EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Stylistic gait synthesis based on hidden Markov models

Joëlle Tilmanne*, Alexis Moinet and Thierry Dutoit

Abstract

In this work we present an expressive gait synthesis system based on hidden Markov models (HMMs), following and modifying a procedure originally developed for speaking style adaptation, in speech synthesis. A large database of neutral motion capture walk sequences was used to train an HMM of average walk. The model was then used for automatic adaptation to a particular style of walk using only a small amount of training data from the target style. The open source toolkit that we adapted for motion modeling also enabled us to take into account the dynamics of the data and to model accurately the duration of each HMM state. We also address the assessment issue and propose a procedure for qualitative user evaluation of the synthesized sequences. Our tests show that the style of these sequences can easily be recognized and look natural to the evaluators.

Keywords: motion capture, hidden Markov models, style, expressivity, gait, motion synthesis

1 Introduction

Human motion is a very complex field of study. The components of our behaviors, which are so natural to the human eye, can hardly be separated into physiological processes, the personal style of every human, or some kind of additional "style" or "mood" that influences the final motion, as presented in many works (see for instance [1]). A given gesture will easily be identified by the human eye as clumsy, elegant, heavy, or any other characteristic. Unfortunately, automatically extracting that information is a very difficult task, as stylistic variation is intrinsically merged with the basic motion, the individuality of the subject and the time-variability of the gesture (two motions by the same subject will never be exactly the same).

A broad field of applications can be found for human motion synthesis. While its use is currently mostly limited to the entertainment industry, with 3D movies, video games, virtual agents, etc., other domains could benefit from a realistic automatic motion synthesis, in the same way as they already benefit from motion capture [2]. Medical applications could use it for instance to control active prostheses, or try to detect and understand the motion of motor impaired individuals [3]. New applications in the field of the animation of virtual characters in 3D could also take advantage of more evolved motion synthesis, for virtual agents interacting in real-time with the user, for 3D animation movies, for video games, etc. [4]. A good synthesizer could for instance enable non-professional animators to produce believable and controllable motion sequences without animation experience nor expensive motion capture equipment. For artistic applications, motion analysis/synthesis could make it possible for an actor or a dancer to interact in real-time on the scene with a virtual character whose motions are correlated with those of the human performer or with any other signal.

In the framework of virtual character animation, several approaches are available to synthesize realistic human motion. Among those, motion capture based approaches have been driving a lot of interest in the last years, especially since motion capture becomes more affordable. Numerous methods have been developed for using and reusing motion capture data [5], a technology that transfers the movements of humans into a numerical form usable by computers. The main problems encountered with motion capture data are its high dimensionality, the choice of the parameterization, and the variability associated with motion in general. All these factors make it hard to retrieve, analyze, adapt, and modify motion patterns either made "on request" or coming from an existing motion database.

* Correspondence: joelle.tilmanne@umons.ac.be
TCTS Lab, Faculté Polytechnique (FPMs), University of Mons (UMons), Boulevard Dolez 31, 7000 Mons, Belgium

Two approaches coexist for using motion capture data for producing animations: the "template-based" and the "model-based" approaches. In the "template-based" approach, a large database of motion sequences is built and algorithms are developed to address the common data-mining issues (like retrieving the required motion segments), edit these motion parts if needed, and blend them together to produce new sequences [6]. Several problems are associated to the "template-based" approach, that rely on a database which is queried for motion segments. The database needs to be stored, which can be a first issue, and to be large enough to contain all the required motion capture segments. But increasing the database size can be a problem for effective searching. In the synthesis process, unrelated motion parts have to be concatenated, and it is difficult to ensure the continuity of the produced motion. Controllability is also an issue, as there is no continuous modeling of styles, only distinct independent examples that can display several characteristics.

The "model-based" approach, sometimes also referred to as the "machine learning" approach, consists in training models based on motion capture data. The models can later be used to synthesize new motion sequences without resorting to the database initially used for training [7-10]. Furthermore, style can be modeled as a parameter of the model, giving the user new possibilities for the control of his synthesized segments, and for the combination of styles not available in the original data. This approach has been used for years in speech processing for example, first for recognition and more recently for synthesis [11].

Our current work falls in the latter category, with the use of model-based techniques, and more precisely of hidden Markov models (HMMs) [12], for the modeling and synthesis of human-like motion. We aim not only to synthesize a plausible human walk but also to isolate some kind of "style" component. Taking into account such a "style" parameter will enable us to synthesize a broad range of styles, and to have an open model where new styles can always easily be added.

In this work, a general model of "neutral" walk is built in a first step, by training a model over a large database. The resulting neutral-style model can then be used as a basis for the adaptive training of any style-specific model using only a small amount of training data from the target style. A new style-adapted model can thus be obtained very easily each time it is required by capturing only about a dozen steps of the desired walk style and running the adaptive training. This technique, which was originally developed for speaker adaptation in speech synthesis (HTS toolkit) [11,13,14], has been adapted to the motion synthesis problem in our work. The main interest of this approach is that it makes it possible to tackle the main problem of model-based techniques, which is the large amount of data needed to train each new model (corresponding to each different walk style in our case). Thanks to this work, it is possible to train representative models for walk styles for which the training of standard models failed because the set of data available for each style was too small. This method also opens interesting paths for style interpolation or for adding style to plain walks.

The article is organized as follows. Section 2 makes a review of HMM-based motion analysis/synthesis. The training databases are then presented in Section 3. Section 4 describes the preprocessing of the data. Section 5 presents the HMM training and adaptation procedure and its use for synthesis of new stylistic walk sequences, along with some results. A qualitative user evaluation is presented in Section 6. Section 7 concludes this article by presenting perspectives and future works.

2 Related work

2.1 HMMs for motion synthesis

Various walk synthesis algorithms use statistical learning techniques to automatically extract the underlying rules of human motion, without any prior knowledge, directly from training on 3D motion capture data. The resulting statistical models can then be used for generating new motion sequences automatically, using only some high-level commands from the user. Such synthesized motions are thus visually different from the training motions but stochastically similar to them. A few studies use principal component analysis (PCA), not for reducing the dimensionality of the angle data, but as a way of modeling motion units composed of a sequence of frames. Thanks to their periodicity, walk cycles are especially well suited for such an algorithm. This approach has been taken for instance by Glardon et al. [8], Troje [15] and in our previous work [10]. But most work in this area use variations of HMMs, Markov chains or other kinds of probabilistic transitions between motions [9,16-18], in order to take the high dynamic complexity of human movement into account.

A HMM consists of a finite set of states, with transitions among the states governed by a set of so-called transition probabilities. In HMMs, each state is associated with an outcome (more generally called observation) probability distribution. Only this outcome is visible for an external observer, not the state that produced it: at each time t, the external observer sees one outcome, but does not know which state produced it. HMMs are hence double stochastic processes, as visible outcomes are determined by the outcome probability distribution associated with the state, and as the state changes at each time according the transition probabilities between states. In our work, the outcome of the HMM are the frames of the motion. A basic

left-to-right HMM with no skips is illustrated in Figure 1, as a model in which the only possible state transitions at each time are either to stay in the same state or to go to the next state.

Motion has been studied with numerous variants of HMMs, whether it was for analysis or synthesis purposes [19]. In the following paragraphs, we will focus on the studies related to the use of HMMs for motion synthesis, and not just motion in general. In some cases, some kind of "style" component is taken into account, but no style parameter has yet been found that can be used to synthesize styles that are very different from the motions on which the system was trained. Increasing the number of styles represented by the system means increasing the complexity of the model and in most cases re-training it completely, with the additional issue that enough data must be available for each style.

Tanco and Hilton [16] describe a model consisting of two hierarchical levels. In the first level, PCA is used to reduce the dimensionality of the data and is followed by a *K*-means clustering of the poses space. The clusters–defining the boundaries of "motion segments" in the original training data–are then used as states of a Markov chain that represents the temporal behavior of the training data. A discrete HMM is only used in the second level to relate the states of the Markov chain to the original full examples of motion sequences from the training database (the Markov chain states are the observations of the HMM and the hidden states of the HMM are the motion examples). During synthesis, a sequence of Markov chain states is calculated given beginning and end poses defined by the user. The second synthesis stage takes the generated state sequence as an input and searches for the most likely sequence of motion segments from the original training data that could have generated that Markov chain state sequence. There is thus no "true" HMM synthesis step, as the database needs to be accessed each time a new motion has to be built. This work is more related to the template-database approach of motion capture animation, as it was described in Section 1, than to the approach we describe in the present article.

Wang et al. [17] go further in their motion modeling by using a "time-striding HMM" (TSHMM), which is also a two-layer model. In the first layer, an approximation of high-level (time-striding) statistical transitions is calculated, with first order transition probabilities. Those "high level" transitions correspond, for example, to the transitions between two different behaviors like walking and running. The high-level states from the upper layer are modeled in the second layer by a set of left-right HMMs. Those HMMs correspond to "atomical movements", i.e., motion segments maximally short, while being long enough to enable the prediction of the next pose. Synthesis is only based on the model without needing to reuse any motion segment from the original database.

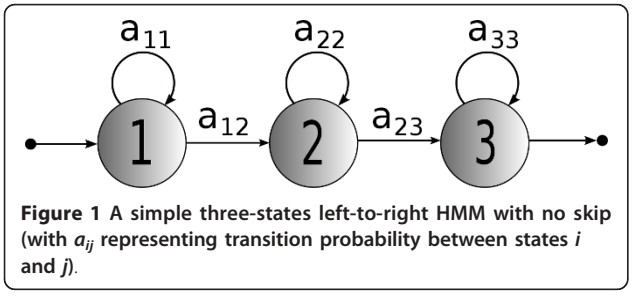

Figure 1 A simple three-states left-to-right HMM with no skip (with a_{ij} representing transition probability between states *i* and *j*).

Li et al. [18] also use the principle of motion decomposition into sub-units connected to each other by transition probabilities, and model each sub-unit individually. Their system, called "motion texture", is a technique for synthesizing complex human motions (like dancing for instance) so that they are statistically similar to the original motion capture data. The model is made of a set of "motion textons", and of their distribution, thereby characterizing the stochastic and dynamic nature of motion captures performed for the training. They define "motion textons" as the repetitive patterns in complex human motion (for instance: spinning, hopping or tiptoeing for dance motion). Each motion texton is modeled by a linear dynamic system (LDS) [18]. The distribution of the textons is modeled by a transition matrix which gives probabilities for transiting from one texton to another. It is thus possible to generate new animations and vary their execution by modifying motion at the texton level, or to synthesize a new choreography by varying the distributions.

2.2 Motion style modeling and synthesis

An interesting approach is chosen by some researchers who try to integrate a "style" variable into their HMM models. It enables the model, during the synthesis step, to vary not only the motion itself, but also the way the motion is performed, i.e. the "style" of the motion.

Wang et al. [20], for instance, use a training algorithm which integrates statistical optimization techniques with the expectation-maximization (EM) learning steps. Their method, called "SOMN-HMM" (which stands for "self-organizing mixture networks" which are used to represent mixture of Gaussians in the HMMs), makes it possible to train basic HMMs as well as parametric HMMs containing a "style" parameter. In [21], output densities are represented by "stylized decomposable triangulated graphs" (mix-SDTG) instead of SOMNs, and they also take into account a style variable.

Among all the models enabling the generation of data representing motion thanks to approximation functions, the "style machine" developed by Matthew Brand [9] is

especially appealing. The major interest brought by this method is that, thanks to its learning algorithm based on the maximization of entropy, it enables to train HMMs for which we do not know the structure in advance, and it does it without having to proceed by successive attempts in order to find the adequate structure. Furthermore, this method integrates a style variable that can vary during the synthesis of a motion sequence. However, in that work the "style" variable is not explicit and it is thus not possible to control directly a given style, but rather to change some intrinsic style-related parameters.

In an approach closely related to ours, Yamazaki et al. [22] synthesize walk using a hidden semi-Markov model (HSMM). The "style" variation they incorporate in their model thanks to multiple regression is the quantitative variations of speed and stride length. There are thus two values that can be controlled but multiple regression is not suited for expressivity modeling which can hardly be quantified with a numerical value. The multiple regression method is trained once for all and it is not possible to add a new "style characteristic" without having to train the whole model again.

One of the problems with motion synthesis is that, unlike for speech which is decomposed into sentences, words, phonemes, etc., which are universal and can be represented as a finite set of possibilities, there is no widely accepted "dictionary" of basic motions. Each research team uses its own terminology and the possibilities are potentially infinite. There is thus no common basis for comparison, and as there is no method to assess the quality or the realism of a synthesized motion, the comparison of methods proposed by each research group is not straightforward. Most studies even lack qualitative assessment of their results.

3 Training databases

In all model-based techniques, the first major issue is to obtain enough representative training data. The quality of models is highly dependent on the quality of the data and how accurately these data describe the studied phenomenon. Motion capture being the only solution to obtain realistic 3D human motion data [2], it is the only way to gather representative training data for statistical modeling of human motion.

In this work we have used two databases recorded with an inertial motion tracking system, the inertial gyroscopic system (IGS-190) from Animazoo [23]. The IGS-190 is a commercial motion capture suit that contains 18 inertial sensors, which each consist of a three axis accelerometer, a three axis gyroscope and a three axis magnetometer. The data from those three sources are integrated and fused directly in the inertial sensor boxes. Angles between the body segments are thus provided straight from the sensors; no mapping is necessary between tracked 3D positions of markers and joint angles, unlike in optical motion capture systems.

Most studies use optical motion capture systems, which usually induce space limitations and where walk is thus recorded on a treadmill. In contrast, the inertial suit IGS-190 does not imply any kind of space limitation. The recorded subject can thus move freely in an open space area and walk can be recorded in a more natural way. This kind of inertial suit is thus especially interesting for the study of expressive walk, as it gives more freedom to the subject who can follow non-straight trajectories and is not constrained to a given constant speed like he would be on a treadmill.

In our databases, like in all motion capture systems, the human body structure is approximated by a kinematic tree of joints modeled as points separated by segments of known constant lengths (see the skeletons in Figure 2). The starting point of that kinematic chain, also called the "root" of the skeleton, is the middle of the hips, at the bottom of the spine. The hierarchy of the skeleton is the same for all our subjects and recordings and contains 18 articulations. This hierarchy and the limb lengths corresponding to the recorded subject are defined for each person prior to the first recording

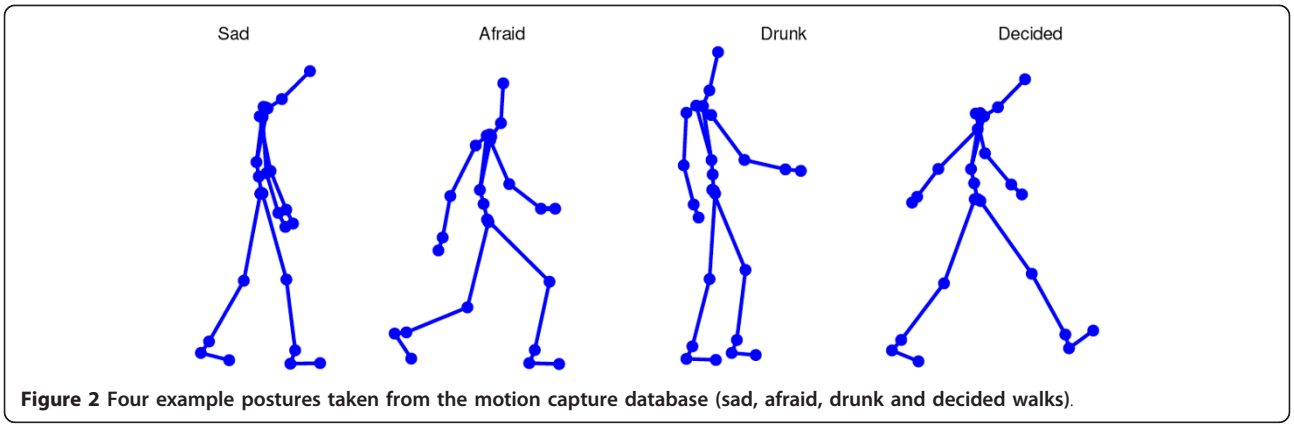

Figure 2 Four example postures taken from the motion capture database (sad, afraid, drunk and decided walks).

and are constant across different motion recordings by the same subject.

There is no 3D position tracking system in the IGS suit, and the absolute position of the subject is calculated by the software given a known initial position, using the length of the skeleton segments from the feet to the hip, the angles recorded between those segments for each frame, and always considering that the lowest point of the skeleton is in contact with the ground.

Our two databases, respectively called "eNTERFACE'08 3D" and "Mockey", were recorded with the same motion capture suit but with different aims, subjects and settings. The eNTERFACE'08 3D database is described in details in [24]. This first database contains 17 walk sequences for 41 subjects. Among these, 12 sequences correspond, for each subject, to three sequences of straight walk over approximately seven meters for four different speed instructions. Those four instructions were "free", slow, middle and fast walks. In the "free" walk, subjects were invited to walk at their usual comfort speed. In the present work, the three free walk sequences of the 41 subjects were used to train our average "neutral" walk model. In that database, the motion was captured at a frame rate of 60 frames per second (fps).

The Mockey database, the second database used in this work, aims to study the "expressivity" of walk [10]. Various walks were performed by the same actor walking on a scene. He was given instructions about the "walking style" he had to act before each walk sequence recording. The 11 different acted styles were the following: proud, decided, sad, topmodel, drunk, cool, afraid, tiptoeing, heavy, in a hurry, manly. These 11 styles were arbitrarily chosen as they all have a recognizable influence on walk, as illustrated in Figure 2. Our "style" component consists thus in exaggerated variations that can be far from plain walk. In this second database, motion was recorded at a frame rate of 30 fps. Depending on the style of walk performed and its corresponding step length, a different number of walk cycles was recorded for each style. The 11 different styles and their corresponding number of left and right steps are presented in Table 1.

Table 1 Mockey database walk styles and corresponding number of steps recorded

Walk	Style	Nbr steps	
		Left	Right
1	Proud	26	24
2	Decided	18	15
3	Sad	35	33
4	Topmodel	28	27
5	Drunk	40	40
6	Cool	25	25
7	Afraid	19	18
8	Tiptoeing	18	20
9	Heavy	24	25
10	In a hurry	20	21
11	Manly	19	20

4 Data preprocessing

In the data format we use, three values per frame give the absolute 3D position (XYZ cartesian coordinates) of the root of the skeleton while the 54 other values represent the 3D angles of the 18 joints of the skeleton. The three values corresponding to the 3D position were discarded as they can be recalculated later using the angles, information about the foot contact with the ground, and the fixed leg segment lengths. The directions of all walk sequences from both databases were then aligned before further processing. The walk sequences were also manually segmented into left and right steps. The boundaries of the steps were arbitrarily defined as the moment the heel touches the ground.

We chose to model the rotations of the 18 captured joints rather than the 3D cartesian coordinates of these joints in order to ensure that the fixed limb length constraints were respected in the synthesized motion: as only rotations are applied to the fixed limb length skeleton definition presented in Section 3, there will be no length deformation in the skeleton after synthesis. This would not be the case with joint cartesian coordinates as nothing would insure that the distance between two successive joints of the skeleton hierarchy remains constant, unless that constraint is explicitly added in the synthesis algorithm.

Once we had chosen to model rotations, the choice of the rotation parameterization was not straightforward. Lots of problems are associated with the different 3D rotation representations that exist, and none of them is ideal in all situations. Rotation matrices, Euler angles, quaternions, axis/angle representation and exponential maps are the most common rotation parameterizations (see for instance [25] for a more detailed presentation of those five representations), but the choice of the parameterization will always depend on the application of interest.

Our data was originally represented by Euler angles, in which each 3D rotation is splitted into three simpler successive rotations around the axes of the local coordinate system associated to the object (X, Y and Z axis). That representation is not well suited for our purpose as, among other issues, there is not always a single representation of each 3D rotation but rather several possible angle combinations that lead to the same rotation. More information about singularities in the Euler angle parameterization can be found in [25,26].

In this work, our angles were converted into the exponential map parameterization which is locally linear and

where singularities can be avoided [26,27]. Exponential maps represent any 3D rotation by a single rotation about an axis. In this parameterization, each vector $\vec{r}$ in $\mathbb{R}^3$ is associated to a single rotation:

$$\vec{r} = \theta.\vec{u}, \tag{1}$$

where the vector $\vec{r}$ is the three-component exponential map, $\vec{u}$ is the unit-length 3D vector corresponding to the axis of rotation, and θ is the rotation angle around the axis. The direction of $\vec{r}$ defines the rotation axis $\vec{u}$, and the magnitude (θ) of the vector $\vec{r}$ is the scalar value of the angle to rotate by. This relationship is completed by associating the zero vector to the identity rotation, making the relationship continuous. For in-depth analysis of the advantages and drawbacks of exponential maps, please refer to [26].

The pose of the skeleton at each frame of the walk cycle is thus described by a vector with a fixed number of variables: 18 tridimensional joint angles, which gives a vector of 54 values per frame to describe the motion.

5 Average model and style adaptation

5.1 Method

As explained before, our objective was to synthesize stylistic walks with few data, starting from a robust neutral walk modeling. Our approach is to start from a procedure originally developed for speaker adaptation in speech synthesis and to adapt it to our motion problem. Both speech and motion fields present strong similarities, like inter-subject variability, stylistic or temporal variations. They are also very different; for instance, motion data do not need feature extraction or temporal windowing, have a much higher dimensionality, and cannot be represented by a finite number of phonemes. This led us to reduce our study to walk synthesis alone, as opposed to motion synthesis in general. In this paragraph, we will briefly explain the different stages of the HMM-based motion synthesis as we used it, based on the HTS framework [11].

5.1.1 Parameter analysis, model structure and labels

Let us assume that our training data C consists in T realizations of our 54-dimensional parameter vector c_t: $C = [c_1, c_2, ..., c_t, ..., c_T]$. As presented in Section 4, our feature vector (c_t) consists in the 54 exponential map parameters describing the skeleton pose at frame t, so we have $c_t = [c_t(1), c_t(2),..., c_t(54)]^\top$. Following the procedure proposed in the HTS framework, the dynamics of the data was taken into account in our models by concatenating c_t with a vector containing the first and second time derivatives of our parameters (for both neutral and stylistic model training) [28]. The observation vector o_t we want to model thus consists of the static feature vector c_t plus the corresponding dynamic feature vectors Δc_t, and $\Delta^2 c_t$, which makes o_t a 162-dimensional parameter vector. Our observation vector o_t can thus be expressed as $o_t = \left[c_t^\top, \Delta c_t^\top, \Delta^2 c_t^\top\right]^\top$, where the derivatives were calculated as follows:

$$\Delta c_t = \frac{1}{2}\left(c_{t+1} - c_{t-1}\right), \tag{2}$$

$$\Delta^2 c_t = \frac{1}{4}\left(c_{t+2} - 2c_t + c_{t-2}\right). \tag{3}$$

Taking into account the T observation vectors, our whole training data can be expressed as $O = i[o_1, o_2,..., o_t,..., o_T]$. Considering matrix W representing the coefficients that link the c, Δc, and $\Delta^2 c$ as expressed in Equations (2) and (3), the relation between the observation matrix O and the static parameter matrix C is:

$$O = WC. \tag{4}$$

In HTS, the time d spent in each state of the HMM is explicitly modeled in duration probability density functions thanks to HSMM [29], a variation of HMMs which takes state duration modeling into account. The schematic representation of an HSMM is represented in Figure 3 and can be compared to the classical HMM of Figure 1. This prevents the probability density of the duration d from being modeled as a decaying exponential like in classical HMMs, as this is inaccurate for most real life problems, like motions in our case. State duration densities were modeled with a multidimensional Gaussian distribution for each HMM. The dimension of these distributions is equal to the number of states in the HMM, set to five in our work, with each dimension corresponding to one HMM state, as explained in [29].

During training, contextual factors related to the position of the step in the whole walk sequence were taken into account, thereby multiplying the number of models to train. However, all model parameters can not be estimated with sufficient accuracy if we only have limited training data. Furthermore, all the possible combinations of contextual factors will not always be present in the training database and unseen models have to be taken into account before the synthesis step. To overcome this problem, both parameter and duration models can be clustered using decision trees. The decision tree is a

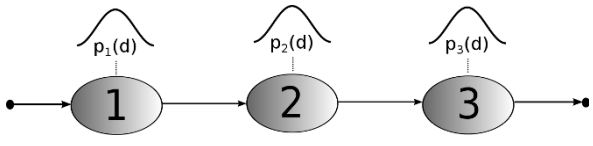

Figure 3 A three-states HSMM (with $p_i(d)$ representing the density probability of the duration of state i).

binary tree, and in each of its nodes, a question splits contextual models into two groups. All possible contextual combinations can be found by traversing the trees. Once the decision tree is constructed, unseen contexts can be taken into account and leaves containing little or very similar data can be merged (for more information on how trees are built and used, please refer to [30]).

5.1.2 Average model training

Using the above HSMM model taking both static and dynamic parameters into account, we train an average walk model on a large set of walkers. This average model will be used in the next step of our procedure as the initial model from which the adaptation will start. In our work, the step boundaries of our segmented database are only used to initialize the parameters of the average walk model (they are not used in the adaptation or synthesis stages). A "walker adaptive training" (WAT) algorithm was used during the training stage of our average model. This WAT training reduces the influence of walk differences among the 41 walkers of our training data on the parameters of the final average model. More information on the WAT training of the average model can be found in [31], where it is referred to as "SAT" for "speaker adaptive training".

5.1.3 Style adaptive training of HSMM models

In the previous paragraph, the general HSMM training scheme has been presented. In some cases, for instance when not enough data is available to perform a conventional training, an adaptive training procedure can be conducted. This adaptive training modifies a general HSMM model, trained with sufficient data, to fit a particular style using only a small amount of data from this target style. Training is performed in this article with constrained structural maximum a posteriori linear regression (CSMAPLR) transformation [32,33]. This transformation is called "linear regression" because the calculated transformation of the HSMM parameters can only be linear. The adapted means $\hat{\mu}(\hat{m})$ and variance $\hat{\sum}(\hat{\sigma}^2)$ of the state output (state duration) distributions can be expressed, given the linear transformation A_o (A_d) and the bias b_o (b_d), under the following form:

$$\hat{\mu} = A_o\mu - b_o, \hat{\Sigma} = A_o\Sigma A_o^\top, \tag{5}$$

$$\hat{m} = A_d m - b_d, \hat{\sigma}^2 = A_d\sigma^2 A_d^\top. \tag{6}$$

The term "constrained" refers to the fact that the linear transformations applied to the means and the linear transformation applied to the variances of the average model (both for durations and observation parameters) are required to be the same, other than the bias. A detailed explanation of the CSMAPLR transformation and how it can be calculated can be found in [32] and [33]. This CSMAPLR transformation is implemented within the HTS framework.

The last step of the adaptation training procedure consists in a mximum *a posteriori* (MAP) [13] adaptation that further transforms the models already linearly adapted by CSMAPLR, modifying the estimation of the distributions having enough training samples, as explained in [32].

5.1.4 HSMM synthesis

In HSMM-based synthesis, the synthesis stage consists in an algorithm which directly generates the optimal parameter sequence from the HSMM in the maximum-likelihood sense. In our HSMMs, the probability density function of the observations was modeled by one Gaussian distribution per state. Given a HSMM (λ) and the sequence of steps we want to generate, the HSMM synthesis consists in finding the parameter sequence $O^* = [o_1^\top, o_2^\top, \ldots, o_T^\top]^\top$ with maximum probability given the HSMM model λ. The problem can thus be mathematically expressed as follows:

$$O^* = \arg\max_O P(O|\lambda). \tag{7}$$

Unfortunately, there is no known algorithm to analytically solve this equation. We can thus only find approximated solutions by using the most likely state sequence. Since $P(O|\lambda) = \sum_{all\ q} P(O, q|\lambda)$, where q is one sequence of states from the set of all possible state sequences corresponding to the walk we want to generate, the problem can be approximated by:

$$O^* \simeq \arg\max_O(\max_q P(O|q, \lambda)P(q|\lambda)). \tag{8}$$

The initial problem of finding the optimal sequence of observations O^* given the HSMM λ and the desired sequence of synthesized walk steps can thus be splitted into two optimization problems:

(1) Find the optimal sequence of states q^* given the HSMM λ and the desired sequence of synthesized walk steps:

$$q^* = \arg\max_q P(q|\lambda). \tag{9}$$

(2) Find the optimal sequence of parameters O given the previously determined optimal sequence of states q^* and the HSMM λ:

$$O^* = \arg\max_O P(O|q^*, \lambda). \tag{10}$$

The optimal sequence of states q^* must first be estimated, according to Equation (9). Knowing the state duration densities thanks to the HSMM modeling, the optimal sequence q^* according to Equation (9) can be

determined [29]. Once the optimal state sequence has been calculated, the optimal sequence of parameters can be determined from Equation (10).

When the constraints between static and dynamic features expressed in Equations (2) and (3) are added to the optimization problem, maximizing $P(O \mid q^*, \lambda)$ with respect to O (Equation (10)) becomes equivalent to maximizing it with respect to C:

$$C^* \arg\max_C P(WC|q^*, \lambda), \tag{11}$$

as $O = WC$ (Equation (4)). In the HTS framework and as explained in details in [28], this problem can be solved using the Cholesky decomposition. The algorithm we just described can thus generate a parameter trajectory of static features that maximizes the likelihood of the parameter sequence containing both static and corresponding dynamic parameters given an HSMM model.

However, the generated parameter sequence is often excessively smoothed due to statistical processing. The sharp variations that appear in the motion and account for a great deal of the style variation tend to disappear and the synthesized walks loose a great deal of their naturalness. In [34], Toda and Tokuda present an algorithm to reduce that effect, by taking into account one of the characteristics of the parameter sequence that was removed statistically: the global variance of the data. The global variance ($gv(C)$) of the static features c_t over a time sequence of T frames is calculated by:

$$\bar{c}(\text{dim}) = \frac{1}{54}\sum_{t=1}^{T}\left(c_t(\text{dim})\right), \tag{12}$$

$$v(\text{dim}) = \frac{1}{54}\sum_{t=1}^{T}\left(c_t(\text{dim}) - \bar{c}(\text{dim})\right)^2, \tag{13}$$

$$gv(C) = \left[v(1), v(2), \ldots, v(\text{dim}), \ldots v(54)\right]. \tag{14}$$

The method proposed in [34] and implemented in the HTS framework considers not only the HMM likelihood for the static and dynamic feature vectors, but also the likelihood of the global variance. The probability to maximize in Equation (7) becomes:

$$P(O|\lambda, \lambda_{gv}) = \sum_{\text{all } q} P(O, q|\lambda)^{\omega} P(gv(c)|\lambda_{gv}), \tag{15}$$

with λ_v a single Gaussian distribution representing the global variance of the data $v(c)$ by a mean vector and a covariance matrix, and ω a constant determining the weight between the two likelihoods. Taking into account the global variance of the data enabled us to avoid oversmoothed synthesized walks.

Once our adapted model is built, we can synthesize as many stylistic walk sequences as we want using the same synthesis procedure as described here. The model gives us joint angles and the displacement of the skeleton can be computed using our knowledge of the limb lengths and the step in which we are (which defines which foot is in contact with the ground).

5.2 Results

5.2.1 Neutral walk modeling

For our HMM training and synthesis, we followed the method explained in Section 5.1 and adapted the functions originally implemented for speech within the HMM-based speech synthesis system (HTS) to our procedure. The implementation of the HTS toolkit (version 2.1) that we used in this work is publicly available on the HTS website [11].

The three sequences of "free" walk of the 41 subjects of the eNTERFACE'08 3D database were used to train our average neutral walk model, which consisted of five-states left-to-right HSMM with no skip for both steps (right and left). The database contains 669 observation sequences for "right step" and 656 observation sequences for "left step". We made the contextual distinction between five positions in the walk sequence for each step: the first, second, last, last-but-one steps of the sequence, and all the other steps. The training began thus with ten models to train (five for each step).

During the training phase, some of the ten initial models were automatically tied by the context-based tree clustering and only six HMMs remained for the whole walk modeling in the average model: two models for the first step of a walk sequence, two for steps inside a sequence, and two for the last step of a sequence (one model for the right step and one for the left step each time).

5.2.2 Style walk modeling

Adaptive training is performed with constrained maximum likelihood linear regression (CMLLR) transformation [33] of our previously trained average neutral walk HSMM model. For each one of the 11 expressive walks of our Mockey database, a separate adaptive training was performed using all of the data available for the target style. The number of observation sequences for each of the stylized walks are given in Table 1. So, for each style, we obtained separate contextual (initial, final and "inside a sequence") models for the right and left steps.

5.2.3 Synthesis of new walk sequences

Each new walk sequence is synthesized by first concatenating HMMs corresponding to the desired succession of steps. The whole parameter sequence is then calculated from that complete sequence of models, taking into account the dynamics of the synthesized parameters

thanks to the first and second derivatives of the parameters. Therefore, the smoothness of the transitions between the successive steps of the walk sequence is ensured.

The parameters generated by the model are only the angles between the body segments, hence no overall displacement of the character is synthesized by the HMMs. But using our knowledge of the boundaries of each synthesized step and the height of each foot (for both heel and toe) given by the known angles and length of the limb parts, we can determine which part of which foot is in contact with the ground. Starting from that fixed 3D point, we can compute the overall displacement of the whole body and ensure at the same time that no foot sliding occurs. Figure 4 illustrates two examples of synthesized walks (sad and topmodel walks). The style difference is already visible in these poses, and the duration difference is also illustrated as more poses (and thus more time) are needed to complete the sad walk step than the topmodel walk step.

Our average model was trained with data recorded at a frame rate of 60fps and adapted in the second phase to data captured at a rate of 30 fps, but that difference was not an issue as the durations were adapted automatically during the average-to-style model adaptation. The synthesized walks, coming from models adapted to the Mockey style data, corresponded to a frame rate of 30fps.

6 Qualitative user evaluation

6.1 Methodology

A recurrent problem with motion data synthesis is the difficulty to evaluate the produced motion sequences. Most studies only present their method without giving the reader information about the quality of the results, or just give a link to an example of synthesized motion.

In this article, we propose three different subjective tests that enabled us to assess the quality of the synthesis results. The basic set of the tested videos consisted in 44 walk sequences: one original walk sequence for each of the 11 styles, the same sequences from which the displacement of the root of the skeleton was removed, one sequence of synthesized walk for each of the 11 styles without adding the overall displacement (called "static" in the next sections), and the same synthesized sequences for which the absolute position of the root was calculated as explained in Section 5.1.4 (called "displacement" in the next sections). Two videos of motion synthesized with the average walk model were added (with and without displacement), which makes 46 videos in total. In the video sequences, motion was performed by a basic blue stick-figure character as shown in Figure 2.

Participants accessed to the evaluation tests through a web browser. They had to start the video themselves by clicking on it, and could watch it as many times as they wanted. If they did not complete the test thoroughly, they could come back later, but the participant's results were saved even if the three tests were not completely finished. Video sequences lasted between 3 and 17 s.

About a 100 naive evaluators took part in the evaluation. The three tests and their respective results are presented in Sections 6.2, 6.3, and 6.4. For each of the three tests, every evaluator was presented a set of ten videos or couples of videos. Those videos were randomly picked by the evaluation program, and were thus different for each evaluator.

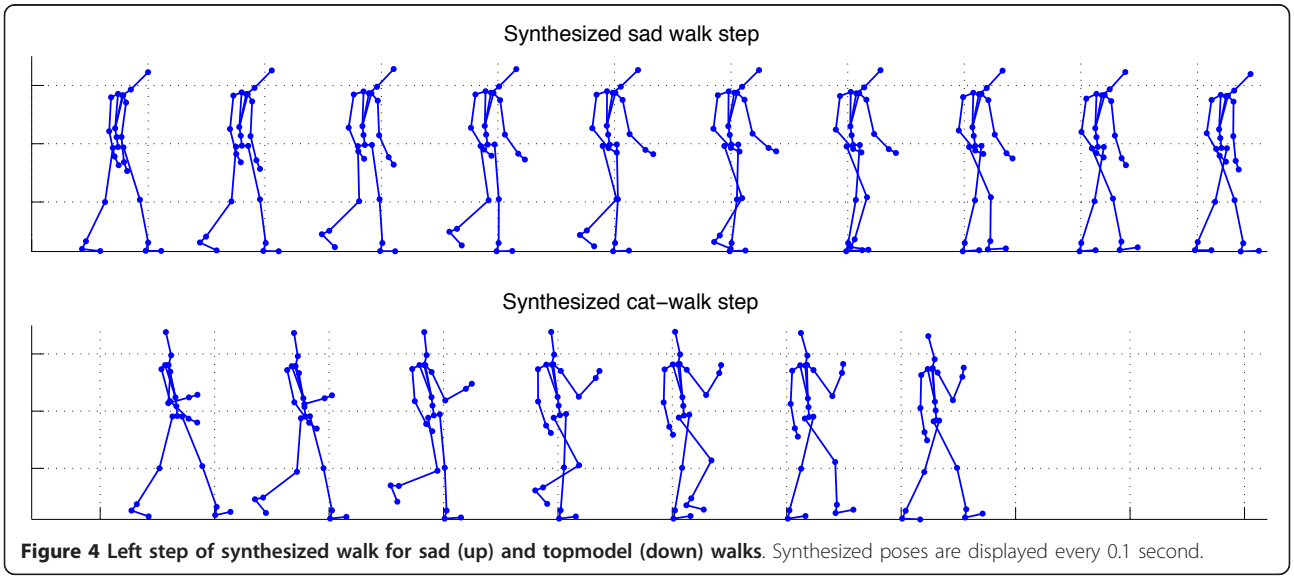

Figure 4 Left step of synthesized walk for sad (up) and topmodel (down) walks. Synthesized poses are displayed every 0.1 second.

6.2 Naturalness evaluation

In the first test, the evaluator was presented one random video at a time. He was asked to choose among three propositions: the stylistic walk in the video seems "real", "synthetic", or "I don't know". The aim of the test was to determine if there was a significant difference in the way the naturalness of the original and the synthesized walks were perceived.

In a first trial, this test was presented to the users in an odd manner and several users reported that they were confused and did not understand the question. The user was asked if the walk was "natural" or "unnatural", which lead most people to perceive nearly all the walks, both original and synthetic, as "unnatural" because of the nature of the data presented: exaggerated walk styles performed by an actor. We reformulated thus the question and only kept the results obtained after that change, which explains why only 500 sequences were evaluated in this first test.

Five-hundred sequences of walk were evaluated in total (246 original sequences and 254 synthetic sequences). The results of the test are presented in Figure 5. 65.45% of the original walks and 50.39% of the synthetic walks were labeled as "real walks", and the user could not decide for 2.44% of the original walks and 6.69% of the synthetic walks. We can thus say that even if the original walks seem a little bit more natural to the evaluators, our synthesized walks looked very natural too, with more than half of the synthetic sequences presented to the evaluators identified as "originals", 15% less than the real original walks. We also verified informally that the degree of unnaturalness between the original and synthesized motions labeled as "unnatural" by the evaluators was not significantly different. This was done a posteriori, by showing five people both original and synthesized videos that had the more often been identified as "unnatural" during our extensive user evaluation, and asking them if some of the videos were significantly less natural than others.

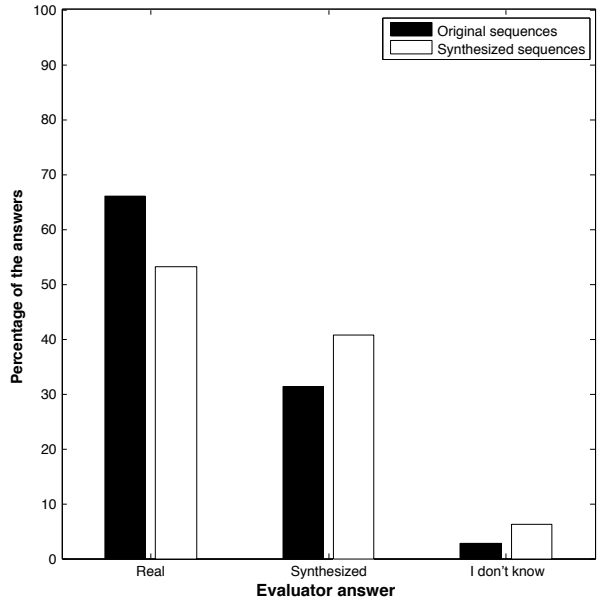

Figure 5 Results of the naturalness test comparing the perception (real, synthetic or "I don't know") of original and synthesized walk sequences.

6.3 Style recognition evaluation

In the second test, the evaluators were again presented one video at a time. They were asked to chose between 13 different style possibilities: the 11 styles, plus "average walk" or "I don't know". A total number of 922 evaluations of videos taken randomly from the set of 46 possible videos were performed.

The recognition rate was of 45.9% for original walks and 38.93% for synthetic walks. Less than half of the styles were properly recognized but this results is easily explained by the fact that no examples of the different possible styles were presented to the users before letting them choose between the 13 proposed answers, and some of the styles were thus subject to the evaluator's interpretation which did not always correspond to the actor's interpretation. Furthermore, some of the styles were very close (for instance "proud" was more often recognized as "cool" or "manly") and were easily confused for one another. The confusion matrix of the classification by the evaluators is presented in Table 2. The confusion matrix shows that when a walk style is wrongly identified as another style, that association is the same for original and synthetic walk and consequently depends more on style interpretation than on the quality of the synthesis. In order to insure that the low style recognition results were not caused by the motion representation (stick figure), we displayed the original motion data on a more realistic 3D body character and asked five subjects to recognize the displayed style in some examples with stick figure and some examples with 3D body character. The 3D body representation did not improve recognition, which comforted us in our analysis that the poor recognition style was due to the variable appreciations of our eleven proposed styles by users and by the actor. Another factor which seemed to influence the results is that in the videos with root displacement, the character displayed was smaller on the screen and the details of the motion were harder to distinguish. Despite these facts, the classification rate is much higher than mere chance: with 13 possible choices for style, a random classification would have given a recognition rate of 7.69%.

The percentages of correctly classified videos for both original/synthetic and with/without root displacement are presented in Table 3. The results correspond to what

Table 2 Confusion matrix of style recognition test for both original walk sequences (first part of the table) and synthesized sequences (second part of the table)

	Evaluators classification (%)												
	Proud	**Decided**	**Sad**	**Topmodel**	**Drunk**	**Cool**	**Afraid**	**Tiptoe**	**Heavy**	**Hurry**	**Manly**	**Average**	**?**
Original (actual style)													
Proud	10	0	3	3	0	27	0	0	3	0	23	27	3
Decided	3	37	3	0	0	5	3	0	3	26	13	5	2
Sad	2	0	68	0	9	2	4.6	0	7	0	2.3	5	0
Topmodel	14	0	0	58	3	11	0	0	0	0	3	11	0
Drunk	0	0	0	0	91	9	0	0	0	0	0	0	0
Cool	9	15	0	4	0	20	0	0	0	0	6	37	6
Afraid	0	0	0	0	6	0	49	36	3	0	0	0	6
Tiptoe	0	0	0	0	0	0	11	71	0	11	0	0	7
Heavy	0	0	15	0	34	15	7	0	17	0	10	0	2
Hurry	0	39	0	10	0	5	0	0	0	19.5	5	15	7
Manly	0	0	6	3	0	19	0	0	11	0	47	6	8
Synthesized (actual style)													
Proud	15	0	0	6	0	23	0	0	6	0	10	37	2
Decided	2	42	2	0	0	0	2	0	9	40	2	0	0
Sad	0	3	71	0	5	0	0	0	13	0	3	3	3
Topmodel	11	0	0	51	17	17	0	0	0	0	2	0	2
Drunk	7	20	2	2	72	5	0	0	2	0	0	0	9
Cool	15	3	0	3	0	38	0	0	0	0	24	12	6
Afraid	0	0	0	0	2	0	44	40	7	0	0	0	7
Tiptoe	0	4	0	8	0	0	15	42	0	15	0	4	12
Heavy	5	2	10	5	7	17	2	0	19	0	19	2	12
Hurry	2	30	0	16	0	5	0	0	0	23	5	18	2
Manly	7	2	11	0	2	11	0	0	9	0	52	2	2

The recognition rate is expressed in percents of the actual style sequences presented to the evaluators, rounded to the unit

could be expected. Original motion were slightly better recognized than synthesized motions. Furthermore, since the displacement of the root enables the evaluator to have a better global view of the scene and of the succession of steps, adding the displacement improves the results for original motions but worsens them for synthesized motions. The values of all style recognition rates for both original and synthetic motions are presented in Figure 6. We can see that for some styles, synthesis worsened the recognition rate (for instance "tiptoe", walk number 8). For others, the style was better recognized in synthesized sequences than in the original ones (for instance the "cool" walk, number 6). The Pearson correlation coefficient between the recognition score of the original motions versus the synthesized motion for each of the 11 styles is 0.8849. This value shows that the recognition rates per style for original motions versus synthetic motions are tightly linked. If one style is well recognized in the original motions, it will be well recognized in the synthesized motions.

Table 3 Percentage of correctly classified walk sequences for the style recognition test

	Original (%)	Synthesized (%)
Static	44.20	40.71
Displacement	47.60	37.15

As Bernhardt and Robinson in [1], we can calculate a more objective measure of the recognition efficiency η by normalizing the achieved recognition rate (or sensitivity) by the recognition rate given by a random classification (sensitivity expected by chance):

$$\eta = \frac{\text{Achieved sensitivity}}{\text{Sensitivity expected by chance}}. \quad (16)$$

The efficiency of the users' recognition is thus equal to η_{orig} = 47.6/7.69 = 6.19 for original walk sequences and to η_{orig} = 37.15/7.69 = 4.83 for synthesized walk sequences. These values are both higher than the human recognition efficiencies cited in [1] (η = 3.72 and η = 3.55 for emotional state classification based on original knocking motions in [35] (four emotions, point-light display) and [36] (five emotions, full video)),

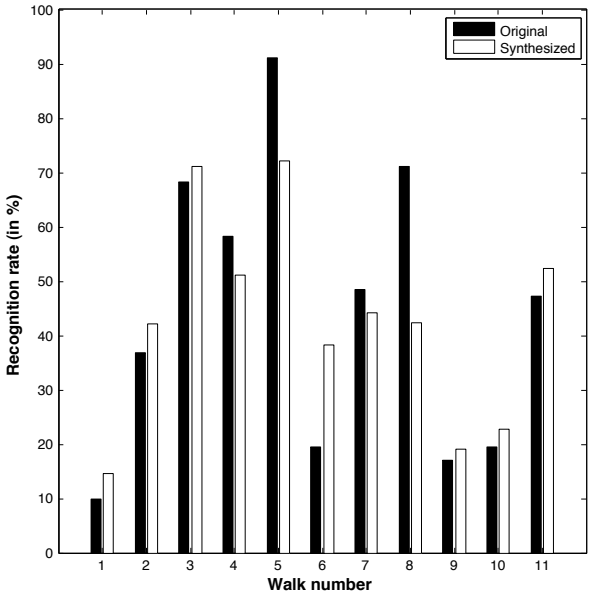

Figure 6 Style recognition rate (in percent) of original versus synthesized walk sequences for the 11 walk styles.

indicating that the style component was accurately perceived in both our original and synthesized sequences.

6.4 Original versus synthesized comparison

In our last test, participants were presented two videos at the same time, the original and synthesized videos corresponding to the same style (either both static or both with displacement). A screenshot of this third test is presented in Figure 7. The order of the videos was randomly determined by the program. Evaluators were asked to choose between five possible qualifications for the level of resemblance between the two videos: "identical", "slightly different", "different", "very different" and "nothing in common". We gave numerical values to these subjective opinions, from four for "identical" to zero for "nothing in common". The best possible score is thus four if all comparisons are found identical and zero if they are all found has having "nothing in common". Eight-hundred and sixty-five comparison tests were performed, leading to a very good global score of 3.15. The detail of the number of answers for each of the five categories is presented in Figure 8. We can see that the most chosen resemblance is "identical" and that the number decreases while the perceived difference increases. These results show that our synthesized walk sequences look very similar to the original training data.

7 Conclusion

Thanks to the method presented in this article, we were able to build HMMs of our 11 different stylized walks, and to use these models to synthesize new walk sequences. Our method produces very convincing synthesized walk sequences where the styles characteristics can be recognized (some examples of synthesized motion sequences can be found in the "Additional file 1" (Video-StylisticGaitSynthesis.mov) or on the author's webpage [37]), even if the walk styles in our stylistic

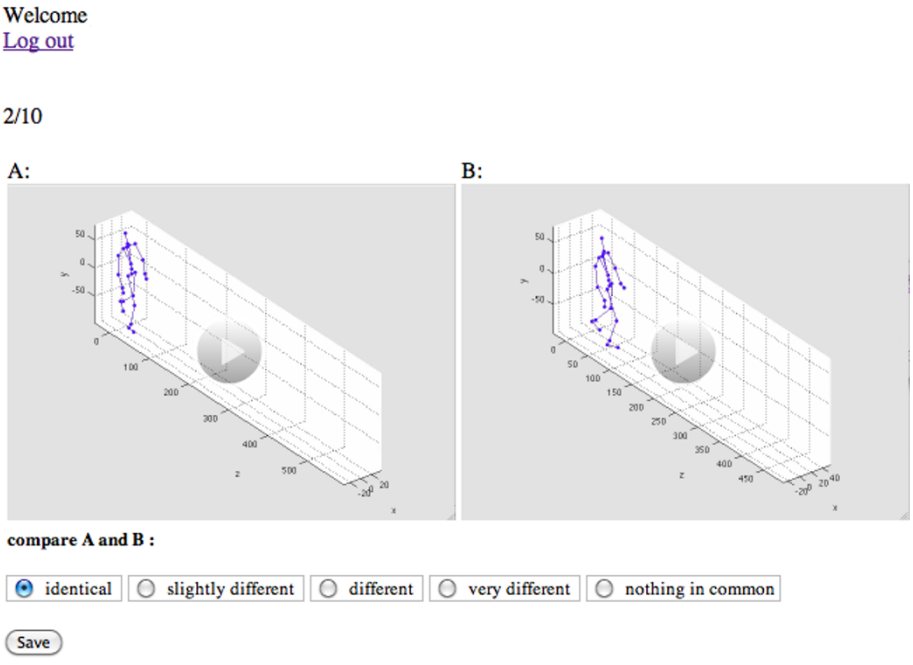

Figure 7 Screenshot of the interface of the third evaluation test: comparison of original versus synthesized walk.

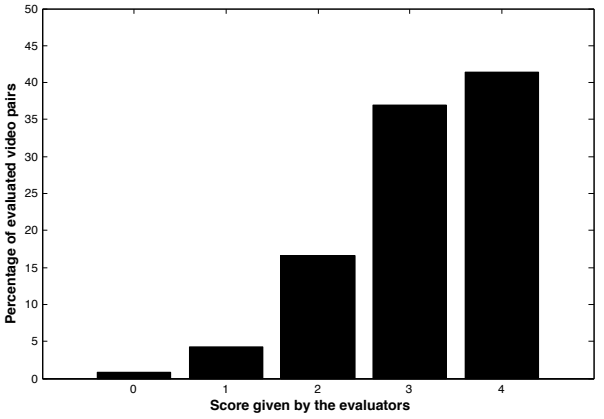

Figure 8 Original versus synthesized comparison test, with resemblance ranging from 0 (nothing in common) to 4 (identical).

database were exaggerated and thus extremely different from each other unlike most motion style studies which concentrate on smaller variations.

We also proposed a setup for a subjective evaluation of the synthesis results, which showed that the synthesized walks were close to the original training data and also pointed out some of the weaknesses of the synthesis, indicating directions for future work. The recognition test showed for instance that adding the displacement to the motion improved the recognition rate for original motions but had the opposite effect on synthesized sequences. We think that this is due to the inter-step variation which is lower in the synthesized sequences than in the original motion, and that should be further improved.

Future work will include further analyses of the evaluation tests that can be used to assess the naturalness of the produced motions, and analysis of the use of the style interpolation/extrapolation using the trained models. One could also study how several parameters influence the perceived results, like the variables of the HMM (number of states for instance), the influence of the number of stylistic steps in the adaptation training phase, the way the results are presented to the user (skinned virtual character versus stick figure), how a reduction of the dimensionally of the original data influences the quality of the results, etc. The adaptation method presented here could also be used to analyze and synthesize walks for different human characteristics that influence the walk style, like gender (male vs. female walk) or age (children vs. elderly or others).

Additional material

Additional file 1: Video-StylisticGaitSynthesis.mov (quicktime movie). This short video present some examples of the stylistic walk sequences that were synthesized in this work and presented to the participants of the user assessment tests.

Acknowledgements
This project was partly funded by the Ministry of Région Wallonne under the Numediart research program (grant N0716631). Joëlle Tilmanne was supported by the "Fonds pour la formation à la recherche dans l'industrie et l'agriculture" (FRIA) during part of this work. The authors would like to thank the comedian Sebastien Marchetti and Thierry Ravet for their participation in the motion capture database recording.

Competing interests
The authors declare that they have no competing interests.

Received: 15 April 2011 Accepted: 26 March 2012
Published: 26 March 2012

References

1. D Bernhardt, P Robinson, Detecting Affect from Non-stylised Body Motions, in *Affective Computing and Intelligent Interaction*. Lecture Notes in Computer Science, vol. 4738 ed. by A Paiva, R Picard (Springer, Berlin, 2007), pp. 59–70
2. A Menache, *Understanding Motion Capture for Computer Animation and Video Games* (Morgan Kauffman Publishers Inc., San Francisco, 2000)
3. D Mena, J Mansour, S Simon, Analysis and synthesis of human swing leg motion during gait and its clinical applications. J Biomech. **14**(12), 823–832 (1981). doi:10.1016/0021-9290(81)90010-5
4. T Pejsa, I Pandzic, State of the art in example-based motion synthesis for virtual characters in interactive applications. Comput Graph Forum. **29**, 202–226 (2010). doi:10.1111/j.1467-8659.2009.01591.x
5. D Forsyth, O Arikan, L Ikemoto, J O'Brien, D Ramanan, Computational studies of human motion: part 1, tracking and motion synthesis. Found Trends Comput Graph Vis. **1**(2–3), 77–254 (2005)
6. W Geng, YG, Reuse of motion capture data in animation: a review. Comput Sci Appl (ICCSA). **2003**, 620–629 (2003)
7. S Calinon, F Guenter, A Billard, On learning, representing, and generalizing a task in a humanoid robot. IEEE Trans Syst Man Cybern B. **37**(2), 286–298 (2007)
8. P Glardon, R Boulic, D Thalmann, PCA-based walking engine using motion capture data. IEEE Comput Graph Int. **2004**, 292–298 (2004)
9. M Brand, A Hertzmann, Style machines, in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. (ACM Press/ Addison-Wesley Publishing Co., New York, 2000), pp. 183–192
10. J Tilmanne, T Dutoit, Expressive gait synthesis using PCA and Gaussian modeling, in *Proceedings of the Third international conference on Motion in games*. MLG'10 (Springer, Berlin, Heidelberg, 2010), pp. 363–374
11. HTS Working Group, The HMM-based speech synthesis system (HTS) Version 2.1. http://hts.sp.nitech.ac.jp/. Accessed 2010
12. L Rabiner, A tutorial on hidden markov models and selected applications in speech recognition. Proc IEEE. **77**, 257–286 (1989). doi:10.1109/5.18626
13. J Yamagishi, T Nose, H Zen, Z Ling, T Toda, K Tokuda, S King, S Renals, Robust speaker-adaptive HMM-based text-to-speech synthesis. IEEE Trans Audio Speech Lang Process. **17**(6), 1208–1230 (2009)
14. B Picart, T Drugman, T Dutoit, Analysis and synthesis of hypo and hyperarticulated speech, in *Proceedings of the Speech Synthesis Workshop 7 (SSW7)*. NICT/ATR, Kyoto, Japan), 270–275 (Sept 2010)
15. NF Troje, Retrieving Information from Human Movement Patterns, in *Understanding Events: How Humans See, Represent, and Act on Events*, vol. 1. ed. by TF Shipley, JM Zacks (Oxford University Press, Oxford, 2008), pp. 308–334
16. LM Tanco, A Hilton, Realistic synthesis of novel human movements from a database of motion capture examples, in *Proc of the Workshop on Human Motion (HUMO'00)*. IEEE Computer Society, Washington, DC, USA, 137 (2000)
17. Y Wang, Z Liu, L Zhou, Automatic 3D motion synthesis with time-striding hidden Markov model, in *Proc International Conference on Machine Learning and Cybernetics (ICMLC'05)*, vol. 3930. (SB Heidelberg, Guangzhou, Aug, 2005), pp. 558–567
18. Y Li, T Wang, H Shum, Motion texture: a two-level statistical model for character motion synthesis, in *Proc of SIGGRAPH'02*. (ACM Press, New York, 2002), pp. 465–472
19. D Ramanan, DA Forsyth, Motion Analysis by Synthesis: Automatically Annotating Activities in Video. (2005), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.94.2069

20. Y Wang, L Xie, Z Liu, L Zhou, The SOMN-HMM model and its application to automatic synthesis of 3D character animation, in *IEEE Conference on Systems, Man, and Cybernetics*. Taipei, Taiwan, 4948–4952 (2006)
21. Y Wang, Z Liu, L Zhou, Learning style-directed dynamics of human motion for automatic motion synthesis, in *IEEE Conference on Systems, Man's and Cybernetics*. Taipei, Taiwan, 4428–4433 (2006)
22. T Yamazaki, N Niwase, J Yamagishi, T Kobayashi, HumanWalking motion synthesis based on multiple regression hidden semi-Markov model, in *2005 International Conference on Cyberworlds (CW'05)*. IEEE Computer Society, Washington DC, 445–452 (2005)
23. IGS-190, Animazoo website, http://www.animazoo.com
24. J Tilmanne, R Sebbe, T Dutoit, A database for stylistic human gait modeling and synthesis, in *Proceedings of the eNTER-FACE'08 Workshop on Multimodal Interfaces*. Paris, France, 91–94 (2008)
25. R R Parent, Technical background, *Computer Animation Complete: Part I: Introduction to Computer Animation* Morgan Kaufmann, Emeryville, 60–68 (2009)
26. F Grassia, Practical parameterization of rotations using the exponential map. J Graph Tools. **3**, 29–48 (1998). doi:10.1080/10867651.1998.10487493
27. MP Johnson, Exploiting quaternions to support expressive interactive character motion. PhD thesis, (Massachusetts Institute of Technology, 2002)
28. K Tokuda, T Yoshimura, T Masuko, T Kobayashi, T Kitamura, Speech parameter generation algorithms for HMM-based speech synthesis, in *Proc ICASSP*. Istanbul, Turkey, 1315–1318 (June 2000)
29. T Yoshimura, K Tokuda, T Masuko, T Kobayashi, T Kitamura, Duration modeling for HMM-based speech synthesis, in *Fifth International Conference on Spoken Language Processing (ICSLP)*. Sydney, 29–32 (1998)
30. S Young, G Evermann, M Gales, T Hain, D Kershaw, X Liu, G Moore, J Odell, D Ollason, D Povey, V Valtchev, P Woodland, *The HTK Book, Version 3.4*. (Entropic Cambridge Research Laboratory, Cambridge, 2009)
31. J Yamagishi, T Kobayashi, Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. IEICE Trans Inf Syst. **90**(2), 533–543 (2007)
32. J Yamagishi, T Kobayashi, Y Nakano, K Ogata, J Isogai, Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans Audio Speech Lang Process. **17**, 66–83 (2009)
33. M Gales, Maximum likelihood linear transformations for HMM-based speech recognition. Comput Speech Lang. **12**(2), 75–98 (1998). doi:10.1006/csla.1998.0043
34. T Toda, K Tokuda, A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Trans Inf Syst. **90**(5), 816–824 (2007)
35. A Kapur, A Kapur, N Virji-Babul, G Tzanetakis, P Driessen, Gesture-Based Affective Computing on Motion Capture Data, in *Affective Computing and Intelligent Interaction*, vol. 3784. Lecture Notes in Computer Science (Springer, Berlin/Heidelberg, 2005), pp. 1–7. doi:10.1007/11573548_1
36. FE Pollick, HM Paterson, A Bruderlin, AJ Sanford, Perceiving affect from arm movement. Cognition. **82**(2), B51–B61 (2001). doi:10.1016/S0010-0277(01)00147-0
37. Joelle Tilmanne's webpage, http://tcts.fpms.ac.be/~tilmanne/

doi:10.1186/1687-6180-2012-72
Cite this article as: Tilmanne *et al.*: **Stylistic gait synthesis based on hidden Markov models.** *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:72.