vision.caltech.edu/bouguetj/calib_doc/`

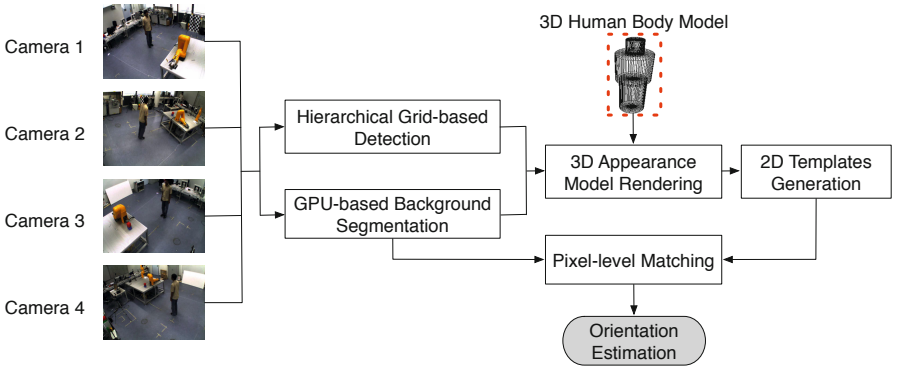



**Fig. 2.** Flow chart of the proposed approach

elliptical section, as illustrated in Fig. 1(b), enclosing the head, the torso and the legs.

The body pose is given by the person location $(t_x, t_y)$ and orientation $\theta_r$, about an axis perpendicular to the ground plane. In our system, the location is independently provided by a multi-camera, multi-people detector and tracker [10], therefore our aim is to estimate only $\theta_r$.

To this aim, a 3D appearance model is reconstructed by back-projecting pixels from each 2D image onto the respective surfaces, at the estimated location. During this phase, in order to obtain the reference orientation, the person is assumed to be more or less at the center of the observation space, and facing a certain direction. By knowing the full pose, each pixel can be back-projected onto a 3D point on the model surface.

In order to avoid collecting also background pixels, we also perform a background subtraction, with a GPU-based method proposed by Griesser et al. [11], which performs an iterative solution on a $3 \times 3$ neighborhood at each pixel, also incorporating darkness compensation. Fig.3(b) shows an example of this algorithm.

The result is a large set of $N$ model points $x_n$ (Fig.3), including position, color values, and local surface normals from the underlying 3D surface. This constitutes a sparse appearance model, that we denote by

$$M \equiv \{(x_1, v_1, n_1), \cdots, (x_N, v_N, n_N)\}\,. \tag{1}$$

## 4 Orientation Estimation

Once the sparse point cloud is available, pixel-level measurements can be obtained by reprojecting the appearance model onto the image planes, through a $(3 \times 4)$ projection matrix onto camera $k = 1, \ldots, 4$

$$y = K_k T_{kw} T_{wo} x_o \tag{2}$$

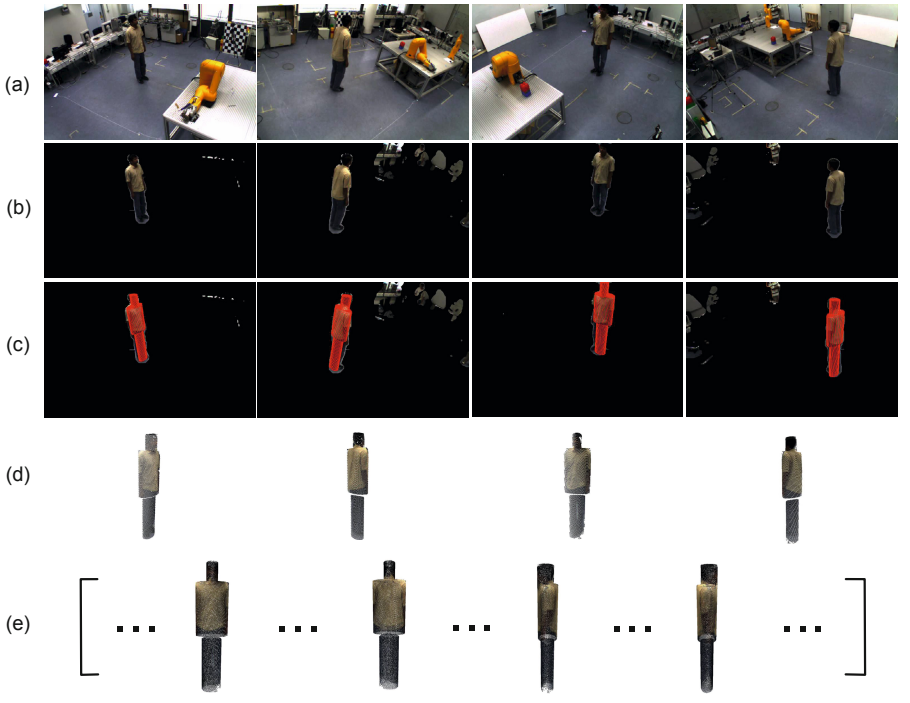



**Fig. 3.** Appearance model reconstruction. (a) Original input frames from 4 views. (b) Corresponding foreground images. (c) Detected target, with geometry model superimposed onto foreground images. (d) Back-projected partial 3D cloud, at each view. (e) Final 3D appearance model, covering 360°, with some key-poses shown.

where $x_o$ is a local model point, in homogeneous coordinates, $y$ is the corresponding image pixel, $K$ is the intrinsic camera projection, known from off-line calibration

$$K = \begin{bmatrix} f_x & 0 & p_x & 0 \\ 0 & f_y & p_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{3}$$

with $f_x, f_y$ the focal lengths, $(p_x, p_y)$ the principal point, and $T_{kw}$, is the camera-to-world constant transform, also known by calibration. Finally, $T_{wo}$ is a homogeneous $(4 \times 4)$ transformation matrix that represents the person location

$$T_{wo} = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \tag{4}$$

where $t = [X, Y, Z]^T$ is the 3-dimensional translation, and the $(3 \times 3)$ rotation matrix $R$ is expressed in terms of $XYZ$ Euler angles (of which only $\gamma$ is updated by the orientation estimation):

$$\begin{aligned} \theta_r &= [\alpha, \beta, \gamma]^T \\ R(\theta_r) &= R_x(\alpha) R_y(\beta) R_z(\gamma) \end{aligned} \tag{5}$$

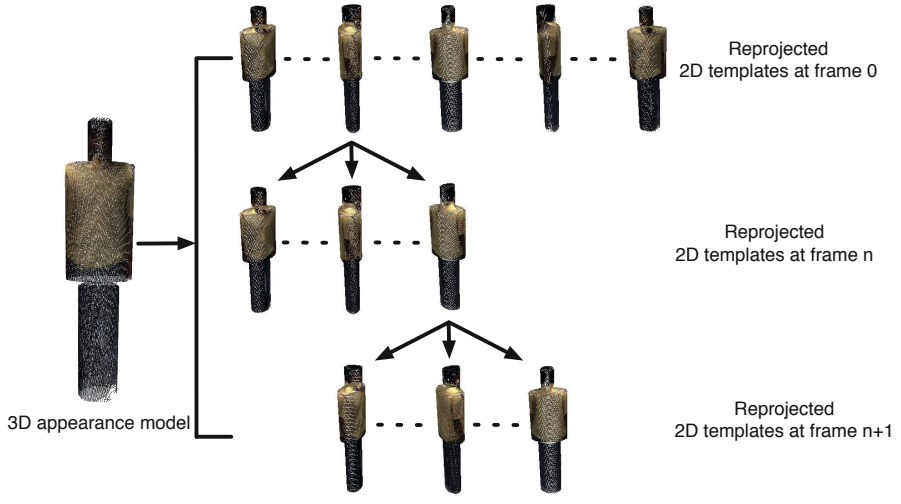



**Fig. 4.** Planar templates are obtained by reprojecting the 3D appearance model in different poses, from different camera views

Given a model point $(x_o, y_o, z_o)$ in world coordinates, the corresponding location $(x_s, y_s)$ in camera coordinate can easily be obtained via (2). During the reprojection, it is worth noting that, due to the rotations of the body, the visible part of the sparse 3D point cloud should change, only a roughly 180° slice of the point cloud is visible in any particular frame, that corresponds with the visible portion of the human body in each video frame.

Thus, it is necessary to test visibility of each point: since the shape is almost everywhere convex, a point $p$ is visible from camera $c$ if the angle between its normal and the camera projection ray through $p$ is less than 90°, that is

$$V_c \cdot n_c < 0 \tag{6}$$

where $V_c$ is the viewing vector (i.e. the position of $p$ in camera coordinates) and $n_c$ is the respective normal direction.

The point projection (2) and visibility test (6) are done at each pose hypothesis $\theta_r$. Fig. 4 provides an example of re-projected templates on one camera view.

Subsequently, template matching simply amounts to evaluate a likelihood function, by comparing color values of the templates with the underlying foreground pixels. For a predicted pose $\hat{T}_t$ at time $t$, our similarity measure between the $N_k$ color pixels $u$ from the reprojected template $h_k(\hat{T})$ onto camera $k$, and the corresponding pixels $v$ of the foreground image, is defined as

$$D = \sum_{k=1}^{4} \frac{1}{N_k} \sum_{u \in h_k} \sqrt{\sum_{c \in (r,g,b)} (u_c - v_c)^2} \tag{7}$$

which is the sum of absolute pixel-wise difference over $(r, g, b)$ channels.

In this formula, the inner sum is an isotropic $L_1$-norm that, compared to classical $L_2$-norm, it is more robust to outliers, such as non-Gaussian noise or erroneous colors sampled from the background. Therefore, the template likelihood corresponds to a Laplacian distribution

$$P(z|s) = \frac{1}{2\sigma} \exp\left(-\frac{D}{\sigma}\right) \tag{8}$$

where the pose is described by $s = (t_{ref}, \theta_r)$, and $\sigma$ is the precision parameter of this distribution.

During orientation estimation, we employ a simple but effective prediction mechanism for computational effiency. As during normal walking, it is unlikely for a person to turn more than 90 degrees over one frame of the sequence; therefore, after the initial detection and modeling phase, reprojection and matching are performed only on the orientations which are in a fixed range around the former estimation, thus saving computation while reducing estimation error. Fig. 4 illustrates this strategy across frames.

## 5 Experimental Results

We evaluate our orientation estimation algorithm through two video sequences, showing multiple people that move and turn freely, as well as interacting and occasionally occluding each other in some views. The sequences have been simultaneously recorded from all cameras, as described in Section 2, with a resolution of $(752 \times 480)$ and a frame rate of 25 fps.

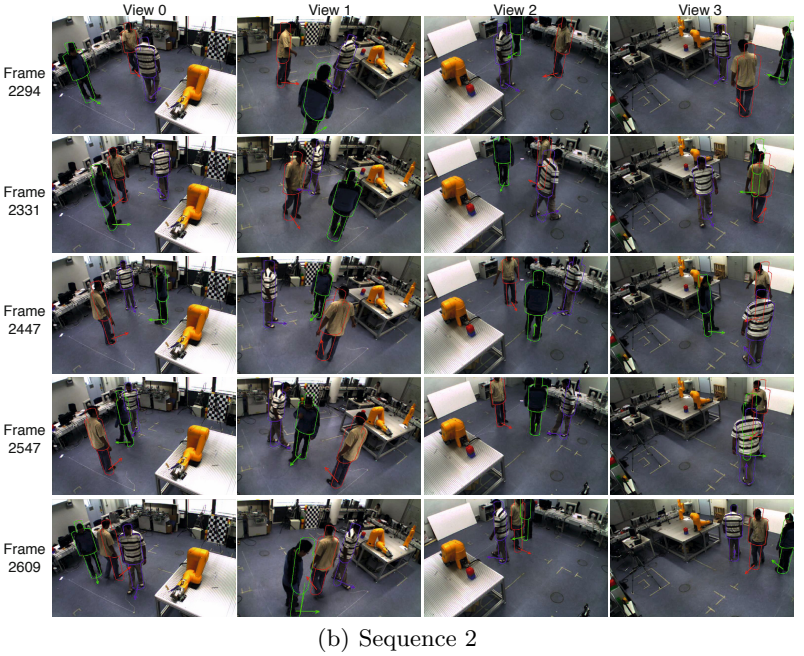
The implementation is done in C++ on a desktop PC with Intel Core-2 Duo CPU ($1.86GHz$), 1GB RAM and an Nvidia GeForce 8600 GT graphic card. Before estimating orientations, we run the hierarchical grid-based detector [10] over the sequences, producing quite reliable target locations on ground plane, which are provided as position reference $t_{ref}$.

Furthermore, during the initialization phase of our system, the 3D appearance model of each target is automatically reconstructed according to the detected 2D location and known orientation, using the technique described in Section 3. During likelihood evaluation, for computational effiency, 12 discrete orientations covering 360° are utilized during the reprojection of the appearance model onto the image planes, while at subsequent frames the model is reprojected only within 90° around the previous estimate.

The first sequence involves two targets, which interact once by shaking hands, then randomly walk around the room. The second sequence involves three targets that are strongly occluded in some views, and occasionally remain static for a few frames.

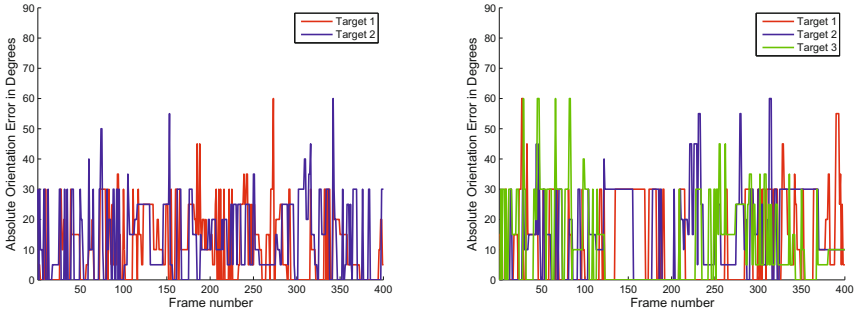
Results of both experiments are shown in Fig. 5, where the orientation of each target is indicated by a colored line, pointing to the estimated direction. The sihouette of the geometry model is also superimposed onto each target, to illustrate the tracked location that is used as reference.


(a) Sequence 1




(b) Sequence 2

**Fig. 5.** Performance of orientation estimation on four camera views




(a) Sequence 1 (b) Sequence 2

**Fig. 6.** Ground truth evaluation

In particular, as can see from Fig. 5(a), during hand shaking (between frame 1127 and 1141), the two orientations are both correctly estimated. Moreover, in Fig. 5(b), we emphasize the challenges due to mutual occlusions, from one or more views. At frame 2609, although people keep very close to each other, our estimation results are still satisfactory. Around frame 2447, all three targets are almost static throughout several frames, however our algorithm successfully estimates their orientations.

In order to evaluate more precisely the performances of our approach, we also manually labeled ground-truth data for each frame of the sequences, by rendering the 3D body model and visually matching it with the target area, in all views where the person can be seen. The most challenging clip, covering 400 frames from both sequences, was selected for ground truth evaluation, and results are shown in Fig. 6, that shows the absolute error between estimated orientation and ground truth for each person.

We can see that orientation errors are most of the time below 30 degrees. As noticed that in our framework, the likelihood is evaluated on a discrete state-space with an interval of 30 degrees, while ground truth data are labeled with an interval of 5 degrees.

## 6 Conclusion

In this paper, we presented a robust algorithm for estimating the body orientation of multiple people simultaneously in a calibrated multi-camera environment. Our method uses a generic 3D human body shape model together with a distinctive appearance model, and employs multiple 2D templates for matching with a robust likelihood function. Experiments over real-world sequences have been performed and also evaluated against ground-truth data.

The proposed methodology can be easily applied to different camera setups and different indoor environments. Future work involves increasing processing speed, robustness and versatility, for example including additional features, such

as motion or edges, in the template likelihood function. In addition, the system output can be used for higher-level tasks, such as trajectory analysis for understanding behaviors, as well as human-robot interaction.

## References

1. Khan, S., Javed, O., Rasheed, Z., Shah, M.: Human tracking in multiple cameras. In: Proceedings of the 8th IEEE International Conference on Computer Vision, Vancouver, Canada, pp. 331–336 (2001)
2. Okuma, K., Taleghani, A., de Freitas, N., Little, J.J., Lowe, D.G.: A Boosted Particle Filter: Multitarget Detection and Tracking. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 28–39. Springer, Heidelberg (2004)
3. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. International Journal of Computer Vision 75, 247–266 (2007)
4. Leibe, B., Schindler, K., Van Gool, L.: Coupled detection and trajectory estimation for multi-object tracking. In: International Conference on Computer Vision (2007)
5. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2010)
6. Lee, M.W., Nevatia, R.: Body part detection for human pose estimation and tracking. In: IEEE Workshop on Motion and Video Computing (2007)
7. Yao, J., Odobez, J.: Multi-camera 3d person tracking with particle filter in a surveillance environment. In: 8th European Signal Processing Conference, EUSIPCO (2008)
8. Gandhi, T., Trivedi, M.: Image based estimation of pedestrian orientation for improving path prediction. In: IEEE IV Symposium, pp. 506–511 (2008)
9. Chen, C., Heili, A., Odobez, J.: Combined estimation of location and body pose in surveillance video. In: IEEE Conf. on Advanced Video and Signal Based Surveillance, AVSS (2011)
10. Chen, L., Panin, G., Knoll, A.: Multi-camera people tracking with hierarchical likelihood grids. In: Proceedings of the 6th International Conference on Computer Vision Theory and Applications, pp. 474–483 (2011)
11. Griesser, A., Roeck, D.S., Neubeck, A., Van Gool, L.: Gpu-based foreground-background segmentation using an extended colinearity criterion. In: Proc. of Vison, Modeling, and Visualization (VMV), pp. 319–326 (2005)