Unsupervised Model Generation for Motion Monitoring

Markus Weber, Gabriele Bleser, Gustaf Hendeby, Attila Reiss, Didier Stricker
Augmented Vision
German Research Center for Artificial Intelligence (DFKI) GmbH
Kaiserslautern, Germany
{firstname.name}@dfki.de

***Abstract*—This paper addresses two fundamental requirements of full body motion monitoring: (a) the ability to sense the input of the user and (b) the means to interpret the captured input. Appropriate technology in both areas is required for an interactive virtual reality system to provide feedback in a useful and natural way. This paper combines technologies for both areas: It develops a sensor fusion approach for capturing user input based on miniature on-body inertial and magnetic motion sensors. Furthermore, it presents work in progress to automatically generate models for motion patterns from the captured input. The technology is then used and evaluated in the context of a personalized virtual rehabilitation trainer application.**

I. Introduction

For the acceptance of virtual reality applications, natural inclusion of the user is, among others, a crucial criterion. This is manifested in the recently increased interest in novel gestural user interfaces that do not depend on conventional input devices, but react on the user acting naturally in the environment — think of the gesture controlled gaming consoles, such as Microsoft XBox Kinect, Wii Motion Plus, Playstation 3 Move. From a technical point of view, such gestural interfaces require two base components, appropriate capturing technology to sense the user actions and appropriate learning and reasoning technology to interpret the captured actions and trigger the expected feedback. This paper combines technologies for both of them.

For capturing user input, a reliable and accurate sensor fusion approach based on miniature body-mounted inertial measurement units (IMUs) is developed. Under full operation, the whole body can be captured with ten IMUs (*cf.* Section II). The captured motion signals are then input to the second part of the paper. Here, a fully automated method for detecting motion motifs is developed. Motions are interpreted as short-time patterns throughout this paper. Using this recurring motion motifs, a hidden Markov model using Gaussian emissions is constructed, which can be used for monitoring user input within virtual environments (*cf.* Section III). The developed technologies are then exploited and evaluated in the context of a personalized virtual rehabilitation trainer application, which supports and monitors previously taught exercises (motions)(*cf.* Section IV). Conclusions and future work are presented in Section V.

II. Motion capture

The pose and motion of the body are contained in the measured accelerations, angular velocities, and magnetic fields from the IMUs attached to it. These measurements are compared to predictions based on a simple biomechanical body model. The pose kinematics are then determined using model based sensor fusion, more precisely using the extended Kalman filter (EKF) [1].

A. Biomechanical body model and calibration

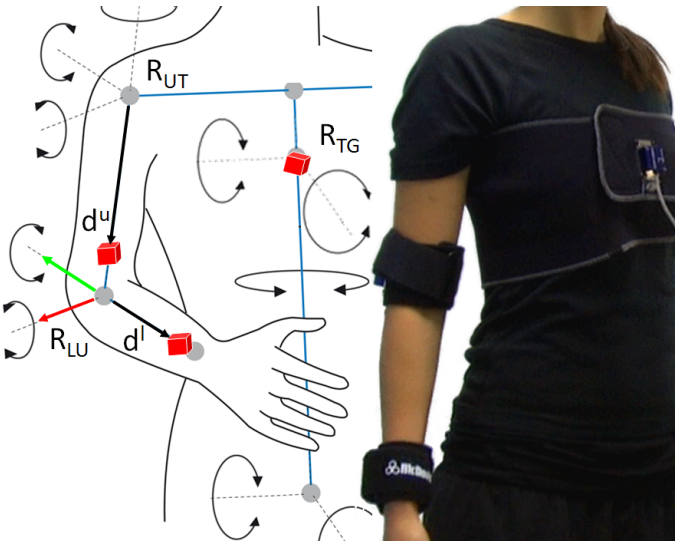


Fig. 1: Functional body model and fixation of IMUs on the upper body of a test subject.

A schematic of the used biomechanical body model is depicted in Figure 1 (left). The complete model consists of ten rigid bodies (bones: torso, pelvis, upper arms, forearms, upper legs, lower legs) connected by anatomically motivated restricted joints. The right side of the figure shows, how IMUs are attached on a test subject. To be able to use the IMU measurements to estimate the body pose, the way the sensors are attached to the segments must be known. The relative positions are determined by measuring the distances along the segments. The calibration procedure for the relative orientations is based on IMU measurements taken under known static poses (*e.g.* [2]).

B. Sensor fusion

The rotations of the torso and pelvis segments are modeled with three degrees of freedom (DOF). They are estimated using a standard attitude and heading reference system approach [3] as implemented in most of the commercially available IMUs. The estimation of limb (arms, legs) motions from two IMUs (one attached on each segment) is handled jointly in one EKF. Contrary to many proposed systems [4], where the upper and lower segment orientations are obtained independently and constraints are enforced in a second step, the approach developed here estimates the limb motion jointly based on forward kinematics only. This has two advantages: (1) the lower segment helps in the estimation of the upper segment, (2) constraints are built-in to start with. Here, the shoulder and hip joints are modeled with three DOF, whereas the elbow is restricted to two and the knee to one DOF.

Subsequently, the state-space model for joint limb motion estimation is described.

Knowing the segment lengths, the limb pose is fully specified by the upper joint rotation, R_{UT}, and the lower joint rotation, R_{LU}. Together with the calibrated quantities (cf. Sections II-A) and the respective torso orientation, R_{TG}, where G denotes the global frame, this information is sufficient to compute the IMU orientations and positions:

$$R_{I^uG} = R_{I^uU} R_{UT} R_{TG}, \quad R_{I^lG} = R_{I^lL} R_{LU} R_{UT} R_{TG} \quad \text{(1a)}$$

$$I^u_G = R^T_{TG} R^T_{UT} d^u, \quad I^l_G = R^T_{TG} R^T_{UT} (d^e + R^T_{LU} d^l) \quad \text{(1b)}$$

with respect to G (*cf.* Figure 1 for the symbols). In order to obtain a minimal parametrization, also with restricted DOF, Euler angles are used to represent the joint configurations. Hence, the system state, $x = [\theta, \dot{\theta}, \ddot{\theta}]$, comprises the joint angles, $\theta = [\theta_{UT}, \theta_{LU}]^T$, with $R_{ab} = \mathrm{rot}(\theta_{ab})$, their velocities, $\dot{\theta}$, and their accelerations, $\ddot{\theta}$. The linear dynamic model assumes constant angular acceleration and zero-mean Gaussian angular acceleration process noise.

The measurement models relate the measured angular velocities, y^ω, accelerations, y^a, and magnetic fields, y^m, in the local IMU frames, I, to the state.

The accelerometers measure a combination of body acceleration, $\ddot{I}$, and acceleration due to gravity, g, in the local IMU frame. Assuming that gravity is the only force acting on the IMUs, the acceleration measurement model is:

$$y^a = R_{IG}\big(\ddot{I}_G(\theta, \dot{\theta}, \ddot{\theta}) - g_G\big) + e^a. \quad \text{(2a)}$$

Here, the body acceleration in the global frame, $\ddot{I}_G$, is a function of θ, $\dot{\theta}$, and $\ddot{\theta}$. It results from differentiating (1b) with respect to time twice and transforming the result to the local IMU frame using R_{IG}. The latter is obtained from (1a). The gyroscope measurement model is:

$$y^\omega = \omega_I(\theta, \dot{\theta}) + e^\omega, \quad \text{(2b)}$$

where the angular velocity in the IMU, ω_I, is obtained by transforming $\dot{\theta}$ to the local frame. The transformation can be derived from the relation $\mathrm{S}(\omega) = (R_{IG} \dot{R}^T_{IG})$, where $\mathrm{S}(\omega)$ is the skew-symmetric matrix of ω [5].

The magnetometers are used as aiding sensors in order to correct for drift due to sensor noise, calibration and model errors. They provide a common forward direction, m_G. The respective measurement equation is:

$$y^m = R_{IG}(\theta) m_G + e^m. \quad \text{(2c)}$$

To simplify the equation and lessen the influence of magnetic disturbances (2c) is reduced to the heading direction. This is achieved by comparing the $\arctan(y, x)$ of both sides of the relation. In (2), e denotes mutually independent zero-mean Gaussian measurement noise.

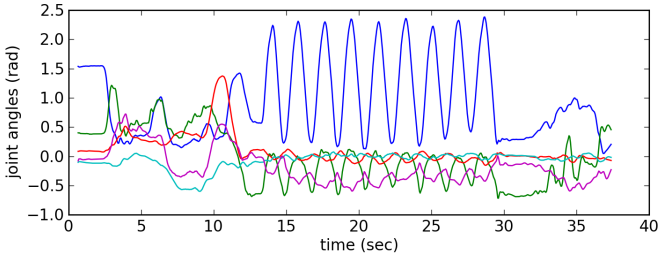
III. Motion analysis and model construction

Based on the joint angles provided by the motion capture system, this section describes a fully automated method for constructing a hidden Markov model (HMM) for a motion pattern from a very short training sequence, such as the one shown in Figure 2a. The training data is assumed to contain a pre-defined number of pattern examples performed by the user during the training step. The HMM representation is chosen for two reasons: (1) it naturally takes variations in motion into account by allowing for time-warping and has thus been successfully applied in domains such as speech, gesture, or handwriting recognition, (2) standard algorithms, such as the Viterbi algorithm, can be used for online monitoring. While the model generation is described and evaluated in this paper, the online monitoring is work in progress and is therefore only indicated here.

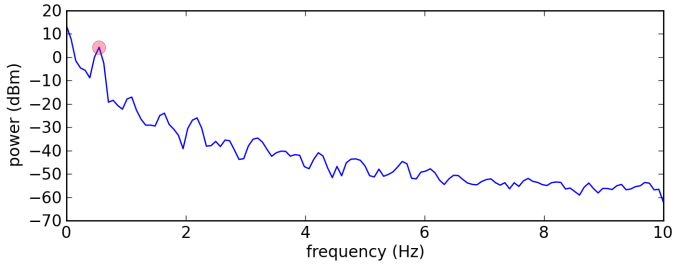
The proposed method for model generation consists of two steps: The first step is to automatically detect motif candidates, *i.e.* the recurring patterns in the training sequence. The second step is to use the detected motifs to construct the model.

A. Motif candidate detection

The problem of locating motifs in real-valued, multivariate time series is a known problem and several approaches have been proposed [6], [7]. However, all of these methods are based on a pre-defined window size. Since the length of the pattern is here unknown, the first step consists in estimating a suitable window size w_{est}. Based on the assumption that the repetitions in the training sequence are performed consecutively with roughly the same speed, a dominant frequency will be present in the signal. This can be extracted using the combined power spectral density (PSD) [8] (*cf.* Figure 2b). The window length w_{est} is then initialized as the wavelength of the dominant frequency, $w_{est} = \lambda = \frac{v}{f_{dominant}}$, with v being the sampling rate. An extended version of Minnen's method [7] is then parametrized with w_{est} to detect the motif candidates: The method collects overlapping subsequences, S_i, of length w_{est} from the training signal, S, and determines the k-nearest neighbors for each subsequence as $\mathrm{kNN}(S_i) = S_{i,1..k}$. Here, k is the pre-defined number of repetitions. In order to reduce the sensitivity to local time shift and slightly varying execution speed, dynamic time warping (DTW) is used as distance measure.



(a) Multi-variate motion data (joint angles of the upper body).



(b) Combined PSD of all channels. The dominant frequency is marked with a red circle.

Fig. 2: Example training sequence.

A real motif should have at least k similar subsequences. Hence, in order to find good motif candidates, for each subsequence, S_i, the distance density is estimated as the reciprocal of the distance to the least similiar neighbor k: $den(S_i) \propto \frac{1}{dist(S_i,S_{i,k})}$ The motif candidates, $cand_i$, are then identified as the local maxima of the densities among its k nearest neighbors:

$$maxima(S_i) = S_i : \ \forall \ S_{i,j} \ den(S_i) > den(S_{i,j}),$$

where $j = [1, k]$.

B. Model generation

The observation probabilities of the HMM are modeled using Gaussian mixtures models (GMM). Here, the different channels of the multivariate signal are handled separately. In a first step, a model is learned for each detected motif candidate $cand_i$. For the i^{th} motif candidate, the respective sequence and its k-nearest neighbors are used as training set $TS_i = \{S_i \cup \text{kNN}(S_i)\}$. Since traditional parameter estimation methods for HMMs, such as the Baum-Welch algorithm, typically fail when applied to too few training examples, a simple construction algorithm is applied to capture the characteristics of each motif. This algorithm builds a HMM with left-right topology, which is a wide-spread approach to model time-varying sequential data. Self-transitions and skip-transitions are added to allow for a faster and slower execution of the pattern. The number of states, N, is chosen as half the estimated window size: $N = ceil(\frac{w_{est}}{2})$. Accordingly, each subsequence is divided into N equal-length adjacent segments and each segment is assigned to a state ST_i (*cf.* Figure 3). For each state, ST_i, a GMM is then trained using an expectation-maximization algorithm on all respective elements of TS_i. Thus, each segment is described by one normal distribution. In order to generate the final model, the best candidate model

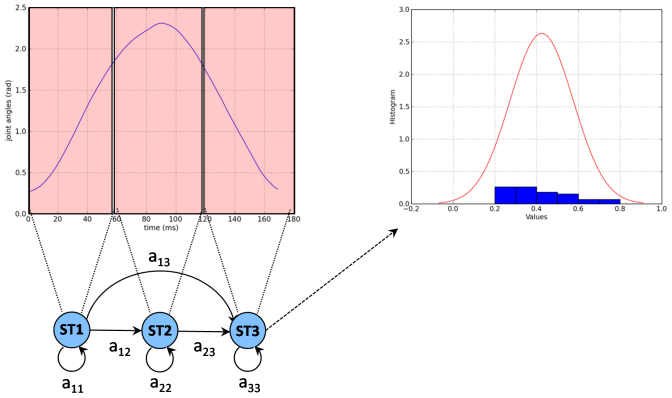


Fig. 3: HMM for one channel of the signal.

must be determined. For this, each candidate HMM is used to refine the detected occurrences of the respective motif. The detection is performed per channel using the standard Viterbi algorithm, which computes for a given observation the Viterbi path and its log likelihood. The Viterbi path represents the most likely sequence of hidden states.

The Viterbi algorithm enables to detect variable-length occurrences of the modeled motion pattern and therefore provides a refined segmentation compared to the overlapping subsequences used in Section III-A. The lower bound for accepting a such detected occurrence is given by the lowest determined log likelihood of the original training set TS_i.

Having refined all training sets, the model which has the best log likelihood score for the pre-defined number of occurrences is chosen as the best model. Finally, the parameters of the HMM are re-estimated using the respective refined training set.

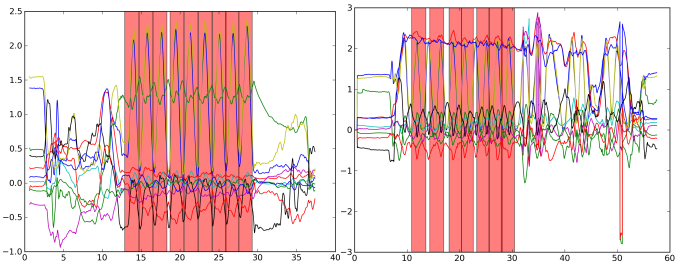
IV. Application, Evaluation and Preliminary Results

The technology presented above is currently used in a virtual trainer for personalized home-based rehabilitation. This is of high interest for, *e.g.* the rehabilitation of stroke patients, which often requires patient-specific and long-term exercise to regain full mobility. The motion patterns to be trained are therefore rehabilitation exercises. The idea behind is the following: A specific patient learns the correct execution of exercises together with a physician. During this training session, GMM-HMMs are constructed for each exercise. At home, these models are then used by the system to monitor the correct execution of exercises, automatically count the number of repetitions, detect anomalies, and, based on this, give immediate feedback to the patient. Figure 4 depicts the system under operation.

Within this application scenario, the developed techniques are evaluated on typical exercises. A test subject is equipped with ten Colibri IMUs from Trivisio [9] (*cf.* Figure 1) and repeats each of the exercises a couple of times. The joint angles are estimated and logged at 100 Hz using the real-time C++

Fig. 4: The virtual trainer in use: The user exercises equipped with two IMUs on the arm in front of a big screen. The movements are transferred to an avatar in a virtual gym environment. The system counts the biceps curls and fills up green bars to indicate the number of repetitions.

implementation of the capturing system described in Section II. The recorded data is used to for the offline training process. GMM-HMMs are constructed from the training sequences as described in Section III. Figure 5 shows the successful discovery of the exercise patterns for two different exercises, biceps curls (*cf.* Figure 5a) and wall pushups (*cf.* Figure 5b). The variable-length pattern occurrences are successfully detected based on the trained GMM-HMMs using the standard Viterbi algorithm with a sliding window. Moreover, as shown in the figures, noisy data in the beginning and the end of the training sequences is correctly ignored. For the online recognition of the motion patterns, we a currently investigate the application of the short-time viterbi algorithm [10].



(a) Biceps curls. (b) Wall pushup

Fig. 5: Recognition results for two different exercises.

V. Conclusion and Future Work

This paper presents an inertial motion capture system using statistical sensor fusion, and a general learning-based approach for recognizing patterns in the captured motion signals, or rather in multivariate signals in general. While the developed methods provide powerful tools for all applications, where previously learned patterns have to be recognized — think of 3D gestures and embodied interaction in general — the technology is here used and evaluated in the context of a personalized virtual rehabilitation trainer application.

The proposed methods are work in progress and there will be a upcoming evaluation of the whole system in a clinical trial. Future work will consist in further developing the online monitoring part, including the development of the feedback system for the users. As the current learning procedure is offline, there is still the need to optimize the model for the real-time system. In order to deal with this real-time aspects of online motion monitoring, we will investigate the application of the short-time viterbi algorithm [10]. This modification of the original viterbi algorithm has been successfully applied on the task of real-time phoneme recognition. The usage of additinal sensors, such as cameras, for improving the precision of the motion capturing system is also planned. Furthermore, the recognition method should be extent with an implict evaluation of how accurate the motion is performed. As for the task of motion monitoring a direct feedback on the deviation would be of interest.

Acknowledgements

This work has been performed within the project PAMAP funded under the AAL Joint Programme (AAL-2008-1). The authors would like to thank the project partners and the EU and national funding authorities for the financial support. For more information, please visit the website http://www.pamap.org.

References

[1] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, ser. Mathematics in Science and Engineering. Academic Press, Inc, 1970, vol. 64.

[2] A. G. Cutti, A. Giovanardi, L. Rocchi, A. Davalli, and R. Sacchetti, "Ambulatory measurement of shoulder and elbow kinematics through inertial and magnetic sensors," *Medical and Biological Engineering and Computing*, vol. 46, pp. 169–178, 2008.

[3] T. Harada, T. Mori, and T. Sato, "Development of a Tiny Orientation Estimation Device to Operate under Motion and Magnetic Disturbance," *The International Journal of Robotics Research*, vol. 26, no. 6, pp. 547–559, 2007.

[4] H. M. Schepers, D. Roetenberg, and P. H. Veltink, "Ambulatory human motion tracking by fusion of inertial and magnetic sensing with adaptive actuation," *Medical and Biological Engineering and Computing*, vol. 48, pp. 27–37, 2010.

[5] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry, *An Invitation to 3-D Vision*, S. Antman, J. Marsden, L. Sirovich, and S. Wiggins, Eds. Springer Verlag, 2003, vol. 26.

[6] D. Minnen, T. Starner, I. Essa, and C. Isbell, "Discovering Characteristic Actions from On-Body Sensor Data," *2006 10th IEEE International Symposium on Wearable Computers*, pp. 11–18, Oct. 2006. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4067720

[7] D. Minnen, C. L. Isbell, I. Essa, and T. Starner, "Discovering multivariate motifs using subsequence density estimation and greedy mixture learning," in *Proceedings of the 22nd national conference on Artificial intelligence - Volume 1*. AAAI Press, 2007, pp. 615–620.

[8] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?\&arnumber=1161901

[9] Trivisio, http://www.trivisio.com.

[10] J. Bloit and X. Rodet, "Short-time viterbi for online hmm decoding: Evaluation on a real-time phone recognition task," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 2121 –2124.