

# Human Posture Tracking and Classification through Stereo Vision and 3D Model Matching

Stefano Pellegrini and Luca Iocchi  
Dipartimento di Informatica e Sistemistica  
University "La Sapienza", Roma, Italy  
E-mail: <lastname>@dis.uniroma1.it

## Abstract

The ability of detecting human postures is very relevant for applications related to the analysis of human behaviors. Techniques for posture detection and classification can be thus very relevant in several fields, like ambient intelligence, surveillance, elderly care, human-machine interaction. This problem has been studied in recent years in the Computer Vision community, but proposed solutions still suffer from some limitations that are due to the difficulty of dealing with complex scenes (e.g., occlusions, different view points, etc.).

In this article we present a system for posture tracking and classification based on a stereo vision sensor that provides both a robust way to segment and track people in the scene and 3D information about tracked people. The proposed method is based on matching 3D data with a 3D human body model. Relevant points in the model are then tracked over time with temporal filters and a classification method based on Hidden Markov Models is used to recognize principal postures. Experimental results show the effectiveness of the system in determining human postures with different orientations of the people with respect to the stereo sensor, in presence of partial occlusions and under different environmental conditions.

## 1 Introduction

Human posture recognition is an important task for many applications in different fields, such as surveillance, ambient intelligence, elderly care, human-machine interaction, etc. Computer vision techniques for human posture recognition have been developed in the last years by using different techniques aiming at recognizing human activities (see for example [10, 15]). The main problems in developing such systems arise from the difficulties of dealing with the many situations that occur when analyzing general scenes in real environments. Consequently, all the works presented in this area have limitations with respect to a general applicability of the systems.

In this article we present an approach to human posture tracking and classification that aims at overcoming some of these limitations, thus enlarging the applicability of this technology. The contribution of this article is to describe a method for posture tracking and classification given a set of data in the form XYZ-RGB, corresponding to the output of a stereo vision based people tracker. The presented method uses a 3D model of human body, performs model matching through a variant of the ICP algorithm, tracks the model parameters over time, and then uses a Hidden Markov Model (HMM) to model posture transitions. The resulting system is able to reliably track human postures, overcoming some of the difficulties in posture recognition, and in particular presenting higher robustness to partial occlusions and to different points of views. Moreover, the system does not require any off-line training phase, it just uses the first frames (about 10) in which the person is tracked to automatically learn parameters that are then used for model matching. During these training frames we only require that the person is in the standing position (with any orientation) and that his/her head is not occluded.

The approach to human posture tracking and classification presented here is based on stereo vision segmentation. Real-time people tracking through stereo vision (e.g., [3, 12, 1]) has been successfully used for segmenting scenes in which people move in the environment and are able to provide not only information about the appearance of a person (e.g. colors) but also 3D information of each pixel belonging to the person.

In practice a stereo vision based people tracker provides, for each frame, a set of data in the form XYZ-RGB containing a 2 1/2D model and color information of the person being tracked. Moreover, correspondences of these data over time is also available therefore when multiple people are in a scene, we have a set of XYZ-RGB data for each person. Obviously, this kind of segmentation can be affected by errors, but the experience we report in this article is that this phase is good enough to allow for implementing an effective posture classification technique as described here. Moreover, the use of stereo-based tracking guarantees a high degree of robustness also to illumination changes, shadows and reflections, thus making the system applicable in a wider range of situations.

The evaluation of the method has been performed on the actual output of a stereo vision based people tracker, thus validating in practice the chosen approach. Results show the feasibility of the approach and its robustness to partial occlusions and different view points.

The article is organized as follows. Section 2 describes some related work, Section 3 presents a brief overview of the system and describes the people tracking module upon which the posture recognition module is based. Section 4 presents a discussion about the choice of the model that has been used for representing human postures. Section 5 describes the training phase, while Section 6 introduces the algorithm used for posture classification and the subsequent sections 7, 8, 9 illustrate in details the steps of the algorithm. Finally, Section 10 includes experimental evaluation of the method. Conclusions and future work will conclude the article.

## 2 Related work

The larger part of the works that deal with human body perception through computer vision can be divided in two major groups: those that try to achieve tracking of the pose (a set of quantitative parameters that defines precisely the configuration of the articulated body) through time and those that aim to recognize the posture (a qualitative assessment that represents a set of predefined configurations) at each frame.

The first category is usually more challenging since it requires a precise estimation of the parameters that define the configuration of the body. Given the inherent complexity of the articulated structure of the human body and the consequent multi-modality of the observation likelihood, one might think that propagating over time the probability distribution on the state, rather than a deterministic representation of the state, could be the right approach. The introduction of the condensation algorithm [14] shows how this approach can lead to desirable results, however revealing at the same time that the computational resources needed for the task are unacceptable. In the following years, there have been many attempts to reduce the problem of the time elapsed, for instance by reducing the number of particles and including a local search [19] or a simulated annealing [8] in the algorithm. Even if the results remain very precise and the time elapsed decrease with these new approaches, the goal of an application that can be used in real scenarios is still far from being achieved due to the still inadmissible time request. Propagating a probability distribution over time yields a robust approach, because it deals effectively with the drift of the tracking error over time. Another approach that addresses the accumulation of the error over time and the ability to recover from error is the one based on the recognition of the components of the articulated body in the single image. These approaches [17, 13, 16] are characterized by the recovery in the images of potential primitives of the body (such as a leg, a head or a torso) through template search exploiting edge and/or appearance information, and then searching for the most likely configuration given the primitives found. While this approach easily allows coping with occlusions, given its bottom-up nature, it still remains limited in the 2D information that it exploits and that it outputs. Other approaches

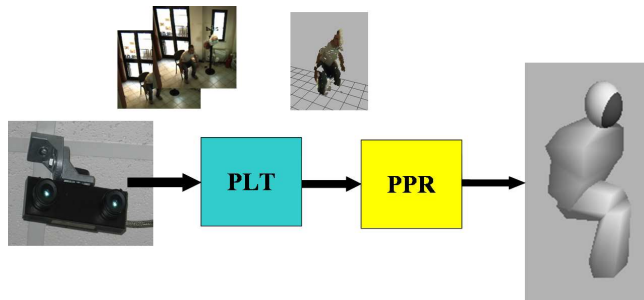


Figure 1: Overview of the system.

try to overcome this limitation, proposing to use a well defined 3D model of the object of interest, and then trying to match these models with the range image, either using the ICP algorithm [7] or a modified version of the gradient search [5]. These approaches are computationally convenient with respect to many others, especially the former that achieves the goal of producing real-time results, even if one can suspect that has problems in dealing with occlusions.

The approaches in the second category, rather than recovering the pose, attempt to classify the posture assumed by the examined person in every single frame, picking up one among a predefined set of postures. Usually this means that some low-level features of the body segment of the image, such as projection histograms [6, 4, 11] or contour based shape descriptors [11], are computed in order to achieve this classification. Otherwise a template is obtained to represent a single class of postures, and then the image is compared with the whole set of templates to find the best match, for example using Chamfer matching [9]. The main difficulty with this kind of solutions is that the sets of different defined postures are not usually disambiguated by the particular low level features that one might choose. Moreover, the set of templates that are used as prototypes for the different classes of postures, do not have enough information to distinguish correctly all the different postures.

Our approach tries to combine aspects of the two categories. We propose a method for posture recognition that does not discard some of the crucial information about the body configuration that we decided to track over time. With respect to methods in the first group, our approach is less time consuming, allowing us to use it in applications such as video surveillance. Indeed, though the output given by our system is not as rich as the one showed in other works [19, 8], we show that there is no need of further analysis of the image when the objective is to classify a few postures. With respect to methods in the second group, our approach is more robust, not relying on low level feature that usually are not distinctive of one single class of postures when the subject is analyzed from different points of view. In fact, we show that the amount of information we used is the right trade-off between robustness and efficiency of the application.

### 3 Overview of the system

The system described in this article is schematically represented in Figure 1. Two basic modules are present in this schema: PLT (People Localization and Tracking), which is responsible for analyzing stereo images and to segment the scene extracting 3D and color information about the person being tracked; PPR (Person Posture Recognition), which is responsible for analyzing this data to recognize and track human postures.

In the rest of this section we briefly describe these modules. Then, since the focus of this article is on the posture recognition module, the detailed description of its design and implementation is delayed to the next sections.

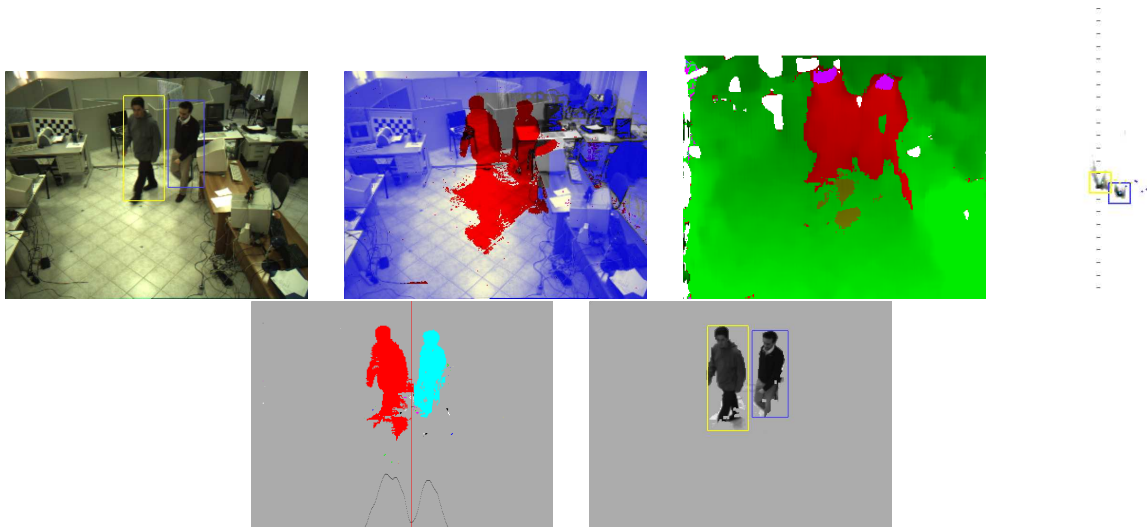


Figure 2: An example of the entire process. From top-left: original image, intensity foreground, disparity foreground, plan-view, foreground segmentation, and person segmentation.

### 3.1 People Localization and Tracking

The stereo vision based People Localization and Tracking (PLT) [1, 12] is composed by three processing modules: 1) segmentation based on background subtraction, that is used to detect foreground people to be tracked; 2) plan-view analysis, that is used to refine foreground segmentation and to compute observations for tracking; 3) tracking, that tracks observations over time maintaining association between tracks and tracked people (or objects).

Background subtraction is performed by considering intensity and disparity components. A pixel is assigned to foreground if there is enough difference between the intensity and disparity of the pixel in the current frame and the related components in the background model. More specifically, With this background subtraction a foreground pixel must have both intensity difference and disparity difference. This allows for correctly dealing with *shadows* and *reflections* that usually produce only intensity difference, but not disparity differences. Observe also that the presence of the disparity model allows for reducing the thresholds, so to detect also minimal differences in intensity, and thus being able to detect foreground objects that have similar colors of the background, without increasing false detection rate due to illumination changes.

Foreground analysis is used to refine the set of foreground points obtained through background subtraction. The set of foreground pixels is processed by: 1) *connected components analysis*, that determines a set of *blobs* on the basis of 8-neighborhood connection; 2) *blob filtering*, that removes small blobs (due for example to noise or high frequency background motion). These processes remove typical noises occurring in background subtraction and allows for computing a set of foreground pixels that contains less errors in representing the set of all pixels belonging to only foreground objects. Therefore it is adequate to be used in the subsequent background update step.

The second part of the processing is plan-view analysis. In this phase, each pixel belonging to a blob extracted in the previous step is projected in the plan-view. This is possible since stereo camera is calibrated and thus we can determine 3-D location of pixels with respect to a reference system in the environment. After projection, we perform a plan-view segmentation. More specifically, for each image blob connected components analysis is used to determine a set of blobs in the plan-view space. This further segmentation allows for determining and solving several cases of *under-segmentation* that occur for example when two people are close in the image space (or partially occluded), but far in the environment. Plan-view blobs are then associated to image blobs and a set of  $n$  pairs (image blob,

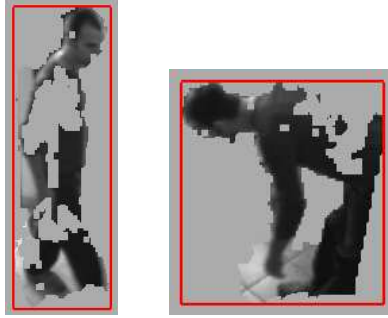


Figure 3: Examples of segmentation provided by the stereo tracker.

plan-view blob) are returned by these process as observations for the  $n$  moving objects (people) in the scene.

Finally, tracking is performed to filter such observations over time. Our tracking method integrates information about person location and color models using a set of Kalman Filters (one for each person being tracked) [12]. Data association between tracks and observations is obtained as a solution of an optimization problem (i.e., minimizing the overall distance of all the observations with respect to the current tracks) based on a distance between tracks and observations. This distance is computed by considering Euclidean distance for locations and model matching procedure for the color models, thus actually integrating the two components in data association.

Tracks in the system are also associated to a finite state automata that control their evolution. Therefore, observations without an associated track generates CANDIDATE tracks and tracks without observations are considered LOST. CANDIDATE tracks are promoted to TRACKED tracks only after a few frames, in this way we are able to discard temporary false detections. While LOST tracks remain in the system for a few frames in order to deal with temporary missing detection of people.

The output of the entire process is thus a set of tracks for each tracked person, where each track contains information about the location of the person over time, as well as RGB-XYZ data (i.e., color and 3D position) for all the pixels that the system has recognized as belonging to the person. Since external calibration of the stereo sensor is available, the reference system for 3D data  $XYZ$  is chosen with the  $XY$  plane corresponding to the ground floor and the  $Z$  axis being the height from the ground. Therefore, for each tracked person, the PLT system provides a set of data  $\Omega^P = \{\omega_t^P, \dots, \omega_{t_0}^P\}$  from the time  $t_0$  in which the person is first detected to current time  $t$ . The value  $\omega_t^P = \{(X_t^i, Y_t^i, Z_t^i, R_t^i, G_t^i, B_t^i) | i \in \mathcal{P}\}$  is the set of XYZ-RGB data for all the pixels  $i$  identified as belonging to the person  $\mathcal{P}$ .

The PLT system produces two kinds of errors in these data: 1) *false positives*, i.e. some of the pixels in  $\mathcal{F}$  do not belong to the person; 2) *false negatives*, i.e. some pixels belonging to the person are not present in  $\mathcal{F}$ . Figure 3 shows two examples of non-perfect segmentation, where only the foreground pixels for which it is possible to compute 3D information are displayed. By analyzing the data produced by the tracking system we estimate that the rate of *false positives* is about 10% and the one of *false negatives* is about 25%.

The posture classification method described in the next sections can reliably tolerate such errors, thus being robust to noise in segmentation that is typical in real world scenarios.

### 3.2 Person Posture Recognition

The Person Posture Recognition (PPR) is responsible for the extraction of the joint parameters that describe the configuration of the body being analyzed. The final goal is to estimate a probability distribution over the set of postures  $\Gamma = \{U, S, B, K, L\}$ , i.e., UP, SIT, BENT, ON KNEE, LAID.

The PPR module makes use of a 3D model of the person. and operates in two phases: 1) a training phase, that allows for adapting some of the parameters of this model to the tracked person; 2) an execution

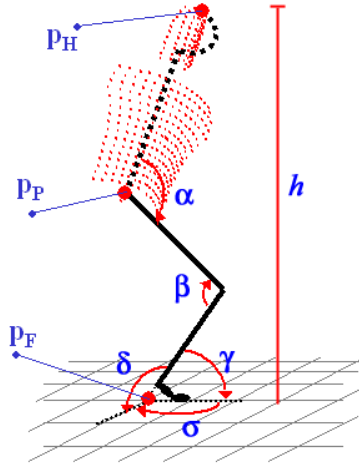


Figure 4: 3D Human model for posture classification.

phase, that is composed by three steps: a) model matching, b) tracking of model principal points, c) posture classification.

The 3D model used by the system, the training phase, and the methods used for model matching, tracking, and classification are described in the next sections.

## 4 A 3D Model for Posture Representation

The choice of a model is critical for the effectiveness of recognition and classification, and it must be carefully taken by considering the quality of data available from the previous processing steps. Therefore, different models have been used in literature, depending on the objectives and on the input data available for the application (see [10] for a review). These models differ mainly for the quantity of information represented.

In our application the input data are not sufficient to cope with hands and arms movement. This is because arms are often missed by the segmentation process, while noises may appear as arms. Without taking into account arms and hands in the model, it is not possible to retrieve information about hand gestures, but is still possible to detect most of the information that allows to distinguish among the principal postures, such as *UP*, *SIT*, *BENT*, *ON KNEE*, *LAID*, etc. Our application is mainly interested in classifying these main postures and then we adopted a model that does not contain explicitly arms and hands.

The 3D model used in our application is shown in Figure 4. It is composed by two sections: a head-torso block and a leg block. The head-torso block is formed by a set of 3D points that represent a 3D surface as shown in Figure 4. In our current implementation, this set contains 700 points that have been obtained by a 180 degree rotation of a curve. Since we are not interested in knowing head movements, we model the head together with the torso in a unique block (without considering degrees of freedom for the head). However, the presence of the head in the model is justified by two considerations: 1) in a camera set-up in which the camera is placed high in the environment heads of people are very unlikely to be occluded; 2) heads are easy to be detected, since 3D and color information are available, and modeled for tracking, since it is reasonable to assume that head appearance can be modeled with a bimodal color distribution, usually corresponding to skin and hair color.

The pelvis joint is simplified to be a planar rotoidal joint, instead of a spherical one. This simplification is justified if one thinks that, most of the times, the pelvis is used to bend frontally. Also, false

positives and false negatives in the segmented image and the distortion due to the stereo system, make the attempt of detecting vertical torsion and lateral bending extremely difficult.

The legs are unified in one articulated body. Assuming that the legs are always in contact with the floor, a spherical joint is adopted to model this point. For the knee is instead used a single planar rotoidal joint.

The model will be built by assuming a constant ratio between the dimensions in the model parts and the height of a person, which is instead evaluated by 3D data of the person tracked.

On this model we define three *principal points*: the head ( $\mathbf{p}_H$ ), the pelvis ( $\mathbf{p}_P$ ) and the legs point of contact with floor ( $\mathbf{p}_F$ ) (see Figure 4). These points are tracked over time, as shown in the next sections, and used to determine measures for classification. In particular, we define an observation vector  $z = [\alpha, \beta, \gamma, \delta, h]$ , that contains the estimation of the four angles  $\alpha, \beta, \gamma, \delta$ , and the normalized height  $h$ , which is the ratio between the height measured at the current frame and the height of the person measured during the training phase (see Figure 4). Notice that  $\sigma$  is not included in the observation vector since it is not useful to determine human postures.

## 5 Training Phase

Since the human model used in PPR contains data that must be adapted to the person being analyzed, a training phase is executed for the first frames in the sequence (ten frames are normally sufficient), to measure the person's height and to estimate the head bimodal color distribution.

We assume that in this phase the person is exhibiting an erect posture with arms below the shoulder level, and with no occlusions for his/her head.

The height of the person is measured using 3D data provided by the stereo vision based tracker: for each frame, we consider as the height of the person, the maximum value of the  $Z_t^i$  in  $\omega_t$ ; the height of the person is then determined by averaging such maximal values over all the training sequence.

Considering that a progressively correct estimation of the height (and, as a consequence, of the other body dimensions) is also available during the training phase, the points in the image whose height is within 25 cm. to the top of the head (we assumed that the arms are below the shoulder level) can be considered as head points. Since the input data provide also color of each point in the image, we can estimate a bimodal color distribution by applying the  $k$ -mean algorithm on head color points, with  $k = 2$ . This results in two clusters of colors  $C_1$  and  $C_2$  that are described by the means of their centers of mass  $\mu_{C_1}$  and  $\mu_{C_2}$  and their respective standard deviations  $\sigma_{C_1}$  and  $\sigma_{C_2}$ .

Given the height and the head appearance of a subject, his or her model can be reconstructed, and the main procedure that will be described in the next sections can be executed for the rest of the video sequence.

## 6 Posture Classification Algorithm

As already mentioned, the PPR module classifies person postures using a three-steps approach: model matching, tracking and classification. The algorithm implementing a processing step of PPR is shown in Table 1.

A couple of data structures are used to simplify the readability of the algorithm.  $\Pi$  contains the three principal points of the model ( $\mathbf{p}_H, \mathbf{p}_P, \mathbf{p}_F$ );  $\Theta$  contains  $\Pi$ ,  $\sigma$ , that is the normal vector of the symmetry plane of the person, and  $\phi$ , that defines the probability of the left part of the body to be on the positive side of the symmetry plane (that is, where  $\sigma$  grows positive).

The input to the algorithm is represented by the structure  $\Theta$  estimated at the previous frame of the video sequence, the probability distribution of the postures in the previous step  $P_\gamma$ , and the current 3D point set  $\omega$  coming from the PLT module. Thus the output will be the new structure  $\Theta'$  together with the new probability distribution  $P'_\gamma$  over the postures.

**Structures** $\Theta = [\Pi, \sigma, \phi]$  $\Pi = [\mathbf{p}_F, \mathbf{p}_P, \mathbf{p}_H]$ **Algorithm**INPUT:  $\Theta, \omega, P_\gamma$ OUTPUT:  $\Theta', P'_\gamma$ CONST:  $\eta, \lambda, CHANGE\_TH$ #  $\eta$ : model  $\lambda$ : learned height

# (these values are computed by the Training phase)

PROCEDURE:

 $H = \max\{Z | Z \in \omega\};$ IF  $((\lambda - H) < CHANGE\_TH)\{$  $\Theta' = \Theta;$  $z = [0, 0, 0, 0, 1];$  $\}$ 

ELSE{

 $[\tilde{\mathbf{p}}_P, \tilde{\mathbf{p}}_H] = \text{ICP}(\eta, \omega);$ 

#

IF  $(!\text{leg\_occluded}(\omega, \mathbf{p}_F))$ 

#

 $\tilde{\mathbf{p}}_F = \text{find\_leg}(\omega, \mathbf{p}_F)$ 

# Detection (Sec. 7)

ELSE

#

 $\tilde{\mathbf{p}}_F = \text{project\_on\_floor}(\tilde{\mathbf{p}}_P);$ 

#

 $\Pi' = \text{kalman\_points}(\Pi, \tilde{\Pi});$ 

#

 $\sigma' = \text{filter\_plane}(\sigma, \Pi');$ 

#

 $\hat{\Pi}' = \text{project\_on\_plane}(\Pi', \sigma');$ 

# Tracking (Sec. 8)

 $\rho = \text{evaluate\_left\_posture}(\hat{\Pi}', \sigma');$ 

#

 $\phi' = \text{filter\_left\_posture}(\rho, \phi);$ 

#

 $z = [\text{get\_angles}(\hat{\Pi}', \sigma', \phi'), H/\lambda];$ 

#

 $\}$  $P'_\gamma = \text{HMM}(z, P_\gamma);$ 

# Classification (Sec. 9)

Table 1: The algorithm for model matching and tracking of the principal points of the model. See text for further details.

A few symbols need to be described in order to easily read the algorithm:  $\eta$  is the model (both the shape and the color appearance),  $\lambda$  is the person’s height learned during the training phase,  $z$  is the observation vector used for classification, as defined in Section 4.

The procedure starts by detecting if a significant difference in the person’s height (with respect to the learned value  $\lambda$ ) occurred at this frame. If such a difference is below a threshold `CHANGE_TH`, that usually is set to a few (e.g., 10) centimeters, then  $z$  is set to specify that the person is standing up without further processing.

Otherwise the algorithm first extracts the position of the three *principal points* of the model. More specifically,  $\mathbf{p}_H$  and  $\mathbf{p}_P$  (head and pelvis points) are estimated by using an ICP variant and other ad-hoc methods, that will be described in Section 7. While  $\mathbf{p}_F$  (feet point) is computed in two different ways depending on the presence of occlusions. The presence of occlusions of the legs is checked with the `leg_occluded` function. This function simply verifies if only a small number of points in  $\omega_t$  are below half of the height of the person (the threshold is determined by experiments and it is about 20% of the total numbers of points in  $\omega$ ). If the legs are occluded,  $\mathbf{p}_F$  is estimated as the projection of  $\mathbf{p}_P$  on the ground, otherwise it is computed as the average of the lowest points in the data  $\omega_t$ .

The second step of the algorithm consists in tracking the principal points over time. This tracking is motivated by the fact that postures (and thus principal points of the model) change smoothly over time, and it allows for increased robustness to the segmentation noise. As a result of the tracking step, the observation vector  $z$  (as defined in Section 4) is computed using simple trigonometry operations (`get_angles`). The tracking step is described in detail in Section 8.

Finally, an HMM classification is used to better estimate the posture for the each frame of the video sequence (Section 9), taking into account the probability of transitions between different postures.

## 7 Detection of the Principal Points

The principal points  $\mathbf{p}_H$  and  $\mathbf{p}_P$  are estimated using a variant of the ICP algorithm (for a review of the variants of the ICP see [18]). Given two point sets to be aligned, the ICP proposes an iterative approach to the absolute orientation problem when the correspondences between the two point sets are not known. In our case the two point sets are  $\omega$ , the data, and  $\eta$ , the model. At each iteration, each point in the model is matched against the closest point in the data set.

The structure of the model  $\eta$  is shown in Figure 4. Since it represents a view of the torso-head block, it can be used only to find the position of the points  $\mathbf{p}_H$  and  $\mathbf{p}_P$ , but it cannot say us anything about the torso direction.

The use of ICP is to estimate a rigid transformation to be applied to  $\eta$  in such a way to reduce to a minimum the misalignment between  $\eta$  and  $\omega$ . The ICP is proved [2] to optimize the function

$$E(\mathbf{R}, \mathbf{t}) = \sum_{i=1}^N \|d_i - \mathbf{R}m_i - \mathbf{t}\|^2 \quad (1)$$

Where  $\mathbf{R}$  is the rotation matrix and  $\mathbf{t}$  is the translation vector that together specify a rigid transformation,  $d_i$  is a point of  $\omega$ ,  $m_i$  is a point of  $\eta$ . We are assuming that the points are assigned the same index if they are correspondent. As already mentioned, such correspondence is calculated according to the minimum Euclidean distance between points in the model and points in the data set. Formally, given a point  $m_j$  in  $\eta$ ,  $d_k$  in  $\omega$  is labeled as correspondent of  $m_j$  if

$$d_k = \operatorname{argmin}_{d_u \in \omega} \operatorname{dist}(m_j, d_u) \quad (2)$$

where the function `dist` is defined according to the Euclidean metric.

The ICP algorithm is applied by setting as initial configuration the pose of the model computed in the previous frame, and for the first frame a straight model corresponding to a standing person. Since

postures do not change instantaneously, this initial configuration allows for quick convergence of the process. Moreover, we limit the number of steps to a predefined number (18 in our current implementation), that guarantees near real-time performance.

Since from the training phase we have also computed the head color distribution, described by the centers of mass of the color clusters  $C_1$  and  $C_2$  and the respective standard deviations  $\sigma_{C_1}$  and  $\sigma_{C_2}$ , the ICP has been modified to take into account these additional data. Indeed, in our implementation, the search for the correspondences of points in the head part of the model is restricted to a subset of the data set  $\omega$  defined as follows

$$\{d_k \in \omega | \text{dist}(\text{color}(d_k), \mu_{C_1}) < t(\sigma_{C_1}) \text{ OR } \text{dist}(\text{color}(d_k), \mu_{C_2}) < t(\sigma_{C_2})\} \quad (3)$$

where  $\text{color}(d_k)$  is the value of the color associated with point  $d_k$  in the RGB color space and  $t(\sigma)$  is a threshold related to the amplitude of standard deviation for each cluster.

Also, since the head correspondences exploit a greater amount of information, we have double their weight. This can be easily done by counting twice each correspondence in the head data set, thus increasing its contribution in determining the rigid transformation in the ICP minimization error phase. Once the best rigid transformation  $(\mathbf{R}, \mathbf{t})$  has been extracted with the ICP, it can be applied to  $\eta$  in order to match  $\omega$ . Since we know the relative position of  $\mathbf{p}_P$  and  $\mathbf{p}_H$  in the model  $\eta$ , their position on  $\omega$  is also known.

For  $\mathbf{p}_F$  we cannot use the same technique, primarily because the lower part of the body is not always visible due to occlusions or to the greater sensibility to false negatives. Since we are interested in finding a point that represent the legs point of contact with the floor, we can simply project the lower points on the ground level, when at least part of the legs is visible. When the person legs are utterly occluded, for example if he/she is sitting behind a desk, we can anyway model the observation as a Gaussian distribution which center is the projection on the ground of  $\mathbf{p}_P$  and which variance is in inverse relation with the height of the pelvis from the floor (function `project_on_floor` in the algorithm).

## 8 Tracking of Principal Points

Even though the principal points are available for each image, there are still problems that need to be solved in order to have good performance in classification.

First, detection of these points is noisy given the noisy data coming from the tracker. To deal with these errors it is necessary to filter data over time and to this end, we use three independent Kalman Filters (function `kalman_points` in the algorithm) to track them. These Kalman Filters represent position and velocity of the points assuming a constant velocity model in the 3D space. Second, ambiguities may arise in determining poses from three points. To solve this problem we need to determine the *symmetry plane* of the person (that reduces ambiguities to up to two cases, considering the constraint on the knee joint), and a likelihood function that evaluates probability of different poses. The symmetry plane can be represented by a vector  $\sigma'$  originating at the point  $\mathbf{p}_F$ . To estimate the plane of symmetry one might estimate the plane passing through the three principal points. However this plane can differ from the symmetry plane due to perception and detection errors. In order to have more accurate data, we need to consider the configuration of the three points, for example co-linearity of these points increases noise in detecting the symmetry plane. In our implementation we used another Kalman Filter (function `filter_plane`) on the orientation of the symmetry plane that suitably takes into account co-linearity of these points. This filter provides for smooth changes of orientation of the symmetry plane. Furthermore, principal points estimated before are projected onto the filtered symmetry plane (function `project_on_plane`), and these projections are actually used in the next steps.

Given the symmetry plane, we still have two different solutions, corresponding to the two opposite orientations of the person. To determine which one is correct we use the function `evaluate_left_posture` that computes the likelihood of the orientation of the person. An example is given in Figure 5, where the

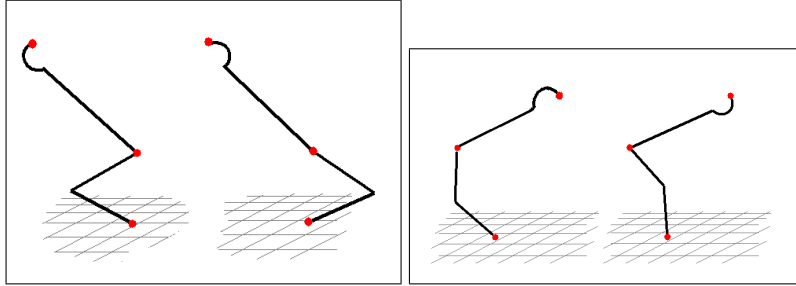


Figure 5: Ambiguities.

two orientations in two situations are shown. We fix a reference system for the points in the symmetry plane and the orientation likelihood function measures the likelihood that the person is oriented on the left. For example, the likelihood for the situation in Figure 5 a) is 0.6 (thus slightly preferring the leftmost posture), while the one in Figure 5 b) is 0, since the leftmost pose is very unnatural. The likelihood function can be instantiated with respect to the environment in which the application runs. For example, in an office-like environment likelihood of situation in Figure 5 a) may be increased (thus preferring more the leftmost posture).

Finally, by filtering these values uniformly through time (function `filter_left_posture`), we get a reliable estimate of the frontal orientation  $\phi$  of the person. Considering that we already know the symmetry plane, at this point we can build a Person Reference System.

This step completes the tracking process and allows to compute a set of parameters that will be used for classification. These parameters are four of the five angles of the joints defined for the model ( $\sigma$  does not contribute to posture detection) and the normalized height (see also Figure 4). Specifically, the function `get_angles` computes the angles of the model for the observation vector  $z_t = \langle \alpha, \beta, \gamma, \delta, h \rangle$ , while the normalized height  $h$  is determined by the ratio between the current height and the height learned in the training phase  $\lambda$ . The vector  $z_t$  is then used as input by the classification step. As shown in the next sections, this choice represents a very simple and effective coding that can be used to make posture classification.

## 9 Posture Classification

Our approach to posture classification is mainly characterized by the fact that it is not made upon low-level data, but on higher level ones that are retrieved from each image as result of the model matching and tracking processes described in the previous sections. This approach grants better results in terms of robustness and effectiveness.

We have implemented two classification procedures (that are compared in Section 10): one is based on frame by frame maximum likelihood, the other on temporal integration using Hidden Markov Models (HMM). As shown by experimental results, temporal integration increases robustness to the classifier, since it allows for modeling also transition between postures.

They use an observation vector  $z_t = \langle \alpha, \beta, \gamma, \delta, h \rangle$ , which contains the five parameters of the model, and the distribution probabilities  $P(z_t|\gamma)$  for each posture that needs to be classified  $\gamma \in \Gamma = \{U, S, B, K, L\}$ , i.e., UP, SIT, BENT, ON KNEE, LAID. These distributions are acquired by analyzing sample videos or synthetic model variations. In our case, since values  $z_t$  are computed after model matching, we used synthetic model variations and manually classified a set of postures of the model to determine  $P(z_t|\gamma)$  for each  $\gamma \in \Gamma$ . More specifically, we have generated a set of nominal poses of the model for the postures in  $\Gamma$ . Then we collected, for each posture, a set of random poses generated as small variations of the nominal ones, and manually labelled the ones that can still be considered in the same posture class. This produces a distribution over the parameters of the model for each posture.

In addition, due to the uni-modal nature of such distributions, they have been approximated as normal distributions.

The main characteristic of our approach is that the measured components are directly connected to human postures, thus making easier the classification phase. In particular, the probability distributions of each pose in the space formed by the five parameters extracted as described in the previous section are uni-modal. Moreover, the distributions for the different postures are well separated each other thus making this space very effective for classification.

The first classification procedure just considers the maximum likelihood of the current observation, i.e.

$$\gamma_{ML} = \underset{\gamma \in \Gamma}{\operatorname{argmax}} P(z_t | \gamma)$$

The second classification procedure makes use of a Hidden Markov Model (HMM) defined by a discrete status variable assuming values in  $\Gamma$ . Probability distribution for the postures is thus given by

$$\begin{aligned} P(\gamma_t | z_{t:t_0}) &= \eta P(z_t | \gamma_t) \sum_{\gamma' \in \Gamma} P(\gamma_t | \gamma') P(\gamma' | z_{t-1:t_0}) \\ P(\gamma | z_{t_0}) &= \eta P(z_{t_0} | \gamma) P(\gamma) \end{aligned}$$

where  $z_{t:t_0}$  is the set of observations from time  $t_0$  to time  $t$ , and  $\eta$  is a normalizing factor.

The transition probabilities  $P(\gamma_t | \gamma')$  are used to model transitions between the postures, while  $P(\gamma)$  is the a priori probability of each posture. A discussion about the choice of these distributions is reported in Section 10.

## 10 Experimental Evaluation

In this section we report experimental results on the presented method. Experimental evaluation has been performed by using a *standard setting* in which the stereo camera was placed indoor about 3 meters high from the ground pointing down about 30 degrees from the horizon, the people in the scene were between 3 and 5 meters from the camera, in a frontal view with respect to the camera, and without occlusions. This setting has been modified in order to explore the behavior of the system in different conditions. In particular, we have considered four other settings varying orientation of the person, presence of occlusions, different heights of the camera and outdoor scenarios.

The stereo vision based people tracker in [1] has been used to provide XYZ-RGB data of tracked person in the scene. The tracker processes 640x480 images at about 10 frame per seconds, thus giving us high resolution and high rate data. The system described in this article has an average computation cycle of about 180 ms on a 1.7 GHz CPU. This value is computed as the average process time for a cycle. However, it is necessary to observe that cycle processing time depends on the situation; when the person is recognized in a standing pose, then no processing on detection and tracking is performed allowing for a quick response. The ICP algorithm takes most of the computational time at each step, but this process is fast, since a good initial configuration is usually available and thus convergence is usually obtained in a few iterations.

The overall system (PLT + PPR) can process about 3.5 frames per second. Moreover, code optimization and more powerful CPUs will allow to use the system in real-time. The overall testing set counts 26 video sequences of about 150 frames each. Seven different people acted for the tests (subject S.P. with 15 tests, subject L.I. with 7 tests, subjects M.Z, G.L., V.A.Z and D.C. with 1 test each). Different lighting conditions have been encountered during the experiments that have been done in different locations and in different days, under both natural and artificial lighting with various intensities. The set of data used in the experiments is shown in <http://www.dis.uniroma1.it/~iocchi/PLT/posture.html> and they are available for comparison with other approaches.

<b>System Ground Truth</b>	<b>UP</b>	<b>SIT</b>	<b>BENT</b>	<b>KNEE</b>	<b>LAID</b>
<b>UP</b>	<b>93.2 %</b>	0.0 %	6.0 %	0.0 %	0.0 %
<b>SIT</b>	0.0 %	<b>86.6 %</b>	13.4 %	0.0 %	0.0 %
<b>BENT</b>	2.0 %	0.5 %	<b>97.5 %</b>	0.0 %	0.0 %
<b>KNEE</b>	0.0 %	22.2 %	0.0 %	<b>77.8 %</b>	0.0 %
<b>LAID</b>	0.0 %	0.0 %	0.0 %	0.0 %	<b>100.0 %</b>

Table 2: Overall confusion matrix with HMM.

The evaluation of the system has been obtained against a ground truth. For each video we built a ground truth by manually labelling frames with the postures assumed by the person. Moreover, since during transitions from one posture to another it is difficult to provide a ground truth (and are also typically not interesting in the applications), we have defined transition intervals, during which there is a passage from one posture to another. During these intervals the system is not evaluated.

This section is organized as follows. First, we will show the experimental results of the system in the standard setting, then we will explore the robustness of the system with respect to different view points, occlusions, change in the height of the camera, and an outdoor scenario. In presenting these experiments we want also to evaluate the effectiveness of the filter provided by HMM, with respect to frame by frame classification.

## 10.1 Standard Setting

The experiments have been performed by considering a set of video sequences, chosen in order to cover all the postures we are interested in. The *standard setting* described above has been used for this first set of experiments and then the results in this setting are compared with different settings.

Both for the values in the state transition matrix and the a priori probability of the HMM, we have considered that the optimal tuning is environment dependant. Indeed, an office-like environment will very likely have different posture transition probability than those of a gym: in the first case, for example, it might be possible to have high values in the transition between the *sitting* and itself; in a gym the whole matrix should have similar values in all its entries, taking in this way into account that the posture often changes. The optimal values should be achieved by training on video sequences regarding the environment of interest. For simplicity purposes, in our application we have determined values that could be typical of an office-like environment. In particular, we have chosen an a priori probability of 0.8 for the *standing* position and  $0.2/(|\Gamma| - 1)$  for the others. This models situations in which a person enters the scene in an initial standing position and that the transition to all the other postures have the same probability. Moreover, we assume that from any posture (other than standing) it is more likely to stand (we fixed this value to 0.15) than to go to another posture. Therefore, the transition probabilities  $T_{ij} = P(\gamma_t = i | \gamma_{t-1} = j)$  have been set to

$$\begin{pmatrix} 0.800 & 0.050 & 0.050 & 0.050 & 0.050 \\ 0.150 & 0.800 & 0.016 & 0.016 & 0.016 \\ 0.150 & 0.016 & 0.800 & 0.016 & 0.016 \\ 0.150 & 0.016 & 0.016 & 0.800 & 0.016 \\ 0.150 & 0.016 & 0.016 & 0.016 & 0.800 \end{pmatrix}$$

	HMM	Maximum Likelihood
<b>UP</b>	93.2 %	90.7 %
<b>SIT</b>	86.6 %	80.0 %
<b>BENT</b>	97.5 %	91.6 %
<b>KNEE</b>	77.8 %	77.8 %
<b>L Aid</b>	100.0 %	100.0 %

Table 3: Classification rates of HMM vs. Maximum Likelihood.

Table 2 presents the total confusion matrix of the experiments performed with this setting. The presence of no errors in the LAID posture is given by the fact that the height of the person from the ground is the most discriminant measure and this is reliably computed by stereo vision, while the ON KNEE posture is very difficult because it relies on tracking the feet, which is very noisy and unreliable with the stereo tracker we have used.

The values of classification obtained by using frame by frame classification are slightly lower (see Table 3). Thus, the HMM slightly improves the performance, however maximum likelihood is still effective, since postures are well separated in the classification space defined by the parameters of the model. This confirms the effectiveness in the choice of the classification space and the ability of the system to correctly track the parameters of the human model.

## 10.2 Different view points

Robustness to different points of view has been tested by analyzing postures with people in different orientations with respect to the camera. Here we present the results of tracking bending postures in five different orientations with respect to the camera. For each of the five orientations we took three videos of about 200 frames in which the person entered the scene, bent to grab an object on the ground and then raised up exiting the scene. Table 4 shows classification rates for each orientation. The first column presents results obtained with HMM, while the second one shows results obtained with maximum likelihood. There are very small differences between the five rows, thus showing that the approach is able to correctly deal with different orientations. Also, as already pointed out, improvement in performance due to HMM is not very high.

## 10.3 Partial occlusions

To prove robustness of the system to partial occlusions, we make experiments comparing situations without occlusions and situations with partial occlusions. Here we consider occlusions of the lower part of the body, while we assume the head and the upper part of the torso are visible. This is a reasonable assumption since the camera is placed in a higher position than people. In Figure 6 we show a few frames of two data sets used for evaluating the recognition of the *sitting* posture without and with occlusions and in Table 5 classification rates for the different postures.

It is interesting to notice that we have very similar results in the two columns (in some cases higher classification rate under partial occlusions). The main reason is that, when feet are not visible, they are projected on the ground from the pelvis joint  $p_P$  and this corresponds to determine correct angles for the postures UP and BENT. Moreover, LAID posture is mainly determined from the height parameter that is also not affected by partial occlusions. For the posture ON KNEE we have not performed these experiments for two reasons: i) it is difficult to be recognized even without occlusions; ii) it is not correctly identified in presence of occlusions since this posture assumes the feet to be not below the pelvis. These results thus show an overall good behavior of the system in recognizing postures in presence of partial occlusions, that are typical for example during office-like activities.

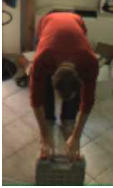




Orientation	HMM	Maximum Likelihood
	91.6 %	86.7 %
	86.0 %	83.1 %
	91.2	89.7 % %
	89.7 %	89.7 %
	90.5 %	88.9 %

Table 4: Classification rates from different view points.

	No occlusions	Partial occlusion
<b>UP</b>	93.2 %	91.5 %
<b>SIT</b>	86.6 %	81.6 %
<b>BENT</b>	97.5 %	93.3 %
<b>KNEE</b>	77.8 %	N/A
<b>LAI</b>	100.0 %	100.0 %

Table 5: Classification rates without and with occlusions.

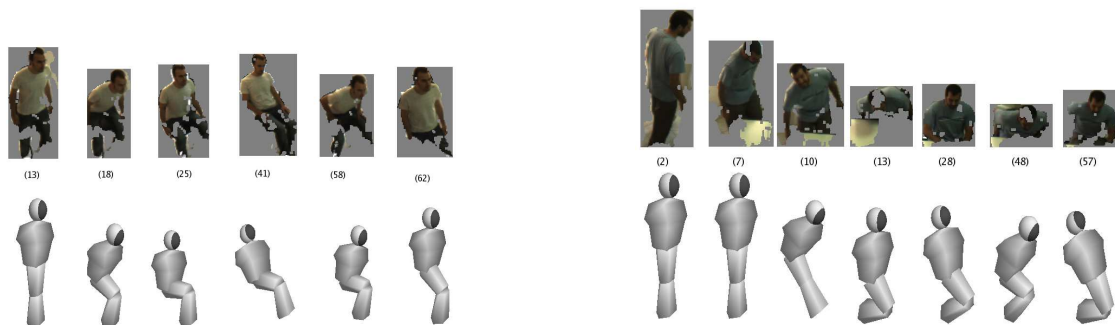


Figure 6: People sitting on a chair (non-occluded vs. occluded).

	<b>3m camera height</b>	<b>1.5 m camera height</b>
<b>UP</b>	93.2 %	96.3 %
<b>SIT</b>	86.6 %	77.0 %
<b>BENT</b>	97.5 %	99.0 %
<b>KNEE</b>	77.8 %	76.9 %
<b>LAID</b>	100.0 %	90.0 %

Table 6: Classification rates with 3 m and 1.5 m camera heights.

	<b>Indoor</b>	<b>Outdoor</b>
<b>UP</b>	96.3 %	95.8 %
<b>SIT</b>	77.0 %	77.1 %
<b>BENT</b>	99.9 %	99.0 %
<b>KNEE</b>	76.9 %	77.0 %
<b>LAID</b>	90.0 %	90.0 %

Table 7: Classification rates in indoor and outdoor environments (1.5 m camera height).

#### 10.4 Camera at different heights

In the previous setting, the camera was placed 3 meters high from the ground. However, we tested the behavior of the system also with different camera placements. In particular, we have put the camera at about 1.5 meters from the ground. In this setting the PLT was able to reliably segment and track the people movement. The classification rates for each posture are summarized in Table 6. From the results, it is clear that there are not significant differences, except for the `SIT` posture that has a relatively lower score. This can be explained by the higher amount of self-occlusions occurring when a person sits in front of a lower camera, that makes model matching more difficult.

Given this problem with the `SIT` posture, we have also performed specific tests with the low camera combined with occlusions. The classification accuracy in this setting was 47.2 %, thus denoting that performance are highly affected by partial occlusions when the camera is low.

#### 10.5 Outdoor setting

Finally we have tested the system on an outdoor scenario. Since it was not possible to put a camera at a height of 3 m. in the outdoor scenario we used tgh 1.5 m camera height configuration. Even though the particular outdoor scenario was not very dynamic, since it is located in a private area, we were nevertheless able to test the robustness of the system against natural lights. The classification rates for this setting are summarized in Table 7 The results do not show a significant degradation of the performance with respect to the low camera height setting, showing the ability of the system to operate appropriately even in an outdoor scenario. However, this experiment highlights a higher difficulty in outdoor scenes, where usually it is not possible to place the camera in the best position for the system.

#### 10.6 Error Analysis

From the analysis of the experimental results reported above, we have highlighted situations in which errors occur. A first class of errors is due to bad segmentation: 1) when this occurs during the initial training phase, a non-correct initialization of the model affects model matching in the following frames, thus producing errors in the computation of the parameters that are used for classification; 2) segmentation errors in the upper part of the body (head and torso) may also be the cause of failures in the model

matching performed by the ICP algorithm. These errors are generated by the underlying tracking system and in case they are not acceptable for an application, it is necessary to tune the tracker and/or to add additional processing in order to provide for better segmentation.

Errors that are more related to our approach are mostly determined by incorrect matching of the ICP algorithm, specially in situations where movements are too quick. This is a general problem for many systems based on tracking. A minor problem arises when the person do not pass through non-ambiguous postures. In fact, until disambiguation is not achieved (as described in Section 8), posture recognition may be wrong.

Finally, the PPR system is quite robust to different view points, partial occlusions and to indoor/outdoor environments. The performance are slightly worse when the camera is placed low in the environment. In particular, low camera setting shows a higher sensitivity to occlusions.

## 11 Conclusions

In this article we have presented a method for human posture tracking and classification that relies on the segmentation of a stereo vision based people tracker. The input to our system is a set of XYZ-RGB data extracted by the tracker and the system is able to classify several main postures with high efficiency, good accuracy and high degree of robustness to various situations. The approach is based on the computation of significant parameters for posture classification, that is performed by using an ICP algorithm for 3D model matching; 3D tracking of these points over time is then performed by using a Kalman Filter in order to increase robustness to perception noise; and finally a Hidden Markov Model is used to classify postures over time.

The experimental results reported here show the feasibility of the approach and its robustness to different points of view, occlusions and different environment conditions, that makes the system applicable to a larger number of situations.

One of the problems experienced was that the people tracker module works very well when people are in standing position, while quality of data worsen when people sit, lay down, or bend. Classification errors may be reduced by providing feedback from the posture classification module to the people tracker one. In fact, given these information the tracker could adapt recognition procedure in order to provide better data.

The work described in this article can be extended to consider other activities (e.g., gestures), when an appropriate segmentation process is executed before it, providing good quality 3D information of the subject. While other activities, like running or jumping, can be recognized by analyzing directly data coming from the people tracking system, since for these cases the 3D model used in this article would be less relevant.

## References

- [1] S. Bahadori, L. Iocchi, G. R. Leone, D. Nardi, and L. Scozzafava. Real-time people localization and tracking through fixed stereo vision. *Applied Intelligence*, 26:83–97, 2007.
- [2] P. J. Besl and N. MacKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1992.
- [3] D. Beymer and K. Konolige. Real-time tracking of multiple people using stereo. In *Proc. of IEEE Frame Rate Workshop*, 1999.
- [4] B. Boulay, F. Bremond, and M. Thonnat. Posture recognition with a 3d human model. In *International Conference on Crime Detection and Prevention (ICDP)*, 2005.
- [5] Matthieu Bray, Esther Koller-Meier, Nicol N. Schraudolph, and Luc Van Gool. Fast stochastic optimization for articulated structure tracking. *Image and Vision Computing*, 25, 2007.
- [6] R. Cucchiara, C. Grana, and A. Prati. Probabilistic posture classification for human-behavior analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, 35(1):42–54, 2005.

- [7] D. Demirdjian, T. Ko, and T. Darrel. Constraining human body tracking. In *International Conference on Computer Vision (ICCV'03)*, 2003.
- [8] J. Deutscher, A. Blake, and I. Reid. Articulated motion capture by annealing particle filtering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 126–133, 2000.
- [9] M. Dimitrijevic, V. Lepetit, and P. Fua. Human body pose recognition using spatio-temporal templates. In *ICCV workshop on Modeling People and Human Interaction*, 2005.
- [10] D. M. Gavrilu. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.
- [11] L. Goldmann, M. Karaman, and T. Sikora. Human body posture recognition using mpeg-7 descriptors. *Visual Communications and Image Processing*, 2004.
- [12] L. Iocchi and R. C. Bolles. Integrating plan-view tracking and color-based person models for multiple people tracking. In *Proc. of IEEE International Conference on Image Processing (ICIP'05)*, volume III, pages 872–875, Genova, Italy, 2005. ISBN: 0-7803-9135-7.
- [13] S. Ioffe and D. Forsyth. Human tracking with mixtures of trees. *Proceedings of International Conference on Computer Vision (ICCV)*, 2001.
- [14] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.
- [15] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding: CVIU*, 81(3):231–268, 2001.
- [16] R. Navaratnam, A. Thayananthan, P.H.S. Torr, and R. Cipolla. Hierarchical part-based human body pose. In *British Machine Vision Conference*, 2005.
- [17] D. Ramanan, D. A. Forsyth, and Andrew Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 271–278, Washington, DC, USA, 2005. IEEE Computer Society.
- [18] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. *Proc. of 3rd International Conference on 3D Digital Imaging and Modeling*, 2001.
- [19] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. *Proc. of the Conference on Computer Vision and Pattern Recognition*, 2003.