HUMAN POSTURE TRACKING AND CLASSIFICATION THROUGH STEREO VISION

Stefano Pellegrini and Luca Iocchi
Dipartimento di Informatica e Sistemistica
Univesità di Roma "La Sapienza"
Via Salaria 113, Roma 00198, Italy

Keywords: Human Posture Classification, Human Activity Recognition, 3D Modeling, Stereo Vision.

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 important in several fields, like ambient intelligence, surveillance, elderly care, etc. 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 paper we present a system for posture tracking and classification that uses a stereo vision sensor, which provides both for a robust way to segment and track people in the scene and 3D information about tracked people. The presented method uses a 3D model of human body, performs model matching through a variant of the ICP algorithm and then uses a Hidden Markov Model to model posture transitions. Experimental results show the effectiveness of the system in determining human postures in presence of partial occlusions and from different view points.

1 INTRODUCTION

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 (Gavrila, 1999; Moeslund and Granum, 2001)) for different kinds of application, including surveillance, ambient intelligence, elderly care. 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 paper 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.

A major distinction among works in human posture recognition is given by the presence of a model of the human body. Methods that do not use human models are based on low-level features extracted from the images. For example, (Cucchiara et al., 2005a) use Projection Histograms and (Goldmann et al., 2004) use also Contour-Based Shape Descriptors to classify some a priori defined postures. The main drawback of these methods is that they rely on correct segmentation of person silhouette and thus are quite sensitive to noise, occlusions and to the view point from which the person is seen, since low-level features tend to ignore some relevant information in the images, e.g., the position of recognizable body parts such as the head or the hands. The work in (Cucchiara et al., 2005b) uses a multi-camera setting to overcome problems related to partial occlusions in posture classification, providing a solution to determine standing postures in presence of occlusions.

Methods based on a human model can be distinguished in two categories: the first includes works using a 2D model analyzing 2D data (monocular camera), the second using a 3D model analyzing either 2D or 3D data (stereo vision or multi-camera settings). Works in the first group are often characterized by the fact that a predefined point of view or some constraints on the movements of the person being analyzed must be specified for the procedure to grant effective results.

The majority of the works based on a human model use a 3D model, both in the case of 2D and 3D input data. In (Sminchisescu and Triggs, 2003) the parameters of a complete 3D model are estimated from monocular images using particle filtering, but the time used for the analysis of each image is prohibitive to real-time applications. In (Boulay et al., 2005) pos-

Pellegrini S. and Iocchi L. (2006).
HUMAN POSTURE TRACKING AND CLASSIFICATION THROUGH STEREO VISION.
In *Proceedings of the First International Conference on Computer Vision Theory and Applications*, pages 261-269
DOI: 10.5220/0001376702610269

tures are searched using a 3D complete model. After image segmentation, the silhouette of the detected image is compared to the virtual silhouette generated by the model in some predefined postures. The posture of the model which has the best match, according to a projection histogram procedure, is chosen to be the right one. (Bregler and Malik, 1998) use a complete 3D model approach that can be used both with a single camera setting, but constrain the human motion to be along a single direction, and with multiple cameras. In any case, the initial position of the joints must be specified by the user. The approaches using 3D models from monocular cameras can better deal with different view points, but have problems with occlusions due to unpredictable variations of the person figures.

The above works deal with posture tracking in different ways. Some of them (Demirdjian et al., 2003; Bregler and Malik, 1998; Sminchisescu and Triggs, 2003) use different tracking techniques for computing and updating the parameters of the model. These works are indeed not focused on posture classification, i.e., determining specific postures. Other works instead propose a two-steps approach: first, model matching or feature-based procedures are used to determine a posture within a predefined set, then a temporal filter is used to integrate these values over time. For example, (Cucchiara et al., 2005b) uses projection histograms to determine postures and then a Hidden Markov Model to track them over time. In general, when the goal is to recognize predefined postures, temporal filtering allows for improving performance with respect to frame by frame classification.

Finally, multi-camera setting has been used for tracking human body movements: (Demirdjian et al., 2003) use stereo vision and a 3D model of the upper human body for real-time 3D tracking of head, torso and arms, while in (Grammalidis et al., 2001) the parameters of an MPEG4 3D model are estimated using the depth image coming from the person being analyzed. However, posture classification has not been explicitly addressed in these works.

The approach to human posture tracking and classification presented here is based on stereo vision segmentation. Real-time people tracking through stereo vision (see for example (Beymer and Konolige, 1999; Bahadori et al., 2005; Iocchi and Bolles, 2005)) 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 are also available; therefore, when multiple people are in the scene, we have a set of XYZ-RGB data over time for each person. Obviously, this kind of segmentation can be affected by errors, but the experience we report in this paper 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 contribution of this paper 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 novel 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 to model posture transitions and to classify among a set of main human postures: *UP*, *SIT*, *BENT*, *ON KNEE*, *LAID*.

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 view points. 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 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 paper is organized as follows. Section 2 describes the data processed by a stereo vision based people tracker that are used as input for the method described here. Section 3 presents a discussion about the choice of the model that has been used for representing human postures, while Section 4 describes the tracking of the principal points and the computation of the parameters of the model. In Section 5 we present the classification method and finally Section 6 includes experimental evaluation of the method. Conclusions and future work will conclude the paper.

2 IMAGE SEGMENTATION AND PEOPLE TRACKING

The method presented in this paper takes as input a sequence of data in the form XYZ-RGB that are relative to a person tracked in the scene. A stereo vision based people tracker (Bahadori et al., 2005; Iocchi and Bolles, 2005) has been used to produce these data. This system has been proved to be robust to illumination changes, partial occlusions, shadows, provides for real-time implementations and is able to deal with multiple people in the scene. The system provides reliable data that can be actually used for posture detection and classification.

For each tracked person, the system provides a set of data $\Omega = \{\omega_{t_0}, ..., \omega_t\}$ from the time t_0 in which the person is first detected to current time t. The value $\omega_t = \{(X_t^i, Y_t^i, Z_t^i, R_t^i, G_t^i, B_t^i) | i \in \mathcal{F}\}$ is the set of XYZ-RGB data for each pixel i identified as a foreground element in the scene (i.e. belonging to a person). The reference system is chosen in order to have the plane XY coincident with the ground floor.

The tracker system produces two kinds of errors in this 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 $\mathcal{F}$. Examples of segmentation are in the upper part of Figure 3, 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 method described in this paper can reliably tolerate such errors, thus being very robust to noise in segmentation that is typical in real world scenarios.

3 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 (Gavrila, 1999) 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 thus we adopted a model that does not contain explicitly arms and hands.

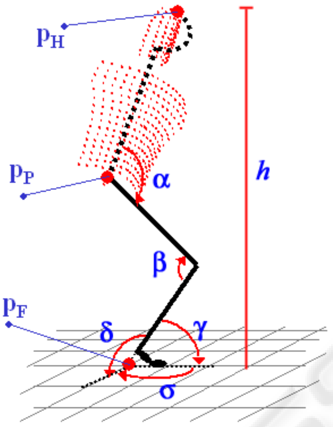


Figure 1: 3D Human model for posture classification.

The model used in our application is shown in Figure 1. It is composed by two sections: a head-torso block and a leg block.

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 a single planar rotoidal joint is instead used.

The model is 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 feet point of contact with floor ($\mathbf{p}_F$) (see Figure 1). These points are tracked over time, as shown in the next section, and used to determine measures for classification. In particular, we are able to estimate the five angles α, β, γ,

δ, σ, and the height h shown in Figure 1.

The parameters of the model that are measured during model matching and then used for pusture classification are $z =< \alpha, \beta, \gamma, \delta, h_{\texttt{norm}} >$, where $h_{\texttt{norm}}$ is the ratio between the height measured at the current frame and the height of the person measured during the training phase, while σ is not included, since it does not contribute to posture detection.

4 TRACKING MODEL PARAMETERS

Detection and tracking of the principal points of the model is important to measure parameters of the model that are used for classification. Our procedure, while attempts to find the position of the head, of the pelvis joint and of the feet point of contact for each image in the sequence, it also analyzes the sequence of observations over multiple frames, so gathering progressively information about the symmetry plane and the direction of the left hand.

Since the human model contains data that must be adapted to the person being analyzed, a training phase must be executed for the first frames in the sequence (ten frames are normally sufficient), to measure the person's height and to estimate the bimodal color distribution. Assuming that in this phase the person is exhibiting an erect posture with arms below the shoulder level, the height is measured for each frame in the training phase and the average value λ over the training sequence is taken.

During the training phase a bi-modal color distribution of the head ξ is also computed. Considering that a progressively correct estimation of the height (and, as a consequence, of the other body dimensions) is available, the points in the image which height is greater than the neck 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 $k = 2$.

After the training phase the Algorithm 1 reported below is used to determine and track over time the parameters of the model. In the algorithm, $\mathcal{M}$ is the model described in the previous section, ξ and λ are the values learned during the training phase described above, and $\lambda_{\texttt{TH}}$ is a threshold. Θ denotes the internal state of the tracker and it is defined by $\Theta = \{\Pi, \tau, \phi\}$, with $\Pi = \{\mathbf{p}_F, \mathbf{p}_P, \mathbf{p}_H\}$ being the set of the three principal points, τ is the symmetry plane of the model, ϕ is the direction of the model within the symmetry plan (it can assume two values: left or right). The algorithm, given the current input data ω, computes the model parameters z and update its internal state (denoted by Θ').

Algorithm 1. ***Track model parameters***

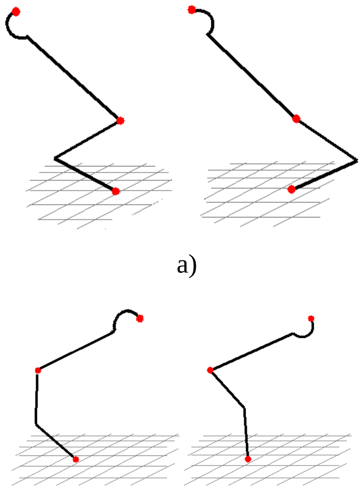
```
INPUT: $\omega$, $\Theta$
OUTPUT: $z$, $\Theta'$
CONST: $\mathcal{M}$, $\xi$, $\lambda$, $\lambda_{\texttt{TH}}$

begin
    $h_{\texttt{M}} = max\{Z^i | (X^i, Y^i, Z^i, R^i, G^i, B^i) \in \omega\}$;
    if ( $\lambda - h_{\texttt{M}} < \lambda_{\texttt{TH}}$ ) {
        $\Theta'$=$\Theta$;
        $z = [0, 0, 0, 0, 1]$;
    }
    else {
        $[\tilde{\mathbf{p}}_P, \tilde{\mathbf{p}}_H] = \texttt{ICP}(\mathcal{M}, \omega, \xi)$;
        if (!$\texttt{leg\_occluded}(\omega, \mathbf{p}_F)$)
            $\tilde{\mathbf{p}}_F = \texttt{find\_leg}(\omega, \mathbf{p}_F)$
        else
            $\tilde{\mathbf{p}}_F = \texttt{project\_on\_floor}(\tilde{\mathbf{p}}_P)$;
        $\Pi' = \texttt{kalman}(\Pi, \tilde{\Pi})$;
        $\tau' = \texttt{filter\_plane}(\tau, \Pi')$;
        $\hat{\Pi}' = \texttt{project\_on\_plane}(\Pi', \tau)$;
        $\rho = \texttt{evaluate\_left\_posture}(\hat{\Pi}', \tau')$;
        $\phi' = \texttt{filter}(\rho, \phi)$;
        $z = [\texttt{get\_angles}(\hat{\Pi}', \tau', \phi'), h_{\texttt{M}}/\lambda]$;
    }
end
```

The first part of the algorithm evaluates the maximum height of the pixels representing the person (we assume that the head of the person is always visible). If the difference between the current height and the person's nominal height is below the threshold $\lambda_{\texttt{TH}}$, then the values of the model are fixed to standard values denoting the erect posture. This simplifies computation in many cases.

Otherwise, the principal points $\mathbf{p}_H$ and $\mathbf{p}_P$ are estimated using a variant of the ICP algorithm (Rusinkiewicz and Levoy, 2001). The model shown in Figure 1 is used to find a match in the image. 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, for example. The ICP is modified to take into account head color information. We consider only those correspondence for the head that are compatible with the color distribution estimated in the training phase. Moreover, since these correspondences, once found, are characterized by a greater amount of information, they have a greater weight, so contributing more in determining the rigid transformation in the ICP minimization error phase.

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 sensitiveness to false negatives. Since we are interested in finding a point that represents the feet 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



a)

b)

Figure 2: Ambiguities.

(this is verified by the function `leg_occluded` in the algorithm), for example if he/she is sitting behind a desk, we can anyway model the observation as a Gaussian distribution with center in the projection on the ground of $\mathbf{p}_P$ and variance in inverse relation with the height of the pelvis from the floor.

After computation of the principal points for each frame, 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 a Kalman Filter to update these values over time. In our implementation we assumed the three components to be independent and a constant velocity model for each of them. This approximation is a good compromise between necessity of filtering and computational cost.

Second, ambiguities may arise in determining poses from three points. To solve this problem we need to determine the *symmetry plane* τ 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 plane passing for the three points 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 temporal filter (`filter_plane`) on the symmetry plane that suitably takes into account co-linearity of these points. Then, principal points are projected onto the filtered symmetry plane ($\hat{\Pi}'$).

Given the symmetry plane τ', we still have two different solutions, corresponding to the two opposite orientations of the person. To determine which one is correct we use a function that computes the likelihood of the orientation of the person. An example is given in Figure 2, where the 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 ρ that the person is oriented on the left. For example, the likelihood for the situation in Figure 2 a) is $\rho = 0.6$ (thus slightly preferring the leftmost posture), while the one in Figure 2 b) is $\rho = 0$ 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 2 a) may be increased (thus preferring more the leftmost posture). By filtering these values uniformly through time, we get a reliable estimate of the orientation of the person ϕ'.

Finally, from $\hat{\Pi}'$, τ', ϕ' the algorithm computes the angles of the model and hence z, that is used in the subsequent classification phase. Moreover, Π', τ', ϕ' denote the internal state of the procedure that will be used in the next cycle.

5 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 a model matching process.

The main feature 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. However, temporal integration of these information allows for a more robust classifier since it allows for modeling also transition between postures.

Therefore, we have implemented two classification procedures (that are compared in Section 6). They use an observation vector $z_t =< \alpha, \beta, \gamma, \delta, h_{\texttt{norm}} >$, 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 can be acquired by analyzing sample videos or synthetic model variations. In our case, since values z_t are computed by 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 $\gamma \in \Gamma$. In addition, due to the uni-modal nature of such distributions, they have been approximated as normal distributions.

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

$$\gamma_{\mathtt{ML}} = argmax_{\gamma \in \Gamma} \;\; P(z_t|\gamma)$$

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

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

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 6.

6 EXPERIMENTAL EVALUATION

Experimental evaluation of the approach presented in this paper has been performed by using an experimental setting formed by a stereo camera placed about 3 meter high from the ground pointing down about 30 degrees from the horizon. A stereo vision based tracker has been used to provide XYZ-RGB data of tracked people 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 paper has an average computation cycle of about 180 ms on a 1.7 GHz CPU. Therefore in combination with the tracker the overall system 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 main objective of the experiments reported here is to evaluate the robustness of the system with respect to occlusions and different view points, as well as the effectiveness of the filter provided by HMM.

The experiments have been performed by considering a set of video sequences, chosen in order to cover all the postures we are interested in, with different people, in different environmental conditions, different orientations and also considering partial occlusions. These videos have been grouped in different ways in order to highlight different characteristics of the system as explained later. For each video we built a ground truth by manually labeling 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. The total number of frames that have been used to compute classification rates in all the experiments is 2085.

For the classification based on HMM we have chosen 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. The transition probabilities $T_{ij} = P(\gamma_t = i|\gamma_{t-1} = j)$ have been set to

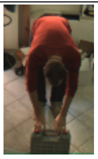
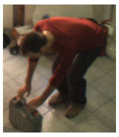
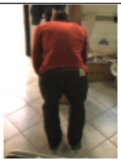
$$\begin{pmatrix} 0.787 & 0.079 & 0.047 & 0.039 & 0.047 \\ 0.147 & 0.827 & 0.019 & 0.004 & 0.004 \\ 0.133 & 0.033 & 0.767 & 0.033 & 0.033 \\ 0.142 & 0.015 & 0.027 & 0.769 & 0.046 \\ 0.148 & 0.037 & 0.037 & 0.037 & 0.741 \end{pmatrix}$$

that have been manually computed by considering more likely transitions from one posture to another, assuming "normal activities" in an office-like environment. In fact, the above matrix is the result of averaging between different transition matrices determined independently by different people. Obviously, this can be customized according to specific application requirements.

Robustness to different view points 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 1 shows classification rates for each orientation. The first row presents results obtained with maximum likelihood, while the second one shows results obtained with HMM. There are very small differences between the five rows, thus showing that the approach is able to correctly deal with different orientations. Also improvement in performance due to HMM is not very high. This is not surprising since postures are well separated in the classification space defined by the parameters of the model.

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 3 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 2 classification rates for the different postures.

Table 1: Classification rate from different view points.

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

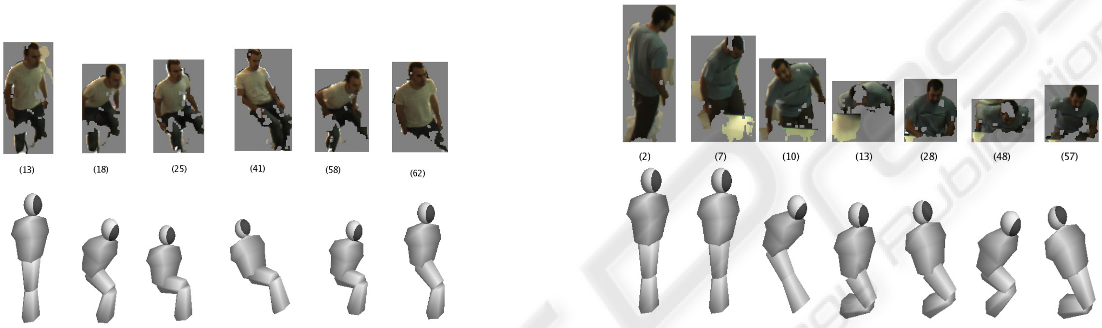


Figure 3: People sitting on a chair (non-occluded vs. occluded).

Table 2: Classification rates without and with occlusions.

	No occlusions	**Partial occlusion**
UP	91.5 %	91.5 %
SIT	88.3 %	81.6 %
BENT	82.8 %	93.3 %
LAID	100.0 %	100.0 %

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 $\mathbf{p}_P$ 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 (see discussion below); 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.

Finally, Table 3 presents the total confusion matrix of all the experiments performed. 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, respectively, 89.5 %, 80.6 % , 88.2 %, 51.7 %, 100 % for the five considered postures. Thus, the HMM slightly improve the performance, however maximum likelihood is still effective, confirming the effectiveness in the choice of the classification space and the ability of the system to correctly track the parameters of the human model. In the *BENT* posture we had better results without the HMM, this was due to a delay in one of the videos in passing from standing to bending position, probably indicating that the transition matrix used here can be optimized for achieving better results. This also shows a possible drawback of using temporal filters: they may introduce delays in switching between postures and thus must be fine tuned.

From the analysis of the classification results, 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

Table 3: Overall confusion matrix with HMM.

System Ground Truth	UP	SIT	BENT	KNEE	LAID
UP	**93.5 %**	1.0 %	5.4 %	0.0 %	0.0 %
SIT	2.1 %	**84.7 %**	7.4 %	5.8 %	0.0 %
BENT	4.2 %	8.3 %	**85.8 %**	0.5 %	1.2 %
KNEE	0.0 %	3.3 %	45.0 %	**51.7 %**	0.0 %
LAID	0.0 %	0.0 %	0.0 %	0.0 %	**100.0 %**

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 4), posture recognition may be wrong.

7 CONCLUSIONS

In this paper 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.

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

One of the problems experienced was that the tracker system works very well when people are in standing position, while quality of data worsen when people sit, lay down, or bend. While the quality of segmentation does not affect classification of the *LAID* posture (that is mainly determined by the pixels height from the ground), segmentation errors are the main causes of classification errors for the other postures. Classification errors may be reduced by providing feedback from the posture classification to the tracker. In fact, given these information the tracker could adapt recognition procedure in order to provide better data.

Additional postures may be considered: interesting cases would be *WALKING* and *JUMPING* that can be detected by analyzing the trajectory of the principal point for the head $\mathbf{p}_H$. Also temporal analysis of the model parameters can be useful to determine for example different ways in which people fall down.

REFERENCES

Bahadori, S., Grisetti, G., Iocchi, L., Leone, G. R., and Nardi, D. (2005). Real-time tracking of multiple people through stereo vision. In *Proc. of IEE International Workshop on Intelligent Environments*.

Beymer, D. and Konolige, K. (1999). Real-time tracking of multiple people using stereo. In *Proc. of IEEE Frame Rate Workshop*.

Boulay, B., Bremond, F., and Thonnat, M. (2005). Posture recognition with a 3d human model. In *International Conference on Crime Detection and Prevention (ICDP)*.

Bregler, C. and Malik, J. (1998). Tracking people with twists and exponential maps. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'98)*.

Cucchiara, R., Grana, C., and Prati, A. (2005a). Probabilistic posture classification for human-behavior analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, 35(1):42–54.

Cucchiara, R., Prati, A., and Vezzani, R. (2005b). Posture classification in a multi-camera indoor environment. In *Proc. of IEEE International Conference on Image Processing (ICIP'05)*.

Demirdjian, D., Ko, T., and T.Darrel. (2003). Constraining human body tracking. In *International Conference on Computer Vision (ICCV'03)*.

Gavrila, D. M. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98.

Goldmann, L., Karaman, M., and Sikora, T. (2004). Human body posture recognition using mpeg-7 descriptors. *Visual Communications and Image Processing*.

Grammalidis, N., Goussis, G., Troufakos, G., and Strintzis, M. G. (2001). 3-d human body tracking from depth images using analysis by synthesis. In *Proc. of IEEE International Conference on Image Processing (ICIP'01)*.

Iocchi, L. and Bolles, R. C. (2005). Integrating plan-view tracking and color-based person models for multiple people tracking. In *Proc. of IEEE International Conference on Image Processing (ICIP'05)*.

Moeslund, T. B. and Granum, E. (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding: CVIU*, 81(3):231–268.

Rusinkiewicz, S. and Levoy, M. (2001). Efficient variants of the icp algorithm. *Proc. of 3rd International Conference on 3D Digital Imaging and Modeling*.

Sminchisescu, C. and Triggs, B. (2003). Kinematic jump processes for monocular 3d human tracking. *Proc. of the Conference on Computer Vision and Pattern Recognition*.