Lying Pose Recognition for Elderly Fall Detection

Simin Wang
NIMTE institute of
Chinese Academy of Science
Email: wangsm@nimte.ac.cn

Salim Zabir
France Telecom, Japan
Orange Labs, Tokyo

Bastian Leibe
UMIC Research Centre
RWTH Aachen University

***Abstract*—This paper proposes a pipeline for lying pose recognition from single images, which is designed for health-care robots to find fallen people. We firstly detect object bounding boxes by a mixture of viewpoint-specific part based model detectors and later estimate a detailed configuration of body parts on the detected regions by a finer tree-structured model. Moreover, we exploit the information provided by detection to infer a reasonable limb prior for the pose estimation stage. Additional robustness is achieved by integrating a viewpoint-specific foreground segmentation into the detection and body pose estimation stages. This step yields a refinement of detection scores and a better color model to initialize pose estimation. We apply our proposed approach to challenging data sets of fallen people in different scenarios. Our quantitative and qualitative results demonstrate that the part-based model significantly outperforms a holistic model based on same feature type for lying pose detection. Moreover, our system offers a reasonable estimation for the body configuration of varying lying poses.**

I. Introduction

Among the major concerns of elderly care, detection of a fall bears immense importance. It can sometimes become a matter of life and death for people living alone, the elderly in particular. It has been reported by CDC [1] that in the US, every year among the 65 years or older people, one in every three experiences a fall. Each year there are cases of 300,000 hip fractures due to fall. More alarmingly, one in every five people breaking their hip die within a year of the fracture. In most of these cases, the damage is due to need of action for calling help from the fallen person's side which is not practical in all the instances.

To address this issue, fall detection systems relying on specialized sensors [24, 14, 11, 13] or with fixed cameras [20, 23] have been proposed. However, they provide neither necessary accuracy in detection nor sufficient protection to privacy. Intelligent robots, usually perceived to be equipped with a camera, can ameliorate these issues due to the relative flexibility that they can offer in terms of positioning. In addition, they are more acceptable to the user, since the humanoid characteristics incorporated in robots allow a smoother interaction with humans.

Considering the limited processing power of current heath-care robots, the problem domain we explore is fall detection in still images captured by robot mounted cameras. In this context, a fall is addressed as a lying pose. In this paper, we therefore propose a novel solution for detection of a fall through lying posture recognition based on mobile robot mounted camera images.

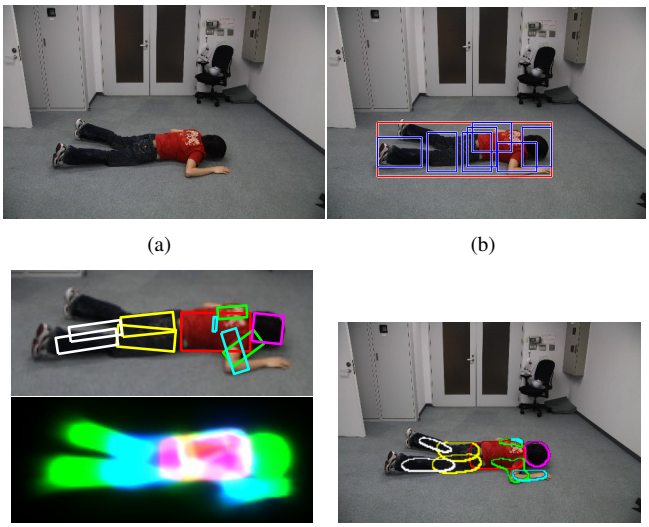


Fig. 1. *Visualization of our proposed pipeline for lying posture recognition. Given an input image (a), we first detect fallen people in different orientations using an extended Deformable Part-Based Model (b). The detected person bounding box (in red) and part locations (in blue) are then used to infer initial limb sizes and locations (c), which are again refined in a Pictorial Structures framework in order to estimate the person's detailed body pose (d).*

However, the challenges for lying pose estimation are huge. Apart from the challenges in upright human detection and pose estimation, lying pose recognition has a much larger degree of foreshortening distortion in perspective than other poses. This makes pose estimation extremely difficult. There are mainly three issues: Firstly, lying poses suffer from the perspective foreshortening effects, so that we lose important cues for detection, *e.g.* the typical head-body omega shape and horizontal body symmetry. Secondly self occlusion and cluttered scenes are frequent cases in the fall detection context. Last but not least, it is difficult to obtain good training data, which ideally spans the variability of person appearances, body poses, and background structures.

Our approach builds upon previous work for part-based detection [7] and pose parsing [8], but adapts and extends it in order to deal with the many challenges associated with lying postures. Our three main contributions reported in this paper are as follows. (1) We extend the detection algorithm to multiple viewpoints and strong perspective foreshortening effects. We explore how an upright person detector needs to be adapted for lying postures under different viewing angles and

devise optimized training procedures for the learning phase of the new detector from few training examples. (2) We propose an algorithm to perform multi-viewpoint pose parsing based on the part-based detection results. Our approach is inspired by the upright upper-body parsing approach of Ferrari *et al*. [8]. The unique property of our approach is that it is capable of parsing a full-body model subject to viewpoint changes and corresponding perspective foreshortening. (3) Since perspectively foreshortened limb sizes become a real problem in the multi-viewpoint case, we propose a method to infer the limb locations and sizes directly from the part-based detection results. This step takes advantage of the Pictorial Structures model available in the detector and naturally connects it to the Pictorial Structures model used for pose inference. We show that the resulting pose estimation performance is superior to the performance of a baseline approach just based on location and segmentation priors.

We experimentally demonstrate that our proposed approach achieves good results for lying posture detection and pose analysis in a variety of challenging settings, as they would be observed by a mobile service robot.

The rest of this paper is organized as follows. Sec. II describes an overview of related works. Sec. III presents the detailed description of our proposed mechanism. We summarize the results from our experiments and discuss observations from them in Sec. IV. Finally, in Sec. V we conclude along with presenting some ideas for future research.

II. Related Works

Many existing approaches attempt to provide the target person with a wearable sensor, such as accelerometers [24] and gyroscopes [14], or even a fusion of multiple sensors [11, 13]. Despite best efforts at stylish and comfortable design, creating user acceptance for such wearable solutions is however still an open issue.

Another stream of solutions is based on computer vision. The use of wall or roof mounted cameras have been proposed in the literature [20, 23, 10]. Miaou *et al*. [20] make use of personal information and try to ensure 360 degree coverage by using omnidirectional surveillance cameras. Williams & Hanson [23] handle the issue of complete coverage by overlapping multiple camera views. However, the use of fixed cameras can be unreliable due to occlusions by furniture items and limitations of the camera viewpoint. In addition, they can be perceived as agents for direct intrusion into privacy.

In approaches based on background modeling [23, 10, 22], decisions of fall detection mainly depend on blob analysis. Williams & Hanson [23] simplify detection of fall as distinguishing non-upright human blobs through size and shape analysis. However, similar planned motions, *e.g*. sitting down, are easy to trigger false alarm in such approaches. Hazelhoff [10] incorporates a head tracking module to enhance detection performance, which is again sensitive to the scene occlusion. We therefore propose to pursue a more detailed analysis by performing full-body pose estimation.

To support the mobility of heath-care robots, our system is designed to recognize lying poses in monocular images without any knowledge about the background or the person's clothing. For this, we build upon recent advances in human detection [3, 12, 7]. To date, however, state-of-the-art works on human detection mainly focus on upright persons. Adapting those approaches to lying poses is non-trivial due to the effects of different body orientations and perspective foreshortening. In Section III-B, we therefore propose an extension of the Deformable Part Based Model by [7] that can recognize lying persons in different orientations.

Several approaches have been proposed for estimating the body pose of upright humans [6, 18, 8, 2, 4]. Ferrari *et al*. [8] propose a coarse-to-fine pipeline that progressively reduces the search space in order to achieve more reliable pose estimation results. The first step is upper body detection, which results in a primary interest region, followed by a foreground segmentation step. This way, the later pose parsing steps [18, 6] are guided to a region focusing on the detected human and with most background removed. This leads to a much faster and robust estimation, though occlusion is still not well modeled. However, this approach heavily relies on expected part locations for the head and upper body in order to initialize the foreground segmentation. As those part locations can significantly vary for lying persons in different orientations, it is therefore not directly applicable here. Moreover, the greater pose variability of lying postures, together with background clutter, introduce additional challenges.

In this work, we therefore propose an extension of Ferrari *et al*.'s pipeline that is targeted at the specific challenges of lying postures. In particular, we propose an inference procedure that takes advantage of the detected part locations from the viewpoint-specific object detector in order to provide an initialization for the limb locations and sizes, taking into account the effects of perspective foreshortening (see Section III-D). As our experiments show, this step significantly improves the pose estimation results and our approach's robustness to clutter.

III. Approach

A. Overview

We formulate the problem of fall detection as lying posture recognition in still images. State-of-the-art pose estimation approaches [6, 18, 8, 2, 4] provide a detailed representation of human bodies, *i.e*., a limb tree, but are likely to suffer from high computation complexity due to the huge searching space. Inheriting the concept of [8], we progressively reduce the searching space by adding two precedent stages, *i.e*., detection and segmentation. More importantly, we take one step further by exploiting information from detection and segmentation stages. Through intensive experiment, we discover that Deformable Part Based Model by [7] is a 'degraded' version of Pictorial Structure for pose parsing [8, 18]. Detections reported by Deformable Part Based Model might imply true limb locations with a much lower computation cost. The main contribution of our approach lies in the way how we model this connection, as shown in Section III-D.

What we propose is a coarse-to-fine pipeline to achieve a more and more detailed representation for lying poses (see Fig. 1). In the first stage, we perform lying pose detection on the input image by searching for bounding box hypotheses containing lying persons over image locations and scales. Humans are described by a part-based star model [7] with a root filter representing the rough outline shape and a set of part filters capturing the important shape details. The detection score is modeled as a sum of appearance confidences from both root and parts, reduced by a deformation cost caused by part displacement. A multiple-component mixture model is trained to separate distinct viewing angles or principal body orientations.

Within each bounding box, we extract a foreground mask via GrabCut [19]. In the context of health care applications, we can make the reasonable assumption that at most one lying person is visible in the robot's field of view. Based on this assumption, we focus on the detailed pose analysis of the foreground image resulting from one hypothesis with the highest confidence. Here, the confidence combines factors from both detection score and foreground probability.

In the final stage, we perform pose estimation based on the detected person location and the estimated foreground segmentation mask. For this, we adopt the pose parsing framework proposed by [16, 8] and extend it to the multi-view case. This framework represents the human body configuration by a tree-shaped body model, which lends itself to an efficient inference procedure [6].

The challenges of lying posture analysis necessitate several important changes to this framework. In order to cope with the effects of perspective foreshortening on the limb configuration, we need to keep separate body models for the different viewpoints. In addition, the larger pose variability of lying postures raises the importance of starting pose inference from a good initialization. A fixed limb location prior, as used in previous papers [16, 8] is no longer sufficient here. We therefore propose to infer the initial limb positions and sizes from the detected part locations of the object detector. As our results show, this step significantly improves the pose estimation results. This initialization also generalizes the pose estimation framework [8] by reducing the number of hard coded parameters.

B. Lying Posture Detection

The human detector proposed by [7] is used for lying pose cases in our system. The basic concept is to represent an object as a star model, which is a collection of rigid patches, named **parts**, whose location is defined independently with respect to a central root part. Looser part placement constraints tolerate local translation and deformation of parts to a certain extent, which offers a benefit when dealing with articulated objects. On the other hand, spatial and visual cues are well fused in a statistical sense.

A model θ with N parts is defined as

$$\theta = (f_0, f_1 \ldots, f_N,\ v_1, \ldots, v_N,\ d_1, \ldots, d_N)$$

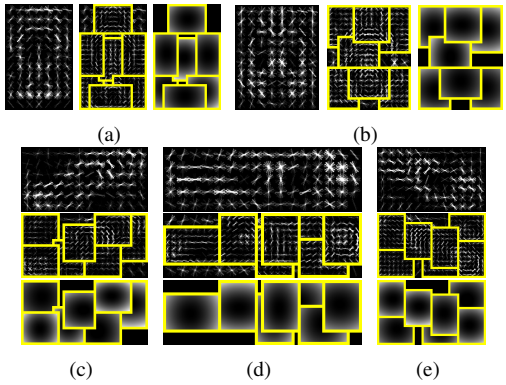


Fig. 2. *Visualization of the seven-parts deformable mixture model for some of the viewpoint classes of lying poses used in our approach. From left to right (a), (b) or from top to bottom (c), (d), (e): the root filter, part filters and corresponding deformation functions.*

Besides the root filter f_0, each part i holds a 3-tuple of parameters, *i.e.* a linear HoG weight filter f_i, an ideal anchor position with respect to the root location v_i, and deformation function coefficients d_i.

Given one image I, a HoG feature pyramid $\mathbf{P}$ is extracted. All possible configurations of object hypotheses $L = \{l_0, \ldots, l_N\}$ are evaluated, where l_i is a 3-dimensional vector in $(x, y, scale)$ space. Linear filters $\{f_0, \ldots, f_N\}$ play a role as shape templates to encode part appearance information. The observation likelihood $P(I|L, \theta)$ is proportional to the product of the response of all part filters in their corresponding locations. The geometric relation between parts and the root $P(L|I, \theta)$ is incorporated into the score function as a quadratic cost.

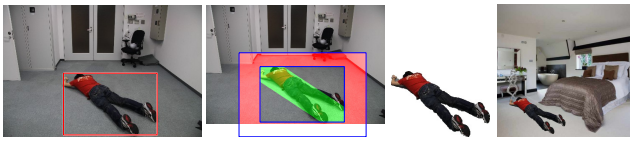
For the detection task, the root location score is formulated as a maximization of object hypothesis scores over part displacements. Maximization is performed for each part independently:

$$S_\theta(I, l_0) = \max_{\{l_1,..,l_N\}} \left\{ f_0 * \mathbf{P}|_{l_0} + \sum_{i=1}^{N} f_i * \mathbf{P}|_{l_i} - d_i \begin{bmatrix} dx \\ dy \\ dx^2 \\ dy^2 \end{bmatrix} \right\}, \quad (1)$$

where $(dx, dy)^\top$ is the location of l_i with respect to its ideal position, *i.e.* $2l_0 + v_i$. Note that part locations and appearances are represented at twice the resolution of the root part.

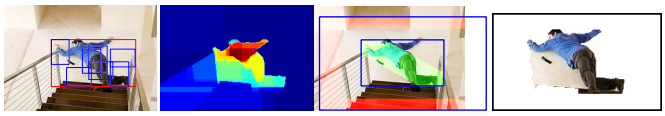
Multi-Viewpoint Detection. In the context of our application, we can assume that a service robot's camera will be mounted at a fixed height and tilt angle. We therefore need to design the detector such that it can recognize fallen people from this perspective. The main difficulty here is the variability of possible body orientations on the floor, which leads to significantly changed person appearance through perspective distortion. This is in contrast to standard human detection tasks, which can rely on the presence of upright body shapes [3, 12, 7].

In order to cope with this difficulty, we propose to use a mixture of 8 distinct detector models, each dealing with one body orientation class covering a $45°$ interval. Our mixture



(a) (b) (c) (d)

Fig. 3. *Synthetic training data creation by background substitution. (a) the original image with annotated bounding box. Based on this box and our object characteristics, the initial mask (b) contains regions marked as foreground (green), background (red) or unknown (no label). (c) the resulting foreground image after two iterations of segmentation assisted by user interaction. (d) a realistic synthetic sample is created by superimposing this foreground onto other floor textures. An optimal substitution center is given by fitting the foreground mask to the new ground plane.*



(a) (b) (c) (d)

Fig. 4. *Reliable foreground segmentation guided by detection. Given the bounding box l_0 from detection (a), the initial mask is defined according to the body orientation class c (c). Accumulating foreground masks from full set of hypotheses gives an overall foreground map $\mathsf{A_f}$ (b) . The darker red color indicates a higher foreground probability of this pixel. Combining (b) and (c), one additional run of Grabcut finalizes segmentation, resulting in a much reduced region of interest (d).*

model with C components is defined as $\Theta = (\theta_1, \ldots, \theta_C)$. We train each individual model component as a separate latent SVM, following the framework of [7], and merge all components into the final mixture. Five components are trained individually to represent the body orientations from degree 0 to 180 (with respect to the y axis of the camera view plane). The other three components $\{225°, 270°, 315°\}$ are obtained by mirroring components $\{135°, 90°, 45°\}$ along the y axis. Each component consists of three elements: the root filter, the placement of the part filters within the root box, and a quadratic function as the deformation penalty. See Fig. 2 for the resulting detector models.

All components perform the detection independently, *i.e.*, the detector fires if any component reports a score higher than its threshold at this root location. Thus, the decision function of a mixture model with C components (Eq. 1) is extended as

$$\mathbf{H} = \{\mathbf{h} = (L, c, S_{\theta_c}(I, l_0)) \,|S_{\theta_c}(I, l_0) \geq T_c\}, \qquad (2)$$

where $\mathbf{H}$ is a list of detection hypotheses $\mathbf{h}$ in image I.

This is again a deviation from the approach of [7], which may also include several mixture components to represent different object aspects (*e.g.* upper body *vs.* full body). As those aspects may share visual characteristics and multiple aspect components may thus be active for the same object, [7] merges the component responses into a single detector response. In contrast, our model uses the components to represent different body orientation classes or viewing angles, which are unlikely to share visual characteristics. We therefore do not perform such a component merging step.

Training Set Enrichment. In order to boost recognition performance, we employ similar schemes as the ones proposed in [21] to create synthetic variants of the original training dataset. Together, this artificial variation enriches our training set by a factor of 120. This includes

Background Replacement. This is a key mechanism in training set enrichment. A foreground mask is obtained from an interactive segmentation tool using *Grabcut* [19]. Based on this, the exact body shape is extracted and superimposed onto other backgrounds. Besides this, we find an optimal position as well as a proper scale according to the ground plane of new background images to make generated synthetic samples more realistic (See one sample in Fig. 3).

Training Sample Shearing. Shearing samples by a small angle, chosen as 3° in our experiments, is a traditional means to augment training sets. To obtain more realistic instances, we perform shearing along the y axis.

Mirror Orientation Association. We associate the horizontally mirrored model to the corresponding body orientation bins. This way, each sample also serves the bin corresponding to its horizontally flipped version.

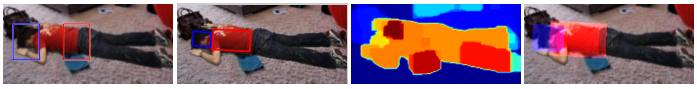
C. Foreground Segmentation

Starting with an initial label mask, Grabcut [19] efficiently extracts a region of interest from the original image. In this paper, we perform an unsupervised version of Grabcut [19] to segment potential foreground regions (See one sample in Fig. 4). Given the detected component index c in hypothesis $\mathbf{h}$, an initial mask is defined according to the corresponding body orientation class. Alternatively, we consider detected part boxes as potential foreground as well. Grabcut performs hard segmentation on a proportionally enlarged version of l_0, which results in a binary mask $\mathsf{F}(\mathbf{h})$.

Foreground segmentation not only reduces the search space for later pose parsing, but also verifies detection hypotheses, in a similar manner as [17]. Recall that a detection hypothesis $\mathbf{h}$ has a representation of the best configuration L, the component index c and its detection score $S_{\theta_c}(I, l_0)$ (Eq. 2). Different from [8], segmentation is performed on the whole set of hypotheses $\mathbf{H}$ individually, instead of only on the best one. Our rationale is that the definition of initial label masks implies the region color coherence. It is less likely that false alarms will have similar foreground/background separation, though they preserve edge characteristics close to our object class.

The resulting foreground masks are accumulated by weighting their detection scores. Detections from the same image I are validated on the basis of this accumulated foreground map, *i.e.*, those bounding boxes having a significant overlap with this accumulated map $\mathsf{A_f}(I, \mathbf{H})$ are more likely to be true detections. This motivates us to re-rank detections by a weight $w_s(\mathbf{h})$, as defined in Eq. 3

$$w_s(\mathbf{h}) = \frac{\mathsf{A_f}(I, \mathbf{H}) \cdot \mathsf{F}(\mathbf{h})}{\sum_p (\mathsf{A_f}(I, \mathbf{H}))} \qquad (3)$$



(a) (b) (c) (d)

Fig. 5. *(a) The first two parts detected. (b) Each limb is predicted by the full set of detected parts via a linear mapping function. Accordingly, the edge model is adjusted to cope with perspective foreshortening effects,* e.g. *head (blue) and torso (red). (d) Adaptive color model initialization. Given the foreground map from segmentation (c) and potential torso and head boxes (b), two region masks are obtained, which lead to more effective color model learning.*

Note that we assume only a single object per image. Thus, there is no normalization over the number of hypotheses.

D. Lying Pose Parsing

The Pictorial Structure model [6] represents the human body as a tree-structured collection of limbs. To distinguish those from parts in the detection stage, limb configuration is written as $V = \{v_1, \ldots, v_n\}$. Given an input image I and the human model ϑ, we search for an optimal configuration that maximizes the posterior $V = \arg\max(P(V|I, \vartheta))$. In the full-body cases, $n = 10$. The tree structure is denoted as E.

The appearance evidence is encoded in each node as $\mathcal{A}(v_i)$. Any two parts are connected if edge $e_{i,j} \in E$. $\mathcal{S}(v_i, v_j)$ captures the geometric constraints between them. The posture posterior is written as a distribution including the contributions from both appearance and spatial information.

$$P(V|I,\vartheta) \propto \underbrace{\prod_{i=1}^{n} \mathcal{A}(v_i)}_{\text{Appearance}} \cdot \underbrace{\prod_{e_{i,j} \in E} \mathcal{S}(v_i, v_j)}_{\text{Spatial Configuration}} \tag{4}$$

Model inference is done by two-pass sum product belief propagation. The first pass starts from the leaves and terminates in the root, whereas the second pass goes the other way around. This procedure gives us the marginal posteriors for all limbs. Pose parsing is performed iteratively so that we have a better guess for the limb regions. The limb location prior is updated by the previous pose posterior, which means pose parsing guides future color model learning, and this again refines the pose posterior.

We extend the parsing framework from [8] to enable multi-view and full-body pose estimation. One critical drawback of the approach proposed in [16, 8] is that no explicit search in scale space is performed for the different limbs, which leads to a rigid edge model in both limb sizes and limb center positions. This constraint might be still acceptable in upright person cases, since they have a regular proportion between limbs. Unfortunately, in lying pose cases, perspective distortion in limb sizes occurs frequently, and also limb anchors with respect to their parent limbs vary as body orientation changes, due to perspective foreshortening. We present two improvements to compensate this weakness in scalability, mainly by exploiting cues from detection and segmentation.

Limb Pre-inference by Detection Parts. In our experiments, we observed a strong correlation between the detected part configuration L and the target limb configuration V. If our detector is well trained, the detected parts will imply limb locations. For instance, in our model, the first two parts roughly locate the waist and head, respectively (see Fig. 5(a)).

This can be explained by the fact that both detection and pose estimation formulate the problems following pictorial structure. Both objective functions are defined in a log-linear form, combining appearance and spatial information. Another interpretation is that Eq. 1 is proportional to a maximum approximation of the posterior distribution, given a star structure model and limbs represented as non-rotating boxes.

Driven by the connection between statistical parts and semantic limbs, a pre-inference step is incorporated to link the detection and pose estimation stages. L given by the best hypothesis $\mathbf{h}$ is utilized to predict an initial pose V_0, which is later optimized by our pose estimation framework. The inference function is modeled as a view-specific mapping matrix A_c in the transformed space, as formulated in Eq. 5.

$$V_0 = \mathsf{A}_\mathsf{c}(L_0) \tag{5}$$

Here, L_0 is the projection of L at original image scale, representing each part as upper-left and lower-right corners, whereas limbs of V_0 have one more dimension to determine their orientations. The training of A_c is rather straight-forward, via a linear regression with the input as latent detections from training samples and the output as ground truth pose annotations.

Sec. IV-D shows that starting with V_0, lying pose estimation works more effectively and yields a substantial improvement on cases with a large degree of perspective distortion.

Adaptive Color Model Initialization. Starting from limb pre-inference, region growing is performed on the foreground probability map using the predicted centers of head and torso as seeds. This step recursively evaluates new neighboring pixels and adds them if their values are close enough to the current mean. Iterated until convergence, two region masks are obtained (see Fig. 5). The initial color models of head and torso are learnt from probability distributions combining both limb priors and generated color masks.

IV. Experimental Results

In this section we present qualitative and quantitative evaluation results from the experiments. All experiments were performed on a Ubuntu 10.04 (64bits) machine with 2.8GHz quad-core Intel Core i7 CPU, 8GB of RAM.

A. Datasets

FT Lying Person Dataset. For the purpose of training, the FT database is constructed in a well-controlled environment, simulating camera conditions from a service robot scenario. It has 507 samples with only one lying person instance for each, containing 21 subjects in total. All images are taken by one monocular camera with similar camera position setting, *i.e.* a camera height of $1.2m - 1.6m$ and a tilt angle of $15°$, pointing down.

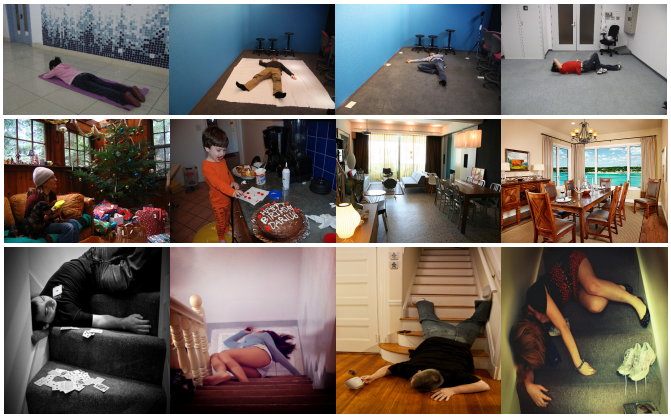


Fig. 6. *The first row: FT lying person dataset as training set. The second row: a negative set containing mainly indoor images. The last row: a challenging web image test set.*

Indoor Image Set for Background Replacement. This set contains 121 indoor images collected from the Internet.

Negative Set. In addition to the image set for background replacement, the negative set also contains 267 indoor images without persons and 72 ones with upright persons from the Pascal09 Database [5].

Test sets. Test set **A** contains a positive set split from FT and enriched by background replacement to a size of 840 images, and a negative set from the MIT Indoor Scene Database [15] with 297 indoor images covering 7 categories (*e.g.*, bedroom, living room and dining room). No human is present in this collection, which allows us to perform the comparison experiment in Sec. IV-C. In order to demonstrate our approach's generalization performance, we also apply it to Test set **B**, which contains 82 challenging lying pose images collected from the Internet.

B. Evaluation Measures

In both detection and pose estimation stages, 'Area of Overlap' criterion [5] , named **AOV** from now on, is adopted to validate hypothesis. A detection hypothesis is accepted as a true detection if its bounding box BB_p has a significant overlap with the ground truth bounding box BB_{gt},

$$a_o = \frac{area(BB_p \cap BB_{gt})}{area(BB_p \cup BB_{gt})} \geq 50\%. \quad (6)$$

In line with [5], detection performance is reported in terms of *average precision* (AP) on *precision-recall* (PR) plots.

Again, due to perspective distortion, representing poses as sticks, as proposed in [8], is no longer suitable. To evaluate pose parsing performance, we propose to measure the **AOV** between estimated limb segments and ground truth. Instead of bounding boxes, the regions of interest are limb masks, generated by segmenting the corresponding posterior maps.

C. Detection Performance

In order to evaluate detection performance, a set of models with various parameter settings are trained on the FT dataset,

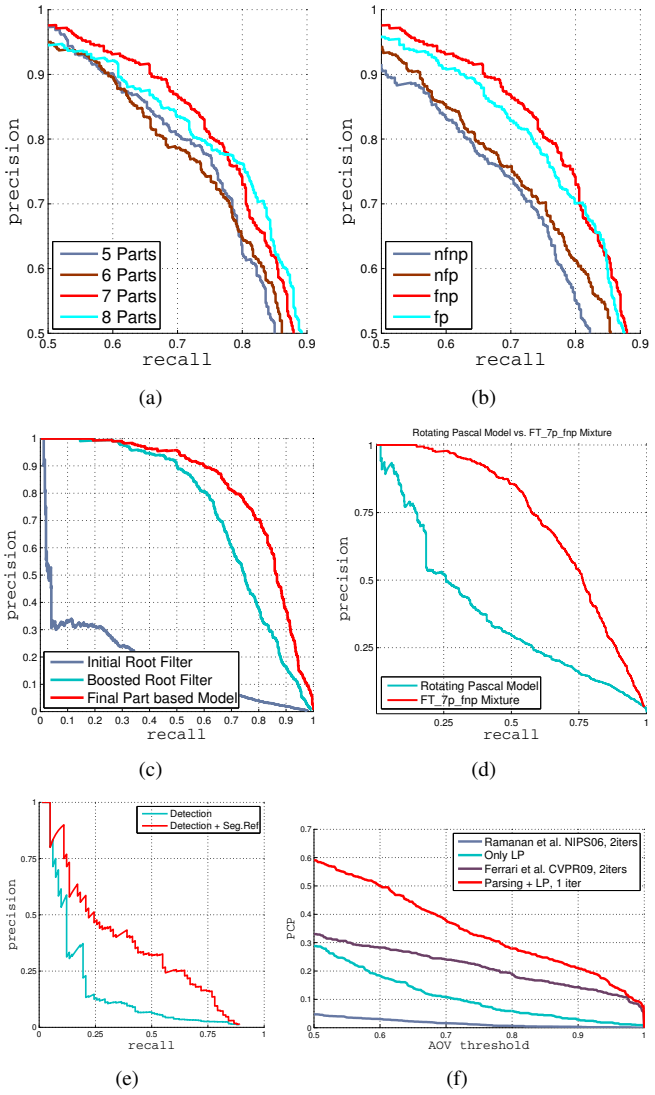


Fig. 7. *(a) (b) The mixture with seven parts has best detection performance, achieving AP = 0.81.* **f/nf** *= with/without Mirror Orientation Association.* **p/np** *= with/without Training Sample Padding. (c) Detection performance comparison of one model set. The initial root filter is equivalent to a HOG+SVM detector, which is enhanced via several rounds of boosting [7]. Adding parts further improves the performance. (d) The FT_7p_fp mixture outperforms the Pascal09 model in lying pose detection. (e) Detection performance on the more challenging Test Set* ***B****. Segmentation re-ranking contributes yields a considerable performance improvement. (f) Quantitative pose estimation performance comparison of our method against [16, 9] on a subset (body orientation class = 0°) of Test Set* ***A*** *given the same detection windows.* **LP** *= limb pre-inference.* **PCP** *= the percentage of correct limb poses. Note that the first two plots have different scales to show a closer view.*

named FT models. As Figs. 7(a) and 7(b) show, the mixture model with seven parts without feature padding and with mirror sample flipping (FT_7p_fp) performs best. The part-based model also outperforms standard HOG+SVM in lying pose detection (Fig. 7(c)).

FT models v.s. Rotating Pascal'09 Model. To study the performance degradation of a general person model on lying poses, we conduct a detection experiment on Test Set **A**

by rotating a model trained on the Pascal'09 person class without pose constraint. To simplify, we name this model as Pascal09 from now on. Since the the Pascal09 model is vertically symmetric and is trained mainly on upright persons, rotating is necessary for detecting multi-view lying poses. Additionally, components in the Pascal09 model represent two aspect classes, upper body and full body. In order not to miss true detections from the upper body component, a polygon annotation is used to properly validate detection hypotheses.

Testing of both the FT_7p_fp mixture and the rotating Pascal09 model on Test Set **A** results in the PR curves shown in Fig. 7(d). Very distinct accuracies are achieved, AP 0.74 from the FT_7p_fp Mixture vs. AP 0.37 by the rotating Pascal09 model. This experiment shows that the perspective foreshortening has a severe impact on detection accuracy. The Pascal09 model fails in samples with significant perspective distortion, whereas it reports aligned boxes that are closer to true body regions in true detection cases.

Segmentation-based Re-Ranking. Segmentation is utilized to link two key elements of our system, *i.e*. detection and pose parsing. As mentioned in Sec. III-C, the foreground probability map implies a better ranking of hypotheses scores. In the experiment on the very challenging Test Set **B**, detection precision is significantly enhanced by segmentation, as shown in Figs. 7(e) and 8. However, segmentation is not able to generate new hypotheses, meaning that misdetections in the first stage are unrecoverable.

D. Pose Parsing Performance

Fig. 9 shows intuitively that the improvements proposed, *i.e*. limb pre-inference by detection boxes and adaptive color model initialization, can compensate for the perspective foreshortening effect and therefore achieve reasonable pose estimation results on very challenging data. Mapping from the detected parts to predicted limbs offers surprisingly good initialization, especially for the torso and head. Moreover, since we reduce the search space progressively, a speed-up factor of more than 4 is achieved, compared to the original image parsing framework by [16].

A final quantitative parsing experiment is conducted to compare our method against state-of-the-art approaches [16, 9]. For a fair comparison, we only perform pose estimation on a subset of Test Set **A** with the principal body orientation close to the y axis of the image plane, which are analogous to upright persons. As shown in Fig. 7(f), our system outperforms the other two approaches, even with less iterations, showing that our adaptive limb prediction brings significant advantages under the high degrees of perspective distortion considered in our application.

V. Conclusion

Driven by the fast involution of vision technology in recent decades, videos and images have become more and more important information sources in robotics community. Our system alerts potential fall events through vision-based lying posture recognition in still images. The whole pipeline extends the current state-of-the-art approaches to multi-view full-body cases. Compared to upright body poses, lying postures are more challenging due to the loss of size proportion between body parts. Our main contributions are: 1) We present a viable approach to extend a part-based model to the multi-view case for viewpoint invariant lying posture detection. 2) We exploit the correlation between detected body parts, in the context of the deformable part-based model [7], and limbs, in the context of the Pictorial Structure model [6], in order to predict the sizes and locations of limbs prior to pose inference. 3) We show how the foreground map from segmentation can assist trained limb location priors to guide appearance model learning. With the help of these three adaptations, our system outperforms state-of-the-art approaches [8, 18] for lying body pose estimation.

Our system can be further improved by incorporating geometric constraints. The knowledge about a scene's ground plane can help reject false alarms that are inconsistent with scene geometry. It would also be interesting to explore the possibility of validating detection by pose estimation. For instance, the sum of pose pixel confidence gives a hint on how reliable the current detection is. On the other hand, fall detections could be validated via biological data of the user, such as ECG wave.

Acknowledgments The authors would like to thank France Telecom/Orange Labs, Japan, and RWTH Aachen University's cluster of excellence UMIC (DFG EXC 89) for facilitating parts of the research. This work is also supported by the NIMTE institute of Chinese Academy of Science.

References

[1] Wellcore unveils fall detection, activity monitor. http://mobihealthnews.com/5923/wellcore-unveils-its-fall-prevention-and-activity-monitor/.

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. In *CVPR'08*, 2008.

[3] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR'05*, 2005.

[4] M. Eichner and V. Ferrari. We are Family: Joint Pose Estimation of Multiple Persons. In *ECCV'10*, 2010.

[5] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2):303–338, 2010.

[6] P. Felzenszwalb and D. Huttenlocher. Pictorial Structures for Object Recognition. *IJCV*, 61(1), 2005.

[7] P. Felzenszwalb, D. McAllester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. In *CVPR'08*, 2008.

[8] V. Ferrari, M. Marin, and A. Zisserman. Progressive Search Space Reduction for Human Pose Estimation. In *CVPR'08*, 2008.

[9] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Pose search: Retrieving people using their pose. In *CVPR'09*, 2009.

[10] L. Hazelhoff, J. Han, and P. H. N. De. Video-based

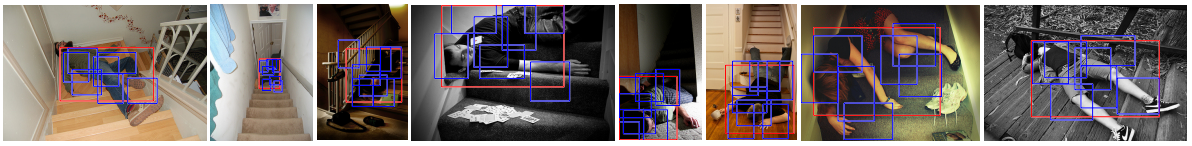


Fig. 8. *Detection Samples on Test Set* ***B***.

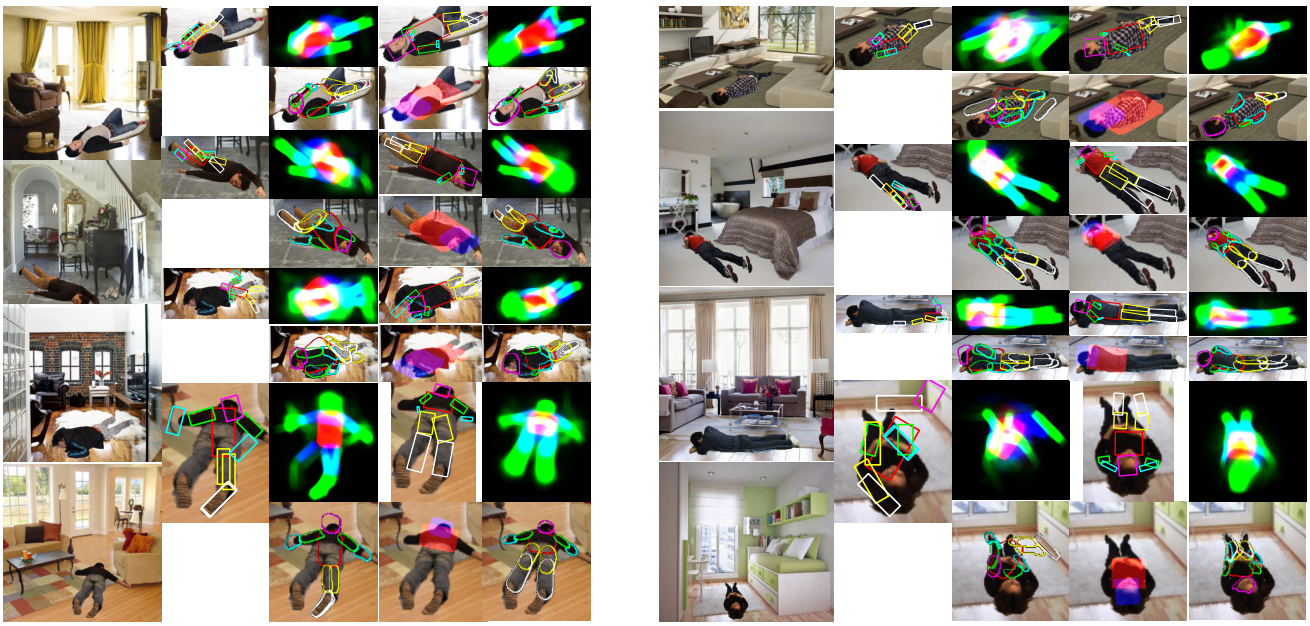


Fig. 9. *Pose parsing samples from Test Set* ***A***. *First column: original Images. Second column: baseline parsing results from [16]. Third column: an extended version of [8] to multi-view full-body pose estimation without limb pre-inference and adaptive color mask learning,* i.e. *upper: posterior; lower: final segments. Last two columns: our pose estimation results and intermediate images,* i.e. *upper-left: limb pre-inference; lower-left: the predicted head and torso masks; upper right: the posterior after two iterations; lower right: the final segments. The baseline method tends to fail in cluttered scenes. The pre-inference limbs enables a more reasonable initialization of limb sizes, thus our approach is able to handle the presence of significant changes in scale.*

fall detection in the home using principal component analysis. In *ACIVS'08*, 2008.

[11] C. Lai, H. C. Chao Y. M. Huang, and J. H. Park. Adaptive body posture analysis using collaborative multi-sensors for elderly falling detection. *IEEE Intelligent Systems*, 2010.

[12] B. Leibe, A. Leonardis, and B. Schiele. Robust Object Detection with Interleaved Categorization and Segmentation. *IJCV*, 77(1-3):259–289, 2008.

[13] A. Leone. A multi-sensor approach for people fall detection in home environment. In *ECCV'08*, 2008.

[14] A. W. Tan M. N. Nyan, F. E. Tay and K. H.Seah. Distinguishing fall activities from normal activities by angular rate characteristics and high speed camera characterization. *Medical Engineering and Physics*, 28(8): 842–849, 2006.

[15] A. Quattoni and A. Torralba. Recognizing Indoor Scenes. In *CVPR'09*, 2009.

[16] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS'06*, 2006.

[17] D. Ramanan. Using segmentation to verify object hypotheses. *CVPR'07*, 0:1–8, 2007.

[18] D. Ramanan, D.A. Forsyth, and A. Zisserman. Tracking People By Learning Their Appearance. *PAMI*, 29(1):65–81, 2007.

[19] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive Foreground Extraction Using Iterated Graph Cuts. In *SIGGRAPH'04*, 2004.

[20] P. H. Sung S. G. Miaou and C. Y. Huang. A Customized Human Fall Detection System Using Omni-Camera Images and Personal Information. In *Transdisciplinary Conference on Distributed Diagnosis and Home Healthcare*, 2006.

[21] H. Schneiderman and T. Kanade. Object Detection Using the Statistics of Parts. *IJCV*, 56(3):151–177, 2004.

[22] N. Thome and S. Miguet. A hhmm-based approach for robust fall detection. In *ICARCV'06*, 2006.

[23] A. Williams, D. Ganesan, and A. Hanson. Aging in place: fall detection and localization in a distributed smart camera network. In *MULTIMEDIA'07*, 2007.

[24] T. Zhang, J. Wang, and L. Xu et al. Using wearable sensor and NMF algorithm to realize ambulatory fall detection. *Lecture Notes in Computer Science, Advances in Natural Computation*, 4222:488–491, 2006.