Using Language Models to Generate Whole-Body Multi-Contact Motions

Christian Mandery, Júlia Borràs, Mirjam Jöchner, and Tamim Asfour

***Abstract*— We present a novel approach for generating sequences of whole-body poses with multi-contacts for humanoid robots, which is inspired by techniques from natural language processing. To this end, we propose a probabilistic n-gram language model learned from observation of human locomotion tasks. Human motion data is automatically segmented according to detected contacts of the body with the environment to provide support, that is, *support poses*, which are further subdivided with regard to whole-body configuration. These poses are subsequently used to train a language model, whose words are the poses, and whose sentences represent sequences of poses. Then, we propose a planning algorithm that, given the constraints imposed by a task, finds the sequence of transitions with the highest probability according to our language model. We have applied our approach to 140 motion capture recordings of locomotion tasks that involve using one or both hands for support. The evaluation demonstrates that our approach is able to generate complex sets of pose transitions, and shows promising results regarding its application to more complex tasks.**

I. Introduction

Whole-body motion planning with multi-contacts for humanoid robots in unstructured environments constitutes an open problem of vital interest for the humanoid robotics community. In this work, we propose a data-driven approach which uses human motion data for the autonomous generation of sequences of whole-body pose transitions for locomotion tasks that use the environment to enhance balance. Given a motion, we automatically detect whole-body *support poses* defined by the body parts used to provide support [1], which are then further subdivided into different *shape poses*. By analyzing large sets of motions showing humans executing locomotion tasks, we can train a language model, whose words are whole-body poses and whose sentences are sequences of these poses that characterize a motion. Using this language model, we can then plan a sequence of whole-body pose transitions, which satisfies the constraints of a given locomotion task. We formulate such a locomotion task as the displacement of a certain distance and the availability of environmental elements that can provide *support and lean affordances*. We assume that such environmental knowledge can be provided by visual perception following our previous works in [2], [3], still under development. Fig. 1 shows an example of a sequence of pose transitions obtained with the proposed approach.

The research leading to these results has received funding from the European Union Seventh Framework Programme under grant agreement no 611832 (WALK-MAN) and grant agreement no 611909 (KoroiBot).

The authors are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany.
{mandery, julia.borrassol, asfour}@kit.edu

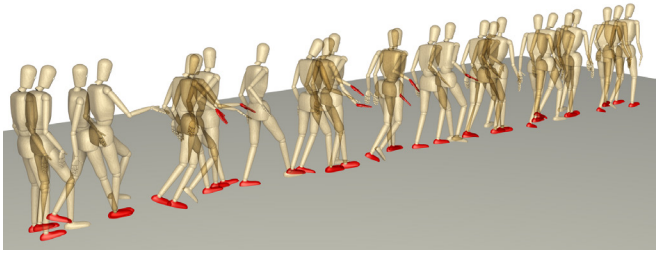

Fig. 1. Visualization of the planned sequence of whole-body shape pose transitions for a straight motion of 6 meters that allows for hand supports from meters 2 to 4 (note that we do not plot the support surface here). In red, we highlight the body segments that are in contact at each pose. Each shape pose is drawn at a distance according to the associated translation of the center of mass.

The problem of planning whole-body motions with multi-contacts has been addressed in the robotics literature by many authors. It is a challenging problem due to the complexity of the kinematic chains, the dynamic constraints, and the multi-dimensionality of the tasks. Some solutions exist that, despite being computationally expensive, are successful in solving the full problem including all the dynamic equations under contact constraints, such as [4], [5], [6]. Other approaches divide the problem into two independent subproblems that can be solved separately. This division consists of first finding a sequence of discrete sets of contacts, called *stances* [7], [8], [9], and second, finding the continuous motion that links the stances [10]. Dividing the problem allows to reduce its complexity, without considerably reducing the quality of the result.

We approach the problem from a different point of view, but still suggest a division of the problem into two steps. Our motivation relies on the fact that humans do not plan specific contact locations in an early stage of the motion execution. The first step of our approach consists of finding sequences of whole-body poses. For each of the transitions between consecutive poses, an associated dynamic movement primitive (DMP) can be learned directly from human motion capture data [11]. Then, in a second step, we will adapt the transition DMPs to satisfy the specific constraints given by the initial and final whole-body poses, the contact constraints, and ultimately the dynamics. The work presented in this paper deals only with the first part of the problem: the planning of pose sequences. Examples in the literature show that the second part can be solved using DMPs as demonstrated in [12].

The idea of sequencing movement primitives (MPs) for

multi-contact motion planning was proposed earlier in the literature. MPs were used to guide the choice of contact points and correct the output motions generated by the algorithm in [13] to appear more human-like. Later works used sequences of MPs to form a continuous motion for climbing a ladder [12]. However, these works focus on the second part of the problem, which is obtaining the continuous motion in joint space that satisfies a set of given constraints, whereas the given set of MP transitions between support poses is specified by hand. In contrast, in our work, we focus on the autonomous generation of stance/pose sequences. As stated in [14], this is considered one of the main challenges in this area of research.

Still, the idea of autonomously sequencing MPs is not new. In this sense, our work is closely related to works by Kulić et al. [15], [16], [17], but with several differences. In [15], the authors build a graph of transitions between MPs directly from segmented human motion data and use this graph to find paths which generate continuous motions. Our work differs from this in several points. First, in our case, the data is segmented according to support poses, while the segmentation introduced in [16], [17] is based on visually recognizable discrete segments of movements. Segmenting using support poses allows us to provide a semantic interpretation of each motion segment that is in accordance with the previously mentioned works on multi-contact motion planning. The graph of transitions built in [15] is similar to the taxonomy of whole-body pose transitions that we presented in previous work [18], [1], with the exception that we focus on transitions between support poses. Another important difference is that instead of planning paths on the graph, our segmented motion data is used to train a language model that allows us to derive a probabilistic model of our data, which can be used for generating new sequences of whole-body poses. It is important to note that for locomotion tasks, the shortest path in our pose taxonomy graph from [18] would not constitute a valid set of pose transitions, since the graph does neither encode the cycles of steps, nor translation information.

Our work can also be related to other high-level planning works in robotics [19], [20], [21]. These works usually build a model based on building worlds with a set of rules. The planning problem is then formulated as finding a sequence of actions from an initial state to a goal state following the stated rules, and the search is guided by optimizing costs and heuristics. Additionally, our work can be related to other applications of linguistic methods in the computer vision and robotics communities. Grammar-based approaches are used in [22] and [23] to recognize and understand human actions. In [24], context-free grammars are used for the representation and verification of robot control policies.

One problem with the mentioned rule-based systems, e.g. STRIPS planning or grammar-based approaches, is that learning the rules gets more challenging with increasing complexity and uncertainty of the environment. Therefore, in fields where natural language is processed, such as automatic speech recognition or machine translation, rule-based and grammar-based systems have been widely replaced by entirely statistical approaches, such as n-gram language models [25]. Techniques from statistical language modeling have also provided solutions to other problems beside natural language [26], [27]. With this work, we hence investigate whether the problem of motion planning for humanoids can also benefit from using statistical language models. To our best knowledge, the use of statistical language models for the sequencing of whole-body motion primitives is novel.

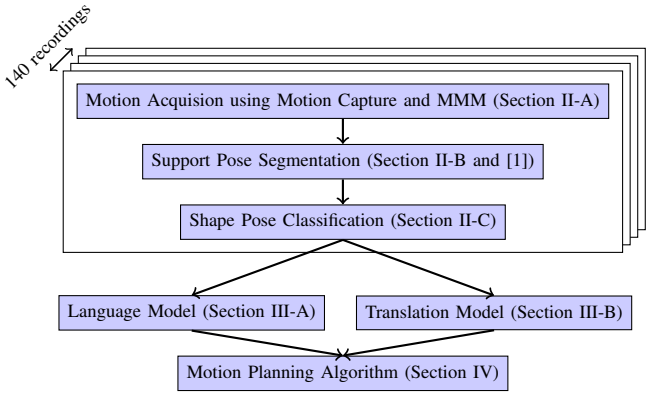

Fig. 2. Relationship between components described in this paper.

This paper is organized as follows. Section II describes our approach for collecting motion data, segmenting it according to whole-body support poses, and classifying these poses with regard to body configuration. In Section III, we explain how we learn n-gram language models and translation models from the motion data. Our planning algorithm is then described in Section IV. Fig. 2 provides an overview of how the motion acquisition and processing, the learned models, and the planning build on each other. Section V presents the results of our approach for three exemplary tasks. Finally, we give conclusions and point out directions of future work in Section VI.

II. Motion Representation

A. Motion Acquisition and Data Set Description

Our motion data has been acquired with optical motion capture by using a Vicon MX10 system equipped with ten T10 cameras running at 100 Hz. Motion recordings are based on the KIT reference marker set, which consists of 56 passive (reflective) markers placed at characteristic anatomical landmarks of the human body. More information about the marker set and the procedures used for motion capture is available in [28] and online[1]. All captured motions are post-processed using the Master Motor Map (MMM) framework described in [28], [29], [30]. The MMM framework provides an open-source framework for the unified representation of human motion, including a reference model of the human body based on biomechanics literature. During reconstruction, the marker-based motion data are mapped to a motion

[1] https://motion-database.humanoids.kit.edu/marker_set/

Fig. 3. Experimental setup for motion capture. Left: Walking with a handrail. Right: Walking on a beam using supports with both hands (on a table and on a handrail).

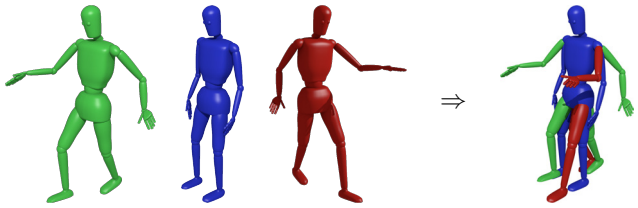

Fig. 4. Exemplary configuration options considered for hands and feet.

of the MMM reference model, providing information about the joint angles, the 6D root pose of the model, and the task space location of every segment of the model. In this work, we are using 40 degrees of freedom for the model kinematics in body torso, extremities, and head while ignoring finger joints and eyes.

In total, we have recorded 140 motions for our evaluation, which are now freely available from the KIT Whole-Body Human Motion Database [28], [31]. Our recordings consist of 20 trials of seven different walking tasks using supports, all demonstrated by the same subject. For each task, 10 trials were walking in one direction, and 10 in the opposite. Hence, the data set is symmetrical concerning left/right hand supports. The tasks include a normal walking task (without hand supports), walking using supports from a handrail or table on one side, walking on a beam using such supports, and finally walking with or without the beam using supports from the handrail and the table on both sides. Fig. 3 shows the experimental setup for the recording of two of these tasks.

B. Motion Segmentation using Whole-Body Supports

In our previous work [1], we have presented a method to segment human motion data according to the body segments that are determined to be in contact with the environment to provide support. One of the features of our MMM framework described above is that it can utilize additional motion capture markers attached to objects and environmental elements, e.g. surfaces of support, to include information about these elements in the motion. This eases the detection of all the support contacts as a combination of distance and velocities of the contact points (see [1] for details). From the set of supporting body segments used by the human subject at a certain time, the respective support pose from our taxonomy of whole-body support poses [18] is then determined and each motion is represented as a succession of transitions between whole-body support poses, which we can visualize as a subgraph of the taxonomy introduced in [18].

C. Shape Pose Classification

In this work, we have added a further differentiation of the support poses based on whole-body configuration. As pointed out in the conclusions of our previous work [1], a support pose can occur in many different body shapes, depending on the task being executed. In order to obtain a better classification of the transitions between poses, we are therefore further subdividing each support pose into several *shape poses* with regard to the configuration of the body during the transition motion.

To explain the concept of shape poses, let us consider a simple walking motion. In such a walking motion, the double foot support pose has three possible shape poses: both feet in parallel when the human is standing (blue pose in Fig. 4), and left/right foot in front of the right/left, respectively. The single foot support poses occur only in one possible shape pose in this motion, with the foot which is not in contact with the floor swinging next to the supporting foot. Defining transitions between these shape poses provides a good classification of motion primitives for such simple walking motions. The shape classification becomes more complex when more contacts are considered, e.g. when the hands are used to reach for a support on the anterior left/right side of the body or to leave a support on the posterior left/right side of the body. To assemble the shape poses, we are considering all combinations of different configuration options for hands and feet, with the options of placing each extremity in front of, near, or behind the body (along the anteroposterior axis), which is exemplified in Fig. 4. For each of these combinations, we have manually selected a representative pose from our data set as the shape pose and saved it in the form of its 40-dimensional representation in joint space. In future work, we are planning to replace this with a data-driven approach for the automatic extraction of useful shape poses from motion data.

For the classification of poses from motion data into shape poses, we found that direct comparison of poses in joint space did not perform well. Therefore, we have defined a metric in task space where each shape pose is represented as a 12-dimensional vector corresponding to the 3D coordinates of the four end-effectors (feet and hands) in the local coordinate system of the model. For each support pose detected as described in Section II-B and [1], we are considering the body configuration at the middle point between when the current support pose is first detected and when the next support pose is detected. Then, we compare this body configuration to the available shape poses for the given support pose and choose the closest shape pose according to the Euclidean distance in the 12-dimensional space. As a result, motions are represented as sequences of shape poses that contain information not only on what body segments are used to provide support, but also indicate a whole-body configuration that will later help to adapt the associated DMPs to the required environmental shape.

TABLE I

NUMBER OF OCCURRENCES AND NUMBER OF DIFFERENT SHAPE POSES FOR EACH SUPPORT POSE FOUND IN OUR DATA SET

Support Pose Name		# Occurrences	# Different Shapes Poses
1Foot	RF	275	16
	LF	384	15
2Feet	-	662	14
1Foot-1Hand	LF-LH	170	6
	LF-RH	130	6
	RF-LH	123	7
	RF-RH	191	9
2Feet-1Hand	LF-RF-LH	244	6
	LF-RF-RH	247	8
1Feet-2Hands	LF-LH-RH	111	8
	RF-LH-RH	89	8
2Feet-2Hands	-	187	8
	Total	**2813**	**111**

In summary, the segmentation of the motions in our data set leads to 2813 detected support poses, which were either single foot, double foot, 1Foot-1Hand, 2Feet-1Hand, 1Feet-2Hands, or 2Feet-2Hands support poses. Table I shows the number of occurrences for these support poses, together with the number of representative shape poses determined from our data set for each of them. Note that not all the possible shape poses (as combinations of hands/feet configurations) appear in our data. In total, our data set contains 111 different shape poses and 442 different shape pose transitions (not given in Table I), defined as a combination of a start and an end shape pose. However, most of these transitions appear only a few times and thus, we will need to collect more data in the future to investigate whether some of these transitions should be ignored. A total number of 74 transitions are used at least ten times in our data set.

III. POSE TRANSITION MODELS

A. Learning the Language Model

Given the representation of human motion explained in Section II, we describe a motion as a sequence of transitions between whole-body shape poses. We propose to approach this problem from a language processing point of view, where one *word* represents a shape pose, and one *sentence* represents a sequence of shape poses, i.e. a motion. In this way, motions can be represented by a language model.

To learn this language model, we take a statistical approach and use an n-gram language model that can be estimated from a textual representation of the motions by using the SRI Language Modeling (SRILM) Toolkit [32]. Such an n-gram language model describes the conditional probability $P\left(w_n | \left(w_1, ..., w_{n-1}\right)\right)$ of observing a certain word w_n given the history of the previous $n-1$ words $(w_1, ..., w_{n-1})$. By multiplying the conditional probabilities of all words in an arbitrary sentence, the probability of observing this sentence can be estimated.

The maximum length of the n-grams considered by an n-gram language model is represented by its order n, where a language model of order n considers the last $n-1$ words to determine possible successor words. Since the amount of necessary training data grows exponentially with the order of the language model, sparsity of data can become a problem. That means, n-grams that are considered as a perfectly valid sequence of words still may not be observed in the training corpus, which is especially true given the sparsity of our data as explained in Section II-C. In natural language processing, this is commonly countered by a technique called *smoothing*, where the probability of *seen* n-grams is slightly reduced and redistributed to allow n-grams not seen in the training corpus to have a non-zero probability. A large number of methods exist to perform such smoothing, with three popular methods being considered by us: Good-Turing smoothing, modified Kneser-Ney smoothing, and Witten-Bell smoothing [33].

TABLE II

N-GRAM LANGUAGE MODEL PERPLEXITIES DETERMINED USING 5-FOLD CROSS VALIDATION FOR DIFFERENT ORDERS (VALUES OF N) AND DIFFERENT SMOOTHING METHODS

Order	Good-Turing	Kneser-Ney	Witten-Bell
2	4.7125	4.7108	4.4668
3	4.1451	4.6747	3.9049
4	4.1759	5.1635	3.7684
5	4.2496	5.7179	**3.7493**
6	4.3471	6.2004	3.7511

For the parametrization of our language model, we perform a grid search across all combinations of language model orders $n \in \{2, 3, 4, 5, 6\}$ and the three aforementioned smoothing techniques. We determine the best combination by using a 5-fold cross-validation, where in each of five rounds, the language model is trained using four fifths of our available data. Then, using the remaining fifth as test data, we compute the perplexity, which is a statistical measure for how well a probabilistic model is able to predict a given input. Table II shows the results of this grid search, with the combined perplexities from all five rounds given and the best value printed bold. As it can be seen, the 5-gram language model using Witten-Bell smoothing exhibits the lowest perplexity and is therefore the type of language model which we consider for the rest of this work.

B. Learning the Spatial Translation for Pose Transitions

In addition to the language model, which describes the probability of certain sequences of shape poses, we want to associate a spatial translation of the whole-body center of mass (CoM) with each shape pose transition and learn these translations also from our motion data. The translation of the CoM can be calculated for each transition by computing the norm of the vector connecting the CoM at the origin pose to the CoM at the destination pose. This is valid because all motions in our data set are in a straight line, and therefore, the spatial translation can be described by a scalar value. However, in the future, we will consider the full 3D direction of motion to allow for movements in a bend or with direction changes. To determine the CoM translation associated with a certain shape pose transition defined as a combination of start and end shape pose, we are taking the mean translation from all occurrences of this transition.

Algorithm 1 Pose Sequence Planning Algorithm

```
activePaths ← heap()
insert Path(startPose) into activePaths
i ← 1
loop
    bestPath ← path with max. score in activePaths
    if (i mod prunePeriod) = 0 then
        pruneDist ← bestPath.distance − pruneThresh
        newPaths ← heap()
        for all path ∈ activePaths do
            if path.distance ≥ pruneDist then
                insert path into newPaths
            end if
        end for
        activePaths ← newPaths
    end if
    if activePath.distance ≥ distance and
  bestPath.endPose = endPose then        // Solution found?
        return bestPath
    else                                 // bestPath is not a solution
        expandedPaths ← expand-path(bestPath)
        for all path ∈ expandedPaths do
            score path using language model and penalty
            insert path into activePaths
        end for
    end if
    i ← i + 1
end loop
```

The determined translation values serve as an indication for the extent of locomotion associated with a transition and are used by the planner to estimate the distance towards the target position covered by a sequence of shape pose transitions. It should be noted though, that since many transitions occur only a few times in the data, the estimated translation should be considered only as an estimate. For the future work of planning transition trajectories using DMPs, these translations are not used and we can use the actual CoM translations associated with the DMPs instead.

IV. Planning Algorithm

Given the language model and the CoM translations associated with shape pose transitions, the process of planning a sequence of whole-body shape poses to satisfy a given locomotion task presents itself as a tree search problem with side constraints. To solve this problem, we assume knowledge about the environment that can be autonomously extracted from visual perception. For instance, approaches like [2] and [3] provide geometric primitives extracted from point clouds with associated affordances for supports with hands or feet. Such geometric primitives contain information not only on the shape, but also on the location of the surfaces for support with respect to the robot camera.

For this work, we assume that we always walk in a straight line and that the robot is located aligned with the considered environmental support objects. We formulate the planning task by specifying sets of allowed body segments, i.e. left/right hand or foot, to provide support contacts for different intervals of distance that depend on the distance from the robot to the support surfaces. For example, we may start a motion with normal walking (only feet supports allowed) until the first support surface, e.g. a handrail, is reached. From that distance until the end of the support surface is reached, hand supports for the left/right hand are also allowed, depending on which side the support surface is located with respect to the robot.

Formally, we want to find the sequence of words $W = (w_1, ..., w_n)$, of any length n, in the space of all possible sentences that solves

$$\arg\max_{W} \big(P(W) - \text{penalty}(W)\big) \tag{1}$$

where P denotes the decimal logarithm of the probability of sequence W according to the language model. Since every word w_i represents a shape pose, the vocabulary size, i.e. the number of valid words, is rather small compared to natural language (111 in our case, see Table I). The penalty function is explained below and serves to bias the planner towards using poses with a higher number of contacts, as these poses are associated with an increased level of stability and robustness. The optimization is subject to the following constraints:

1) The translations associated with the used shape pose transitions must allow to cover the distance from the start to the target position:
$$\sum_{i=1}^{n-1} \text{translation}(w_i \rightarrow w_{i+1}) \geq |\mathbf{p}_{end} - \mathbf{p}_{start}|$$
2) Each shape pose represented by the words $w_1, ..., w_n$ must respect the contact restrictions. That is, the shape pose represented by w_j, located at
$$\sum_{i=1}^{j-1} \text{translation}(w_i \rightarrow w_{i+1}),$$
may contain only support contacts using the body segments allowed for the distance where it is located.
3) Any subsequence of W with the same body segment used as a continuous support contact must not cover a distance longer than a given maximum. Kinematically, a fixed contact allows only for a certain amount of translation before it has to be released to be able to move forward. Since our approach to high-level planning does not explicitly model kinematic constraints, this constraint is necessary to ensure that the same contact is not continuously maintained for too many transitions, e.g. continuously holding a handrail while moving forward two meters.
4) The start and the end pose of the motion, represented by w_1 and w_n respectively, must satisfy the given task. In this paper, we are always assuming the neutral double foot pose, shown in blue in Fig. 4.

The planning algorithm for expression (1) is shown as pseudocode in Algorithm 1. We use an informed breadth-first search with pruning to search the tree spanned by shape pose transitions. Each path through this tree corresponds to a possible sequence of words, i.e. shape poses. This approach

somehow resembles the beam search used in systems for automatic speech recognition or statistical machine translation. In our algorithm, *activePaths* is used to track the currently considered paths to all tree nodes that have not been expanded yet. We use a heap data structure, which allows us to find the active path with the highest score in $\mathcal{O}(1)$ (line 5). The score of a path is composed of the path probability determined by the language model and a penalty value, as shown in expression (1). To compute the penalty of a path, we iterate over its words and accumulate the penalty according to *unused support contacts* in each of the poses represented by the words, i.e. body segments that are not used to provide support in a certain pose, although our environmental knowledge would allow their use at the respective position. In each iteration of the algorithm, the best active path (*bestPath*) is retrieved from *activePaths* and expanded (line 19), resulting in the creation of new possible paths, which are formed by adding to the best path all words that represent allowed shape poses regarding constraint 2. These new paths are then scored and added to the heap (lines 21 and 22). As breadth-first search in general has an exponential time complexity with respect to the length of the paths, we employ pruning techniques (lines 6 to 15) that are run at regular intervals, determined by *prunePeriod*. The pruning considers the translation associated with paths and discards all paths that fall behind the current best path by more than a given threshold (*pruneThresh*). The algorithm terminates if a valid solution for the planning problem has been found (line 16 and 17). Since $P(W)$ can only decrease and penalty(W) can only increase when a path W is expanded by adding another shape pose, the score of *bestPath* is monotonically decreasing with each iteration. Therefore, it can be ruled out that better solutions could still be found when the algorithm terminates.

V. Results

To evaluate the ability of our approach to generate realistic pose transitions, we have designed three exemplary scenarios with different requirements. For reasons of simplicity, we are using the morphology of the MMM reference model for this evaluation. Of course, the kinematic of a humanoid robot could be used as well, as long as it possesses an anthropomorphic morphology, and thus, shape poses learned from the human can be used. These shape poses could then be transferred to the robot morphology by using the retargeting procedure provided by the MMM framework [28]. Also, the robot size should be considered in two ways. First, the capabilities of the robot are implicitly considered in the preceding determination of possible support contacts [2]. Second, the CoM translations associated with shape pose transitions should be scaled proportionally if the height of the robot deviates from the height of the human from which these CoM translations have been learned.

The first task consists of walking a total of 6 meters, allowing right hand contacts from meters 1 to 3. The sequence of pose transitions that is generated by our planner is shown in the upper part of Fig. 5 together with a 3D visualization of the scenario. In red, we highlight the body segments that are used as a support contact. Since each pose transition has an associated CoM translation, we show each shape pose at the corresponding accumulated distance. However, as we mentioned before, this is just an indicative location that helps our planner to decide how many steps are needed to reach the required distance. For task 1, the total distance covered by the CoM translations for the planned transitions is exactly 6.0m. We can see that most of the shape poses at the origin of a transition anticipate hand contacts at the destination pose. We believe that this will facilitate the adaptation of DMPs. In addition, Table III shows details on the generated motion for task 1, providing the list of shape poses together with their respective CoM translation, the total distance traveled, and the penalty value associated with each destination pose. The suffixes of the pose names enumerate different shape poses of the same support pose. Additionally, the column "LM Prob." shows the probability (between 0 and 1) of observing the transition to the destination pose in the respective line given the known history of the poses before, according to our language model. Since we are using a 5-gram model, the previous four poses are considered for this history.

TABLE III
Details on the planning results of task 1

	Origin Pose	Destination P.	Transl.	Dist.	LM Prob.	Pen.
1	LFRF_1	LF_1	0.07m	0.07m	0.37	-2
2	LF_1	LFRF_2	0.25m	0.32m	0.60	0
3	LFRF_2	RF_4	0.09m	0.41m	0.17	-2
4	RF_4	LFRF_10	0.55m	0.96m	0.31	0
5	LFRF_10	LFRFRH_2	0.04m	1.00m	0.79	0
6	LFRFRH_2	LFRH_4	0.11m	1.11m	0.79	-2
7	LFRH_4	LF_3	0.38m	1.49m	0.79	-4
8	LF_3	LFRF_5	0.41m	1.91m	0.01	-2
9	LFRF_5	LFRFRH_5	0.10m	2.01m	0.54	0
10	LFRFRH_5	RFRH_2	0.09m	2.10m	0.05	-2
11	RFRH_2	LFRFRH_1	0.50m	2.61m	0.30	0
12	LFRFRH_1	LFRH_4	0.07m	2.67m	0.42	-2
13	LFRH_4	LF_3	0.38m	3.05m	0.08	-4
14	LF_3	LFRH_1	0.37m	3.42m	0.02	-2
15	LFRH_1	LF_3	0.58m	4.00m	0.15	-2
16	LF_3	LFRF_4	0.09m	4.10m	0.87	0
17	LFRF_4	RF_2	0.12m	4.22m	0.22	-2
18	RF_2	LFRF_3	0.50m	4.72m	0.55	0
19	LFRF_3	LF_1	0.11m	4.82m	0.42	-2
20	LF_1	LFRF_2	0.25m	5.08m	0.93	0
21	LFRF_2	RF_2	0.11m	5.19m	0.93	-2
22	RF_2	LFRF_3	0.50m	5.69m	0.93	0
23	LFRF_3	LF_2	0.10m	5.80m	0.50	-2
24	LF_2	LFRF_1	0.21m	6.00m	0.74	0

Task 2 requires a total distance of 8 meters, and allows right hand supports from meters 1 to 3 and left hand supports from meters 4 to 6. Finally, task 3 covers a distance of 6 meters, allowing hand supports on both left and right side from meters 2 to 4. The resulting sequences of shape poses for tasks 2 and 3 are also shown in Fig. 5. The total distance covered by the CoM translations for the used transitions is 8.06m and 6.34m, respectively. The video attachment to this paper demonstrates how our planning algorithm finds its solution for the three presented examplary scenarios by showing the path with the highest score (*bestPath* in Algorithm 1) for every 100 iterations of the algorithm. Additionally, the used support poses are depicted at the top

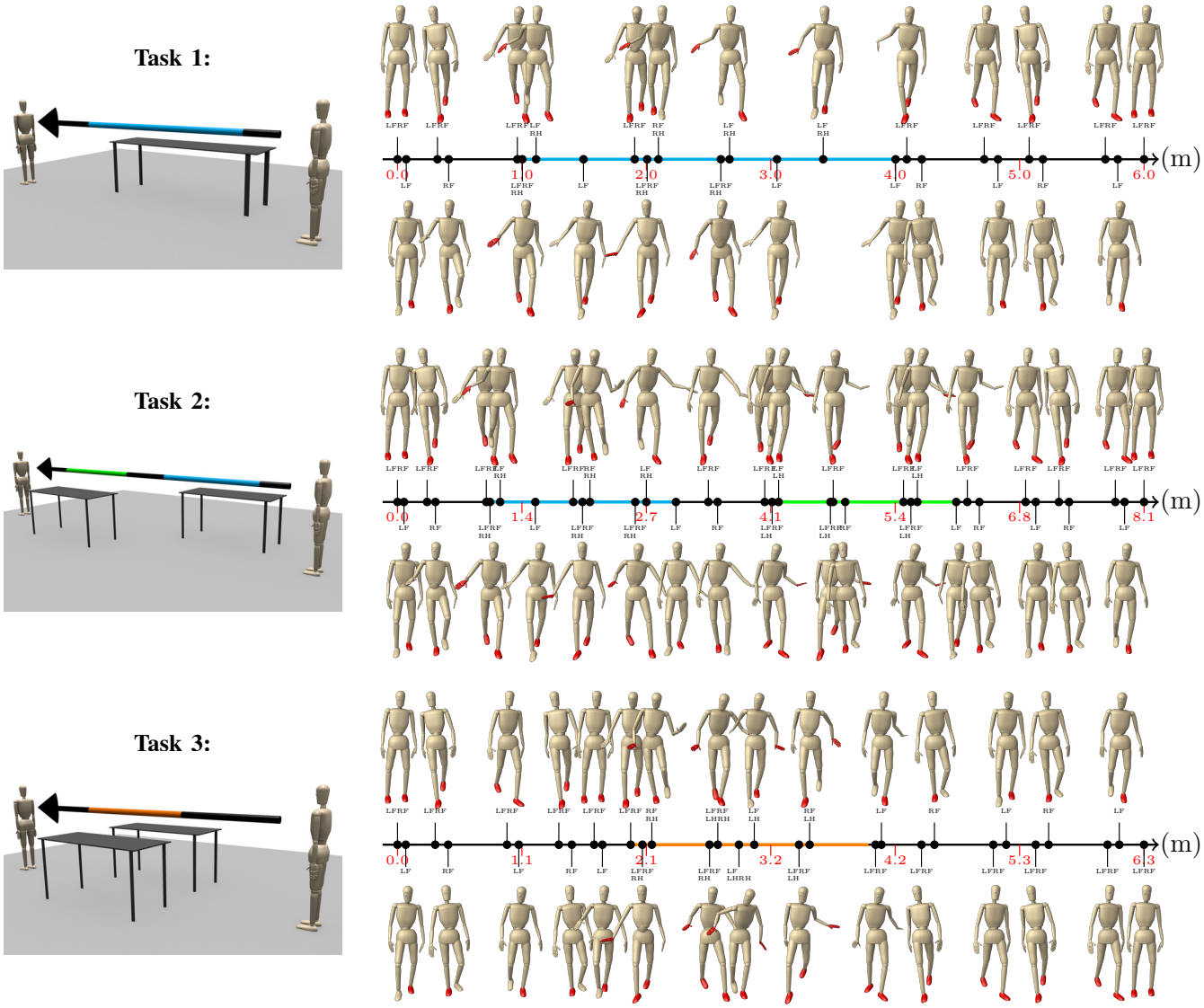

Fig. 5. Visualization of the three different tasks with the tables affording hand supports, and the resulting sequences of shape poses along the translation axis. The blue/green/orange lines represent the intervals where right/left/both hand supports are available, respectively. Labels indicate the body segments used for supports, also highlighted in red in the figures. LF/RF stand for left/right foot, and LH/RH stand for left/right hand, respectively.

margin of the video using the icons introduced in [18].

We believe that the presented results are very promising because they show that our approach is able to provide complex pose transitions that resemble the way humans walk. Despite its simplicity, our approach is able to plan non-trivial pose transitions that provide a very good starting point to plan movement primitives, which will result in specific contact points and joint angle trajectories. The planning algorithm has the advantage that it is computationally inexpensive, and only requires 5171, 7046, and 14052 iterations for the tasks 1, 2, and 3, respectively. While we are currently using only an Python-based prototype implementation of the planning algorithm that is not optimized in any way, future implementations compiled to native code should offer a performance that allows their use in real-time, e.g. for periodic re-planning several times per second.

VI. Conclusions and Future Work

Multi-contact motion planning is a challenging problem in humanoid robotics. In this paper, we have investigated whether techniques from statistical language modeling can be applied to the problem and proposed an innovative approach that learns from human motion capture data and allows the planning of whole-body pose transitions at a high level. This approach is based on a probabilistic n-gram language model, which is learned from sequences of whole-body pose transitions extracted from human motion data. Our results have shown that our approach is able to successfully generate a complex sequence of pose transitions for locomotion along a straight line with left or right hand supports, or both at the same time. In the future, we will collect more motion capture data to extend this to other types of locomotion, including going up and down stairs or climbing a ladder.

Despite the promising results, some limitations of our approach need to be addressed. The probabilistic approach alone does not ensure kinematic and dynamic feasibility of the resulting motion, and some transitions might need an additional intermediate pose to make them feasible. These limitations can be addressed by adding additional constraints to our planner. Furthermore, as stated in the introduction, the work presented here covers only the first part of the problem. We are dividing the problem into a first step of pose transition planning and a second step of DMP integration, and our work has shown the feasibility of an efficient solution for the first part. For the second part of the problem, we have already started working on the extraction of DMPs for each type of pose transition and we will use the results of the first step as the starting point to adapt learned movement primitives to the specific situation and shape of the environment.

REFERENCES

[1] C. Mandery, J. Borràs, M. Jöchner, and T. Asfour, "Analyzing whole-body pose transitions in multi-contact motions," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2015, pp. 1020–1027.

[2] P. Kaiser, N. Vahrenkamp, F. Schültje, J. Borràs, and T. Asfour, "Extraction of whole-body affordances for loco-manipulation tasks," *International Journal of Humanoid Robotics*, vol. 12, no. 3, 2015.

[3] P. Kaiser, M. Grotz, E. E. Aksoy, M. Do, N. Vahrenkamp, and T. Asfour, "Validation of whole-body loco-manipulation affordances for pushability and liftability," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2015, pp. 920–927.

[4] L. Sentis and M. Slovich, "Motion planning of extreme locomotion maneuvers using multi-contact dynamics and numerical integration," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2011, pp. 760–767.

[5] S. Lengagne, J. Vaillant, E. Yoshida, and A. Kheddar, "Generation of whole-body optimal dynamic multi-contact motions," *The International Journal of Robotics Research*, vol. 32, no. 9–10, pp. 1104–1119, 2013.

[6] L. Saab, O. E. Ramos, F. Keith, N. Mansard, P. Soueres, and J. Fourquet, "Dynamic whole-body motion generation under rigid contacts and other unilateral constraints," *IEEE Transactions on Robotics*, vol. 29, no. 2, pp. 346–362, 2013.

[7] K. Bouyarmane and A. Kheddar, "Multi-contact stances planning for multiple agents," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 5246–5253.

[8] ——, "Static multi-contact inverse problem for multiple humanoid robots and manipulated objects," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2010, pp. 8–13.

[9] K. Hauser, T. Bretl, and J.-C. Latombe, "Non-gaited humanoid locomotion planning," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2005, pp. 7–12.

[10] K. Bouyarmane and A. Kheddar, "Humanoid robot locomotion and manipulation step planning," *Advanced Robotics*, vol. 26, no. 10, pp. 1099–1126, 2012.

[11] S. Schaal, "Dynamic movement primitives - a framework for motor control in humans and humanoid robotics," in *Adaptive Motion of Animals and Machines*. Springer, 2006, pp. 261–280.

[12] Y. Zhang, J. Luo, K. Hauser, R. Ellenberg, P. Oh, H. A. Park, and M. Paldhe, "Motion planning of ladder climbing for humanoid robots," in *IEEE International Conference on Technologies for Practical Robot Applications (TePRA)*, 2013, pp. 1–6.

[13] K. Hauser, T. Bretl, K. Harada, and J.-C. Latombe, "Using motion primitives in probabilistic sample-based planning for humanoid robots," in *Algorithmic Foundation of Robotics VII*. Springer, 2008, pp. 507–522.

[14] K. Hauser, "Large motion libraries: Toward a google for robot motions," *RSS Workshop on Robotics Challenges and Vision*, pp. 5–8, 2013.

[15] D. Kulić, C. Ott, D. Lee, J. Ishikawa, and Y. Nakamura, "Incremental learning of full body motion primitives and their sequencing through human motion observation," *The International Journal of Robotics Research*, vol. 31, no. 3, pp. 330–345, 2012.

[16] D. Kulić, W. Takano, and Y. Nakamura, "Incremental learning, clustering and hierarchy formation of whole body motion patterns using adaptive hidden markov chains," *The International Journal of Robotics Research*, vol. 27, no. 7, pp. 761–784, 2008.

[17] ——, "Online segmentation and clustering from continuous observation of whole body motions," *IEEE Transactions on Robotics*, vol. 25, no. 5, pp. 1158–1166, 2009.

[18] J. Borràs and T. Asfour, "A whole-body pose taxonomy for loco-manipulation tasks," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 1578–1585.

[19] S. Richter and M. Westphal, "The LAMA planner: Guiding cost-based anytime planning with landmarks," *Journal of Artificial Intelligence Research*, vol. 39, no. 1, pp. 127–177, 2010.

[20] R. P. Petrick and F. Bacchus, "A knowledge-based approach to planning with incomplete information and sensing," in *International Conference on Artificial Intelligence Planning and Scheduling (AIPS)*, 2002, pp. 212–222.

[21] M. Botvinick and M. Toussaint, "Planning as inference," *Trends in Cognitive Sciences*, vol. 16, no. 10, pp. 485–488, 2012.

[22] Y. Yang, Y. Aloimonos, C. Fermüller, and E. E. Aksoy, "Learning the semantics of manipulation action," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015, pp. 676–686.

[23] D. Summers-Stay, C. L. Teo, Y. Yang, C. Fermüller, and Y. Aloimonos, "Using a minimal action grammar for activity understanding in the real world," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012, pp. 4104–4111.

[24] N. Dantam and M. Stilman, "The motion grammar: Analysis of a linguistic method for robot control," *IEEE Transactions on Robotics*, vol. 29, no. 3, pp. 704–718, 2013.

[25] R. Rosenfield, "Two decades of statistical language modeling: Where do we go from here?" *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1278, 2000.

[26] L. Wu, M. Li, Z. Li, W.-Y. Ma, and N. Yu, "Visual language modeling for image classification," in *International Workshop on Multimedia Information Retrieval (MIR)*, 2007, pp. 115–124.

[27] W.-T. Chu, Y.-L. Lee, and J.-Y. Yu, "Visual language model for face clustering in consumer photos," in *ACM International Conference on Multimedia*, 2009, pp. 625–628.

[28] C. Mandery, O. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour, "Unifying representations and large-scale whole-body motion databases for studying human motion (to appear)," *IEEE Transactions on Robotics*, 2016.

[29] O. Terlemez, S. Ulbrich, C. Mandery, M. Do, N. Vahrenkamp, and T. Asfour, "Master Motor Map (MMM) - framework and toolkit for capturing, representing, and reproducing human motion on humanoid robots," in *IEEE-RAS International Conference on Humanoid Robotics (Humanoids)*, 2014, pp. 894–901.

[30] P. Azad, T. Asfour, and R. Dillmann, "Toward an unified representation for imitation of human motion on humanoids," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2007, pp. 2558–2563.

[31] C. Mandery, O. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour, "The KIT whole-body human motion database," in *International Conference on Advanced Robotics (ICAR)*, 2015, pp. 329–336.

[32] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *International Conference on Spoken Language Processing*, 2002, pp. 901–904.

[33] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 13, no. 4, pp. 359–393, 1999.