Learning to recognize parallel combinations of human motion primitives with linguistic descriptions using non-negative matrix factorization

Olivier Mangin and Pierre-Yves Oudeyer

***Abstract*— We present an approach, based on non-negative matrix factorization, for learning to recognize parallel combinations of initially unknown human motion primitives, associated with ambiguous sets of linguistic labels during training. In the training phase, the learner observes a human producing complex motions which are parallel combinations of initially unknown motion primitives. Each time the human shows a complex motion, he also provides high-level linguistic descriptions, consisting of a set of labels giving the name of the primitives inside the complex motion. From the observation of multi-modal combinations of high-level labels with high-dimensional continuous unsegmented values representing complex motions, the learner must later on be able to recognize, through the production of the adequate set of labels, which are the motion primitives in a novel complex motion produced by a human, even if those combinations were never observed during training. We explain how this problem, as well as natural extensions, can be addressed using non-negative matrix factorization. Then, we show in an experiment in which a learner has to recognize the primitive motions of complex human dance choreographies, that this technique allows the system to infer with good performance the combinatorial structure of parallel combinations of unknown primitives.**

I. Introduction

Personal and social robots, and the ambient intelligent systems in which they shall be integrated, are expected to play an important role for assisting humans in everyday environments, ranging from educational contexts to helping people with physical or cognitive disabilities. Recognizing and understanding the behaviour of humans is a key capability for these systems. Yet, building such capabilities entails diverse and difficult technical challenges [1], partly due to the fact that each human user and each task context has its own particularities, i.e. different humans often behave differently in similar contexts, and vice versa the behaviour of one human, for example its motions, can have different properties and structures in different contexts. Thus, there is a strong need for capabilities to adapt to users, which importantly can be achieved through learning [1].

Human behaviour is highly diverse and multi-modal, and the properties of its structure and interpretation are equally diverse. Body motions are important parts of such behaviours, and their structure can be quite complicated. They are often composed of primitive motions, forming an action vocabulary that is combinatorially re-used to build, parse and recognize complex motions. This recombination for forming complex motions typically happens through a mixture of sequencing and parallel or concurrent execution, such as for example walking forward *after* or *while* grasping an object or saying "hello" with the hand, or for example clapping with two hands while smiling with the face. One problem for artificial systems to learn this structure and recognize both the primitive and complex motions, is that humans only rarely produce motion primitives one by one and do not provide "class/linguistic labels" associated unambiguously with them. Rather, humans typically produce only mixtures of these primitives, and in any given mixture the primitives are blended together in such a way that it can be difficult to segment them apart from the observation of one single complex motion. Also, while humans can provide multi-modal cues such as linguistic labels, they are often ambiguous: the human might describe with multiple words his complex motion, but with no easy way to guess the direct relation between each word and its possible corresponding subparts in the complex motion.

In this article, we address aspects of these issues by focusing on the problem of learning to recognize a particular family of combinatorial full body motions (encoded as high-dimensional flow of continuous values) associated with the production of sets of linguistic labels (here encoded symbolically). More precisely, we consider a human producing complex motions (such as dance movements), that can also be called complex gestures or choreographies, which are parallel combinations of motion primitives (we do not consider sequencing here). Each motion primitive corresponds to a coordinated motion over a subset of the dimensions of the human limbs, and two motion primitives might operate on different subsets. Producing parallel combinations consists here in producing concurrently and in parallel several motion primitives (typically 2 or 3 at the same time taken in a repertoire of 20 to 50 primitives). We assume here that the human only combines independent primitives in each single complex motion (i.e. their subset do not intersect in a given demonstration and the execution of one has little effect on the execution of the other).

In the training phase, the learner observes a human producing examples of such unsegmented complex motions. Each time the human shows a complex motion, he also provides high-level linguistic descriptions, consisting of an unordered set of labels. Each label corresponds to the name of one of the primitive motions composing the complex motion, but the labels are ambiguous since the primitives are neither segmented nor known by the learner initially (i.e. he does not know how many there will be, on what limb dimensions each one operates, and what are their values on those unknown dimensions). From these multi-modal observations,

Flowers Team, INRIA, ENSTA ParisTech, France
Université Bordeaux 1, France

the learner has to infer internal representations that allow it to later be able to recognize, through the production of the adequate set of labels, which are the motion primitives in a novel complex motion produced by a human, even if those combinations were never observed during training.

We propose here that the learning technique of Non-negative Matrix Factorization (NMF) [2] can be efficiently used to solve this problem. It has been used in the past in various machine learning and perceptual processing tasks, such as image parts learning and semantic feature extraction from text [3], acoustic source separation [4] and language acquisition [5]. Several of its central properties, such as the ability to infer primitive percepts from the observation of combined percepts, the ability to process and leverage multi-modal data, and its ability to achieve multi-way reconstruction of data and modalities, are particularly suited to our problem. Furthermore, it has been used successfully in past work for two particular problems which, while different in several aspects, also share similarities with our problem. In Driesen et al. [5], NMF has been used to learn to recognize *speech* keywords (primitives) in continuous complex sentences from the observations of complex unsegmented speech sentences (i.e. *sequencing* combinations of words) associated with ambiguous sets of semantic labels corresponding to these keywords. Our work shows that NMF can be also used to learn to recognize and classify combinatorial *motions* from the observation of *parallel* combinations of motions associated with ambiguous sets of semantic labels. In Hellbach et al. [6], NMF has also been used successively to achieve forward prediction of unimodal and globally unstructured motions. Our work shows how NMF can recognize and classify combinatorial motions leveraging ambiguous multi-modal cues.

In the next section, we discuss in more details the related work. Then, we present a formal definition of the problem we attack, before describing both the NMF technique and how it can be used to solve this problem. Finally, we present experiments showing how this technique can allow to learn to recognize initially unknown primitive dance motions in complex dance choreographies produced by a human, even when the combinations of primitives were never observed during training. We conclude by discussing the limits, as well as the potential extensions of the system. We explain in particular how the system could naturally be extended to deal with real unsegmented speech streams as the linguistic cues.

II. Related work

In this article we focus on the problem of learning to recognize structured human behaviours from full body motion and linguistic descriptions. This particular problem illustrates two main challenges of human behaviour understanding presented by Salah et al. [1].

First it requires dealing with the structural complexity of human behaviours, a problem that has been largely explored in works from both action recognition and reproduction communities, as discussed in Section II-A.

Then, using social or linguistic cues in a multi-modal learning framework, also introduces difficulties, discussed in Section II-B such as dealing with language structure and ambiguity and relating this information to other modalities such as the learning of motions.

A. Gesture recognition and reproduction

Full body motions are high-dimensional signals where interesting patterns occur both simultaneously, in sequences and at different time scales. A practical approach to the associated learning problem is to decompose complex motions into simpler elements called **motion** or **motor primitives** that can be composed to represent observed actions or produce new ones. Motion primitives are however not observed separately in natural behaviours, and it is an important feature of a recognition or reproduction system not to require demonstrations to be explicitly segmented into primitives.

Hidden Markov models (HMM) have been largely used to learn sequences of primitives. For example, Kulic and Nakamura have proposed in [7] a method that first performs an unsupervised segmentation of the motion signal into small blocks through a first HMM, and then performs clustering over a HMM representation of the found blocks, thus learning motion primitive as clusters. Kruger et al. [8], have proposed to first discover primitives by clustering action effects on manipulated objects and then use the found clusters and associated segmentation to train parametrized hidden Markov models that allow recognition and reproduction. Finally Calinon et al [9] and Butterfield et al. [10] use Gaussian mixture models to represent motion primitives and HMMs to discover and represent the transitions and sequential combinations of primitives.

Dictionary learning techniques have also been applied to the discovery of motion primitives from sequences. For example, Li et al. [11] have used orthogonal matching pursuit to decompose complex motions into simple motion patterns activated briefly along the time dimension. The decomposition can then be used to perform both compression and classification. Hellbach et al. [6] have also used non-negative matrix factorization to perform a decomposition of globally unstructured motions in low level components and use it in a prediction problem.

In this article we address the complementary problem of learning to recognize motion primitives from demonstrations in which primitives are active simultaneously. Our contribution differs from previous work by Hellbach et al. [6] by the kind of combinatorial learning our system is able to perform, by the multi-modal setting that we present, and by the fact that we use NMF to achieve classification.

B. Multi-modal learning with linguistic signal

In [1] Salah identifies the key role played by social signal in understanding human behavior. On the other hand learning motions with linguistic guidances has also been shown to benefit the motion learning itself. As an example, Tuci et al. [12] have shown that learning a compositional structure shared between action and language can allow robotic agents

to achieve behaviours that were not encountered in training. Furthermore Massera et al. [13] have demonstrated that providing linguistic instructions can facilitate the acquisition of a behavioural skill, in comparison to pure motor learning.

Other models of joint learning of behavioural and linguistic knowledge have been developed by Sugita and Tani [14], using recurrent neural networks, and more recently by Cederborg and Oudeyer [15], using a clustering algorithm. However both experiments involve very simple (one verb and a noun from sets of three words and three nouns for each demonstration in [14]) or no combinatorial structure in language and behaviours (single word, single gesture demonstrations for [15]).

In [5], Driesen et al. have successfully used NMF to learn to recognize speech keywords from unsegmented speech signals. More precisely, in their experiment, full speech sentences are presented to the learner while associated with a semantic symbolic labels. After some training, the learner is able to reconstruct the symbolic label associated to a given audio stream.

In the present article we explore the complementary experiment in which the semantic part, consisting of motions, is learnt from demonstrations of real human motions and the linguistic part is symbolic. The simplified symbolic linguistic representation used in this article is however similar to those used by Sugita and Tani [14] and Tuci et al. [12]. Furthermore we extend the method presented in Driesen et al. to learn complex motions with many ambiguous labels.

Finally exploring the use of NMF for motion learning, as similar algorithms have already been developed for language learning, is a useful step towards discovering similar structures in both language and motion learning and performing joint language and motion multi-modal learning.

III. Task definition

In this article the learning occurs between two modalities. The first one consists in complex human demonstrated movements in which several primitive motions are active at the same time. The second one consists in linguistic descriptions of the movements where each sentence contains several speech keywords (represented symbolically here), each describing a primitive gesture.

We consider, for illustration and experiments, complex human movements as choreographies (see Figure 1). Each choreography is composed of several primitive dance motions (typically two or three in our setting), for example, one leg gesture combined with one left arm gesture and one right arm gesture, or one leg gesture combined with a gesture involving the coordination of the two arms (such as clapping one's hands). In this article we only consider choreographies involving compatible combinations of dance gestures, which means no choreographies contains two *left arm* movements at the same time, for example.

From a high level point of view, given a set $\mathcal{G}$ of primitive dance gestures, which constitutes the *dance vocabulary*, possible choreographies are described by a subset $\mathcal{C}$ of the parts of the vocabulary, $\mathcal{C} \subset \mathcal{P}(\mathcal{G})$ (typically all sets of two or three compatible dance gestures). This set can be seen as a representation of our *dance grammar*.

Since the learner is only exposed to real human demonstrations of the choreographies, it does not observe choreographies in that high level form but only through their realisations (which are subject to noise and variation). Each demonstration x^i is a sequence of values $(x^i_0, x^i_1, ..., x^i_{T_i})$, where each x^i_t is an observation (typically a vector of 2D or 3D marker positions or joint angles for all joints), and T_i is the length of the demonstration.

The language description shares a somehow similar structure: keywords or labels from a set $\mathcal{L}$ are associated to gestures, and combined into sentences from a set $\mathcal{S} \subset \mathcal{P}(\mathcal{L})$. In this article we only consider symbolic labels. More precisely when the sentence $s = \{l_1, l_2, l_3\} \in \mathcal{S}$ is used to describe a choreography, the system observes a vector $y^i \in \mathbb{R}^L$ (L is the total number of labels, $L = |\mathcal{L}|$) such that for $j = 1, ..., L$, y^i takes value 1 if $l_j \in s$, and 0 elsewhere. For example if 5 labels are considered, the sentence containing labels 1 and 3 would be represented by vector: $(1,\ 0,\ 1,\ 0,\ 0)^T$.

The learning problem considered in this article consists in two phases illustrated in Figure 1. In a training phase the robot observes joint motion demonstrations and linguistic descriptions, that is to say each demonstration consists in a couple of vectors (x^i, y^i).

Then in the testing phase, the system is only given a motion demonstration, that is to say a vector x^{test} and has to produce the associated linguistic description, that is to say the vector y^{test}.

IV. Learning joint language and gesture structure with NMF

Non-negative matrix factorization (NMF [2], [3]) is a class of machine learning problems and methods, often used to solve dictionary learning problems similar to the one presented in this article.

More precisely given a set of examples represented by vectors $v^i \in \mathbb{R}^m$ ($1 \leq i \leq n$), a dictionary learning problem consists in finding both a dictionary containing vectors $w^j \in \mathbb{R}^m$ ($1 \leq j \leq k$), called **atoms**, and **coefficients** $h^i \in \mathbb{R}^k$ such that each example can be represented as a linear combination of atoms with these coefficients: $v^i = \sum_{j=1}^{k} h^i_j w^j$.

If example vectors are stacked as columns of a data matrix $V \in \mathbb{R}^{m \times n}$, this problem can be written as finding matrices $W \in \mathbb{R}^{m \times k}$ (which columns are atoms) and $H \in \mathbb{R}^{k \times n}$ (which columns are coefficients) such that:

$$V \simeq W \cdot H.$$

When the inner dimension k of the product is smaller than original dimensions m and n of the data, this representation achieves data compression by capturing structure in the matrix W. Thus the reconstruction is not always exact and the "$\simeq$" takes the form of a minimization of a reconstruction error.

The non-negative matrix factorization problem focuses on the case where V, W and H have non-negative coefficients, a

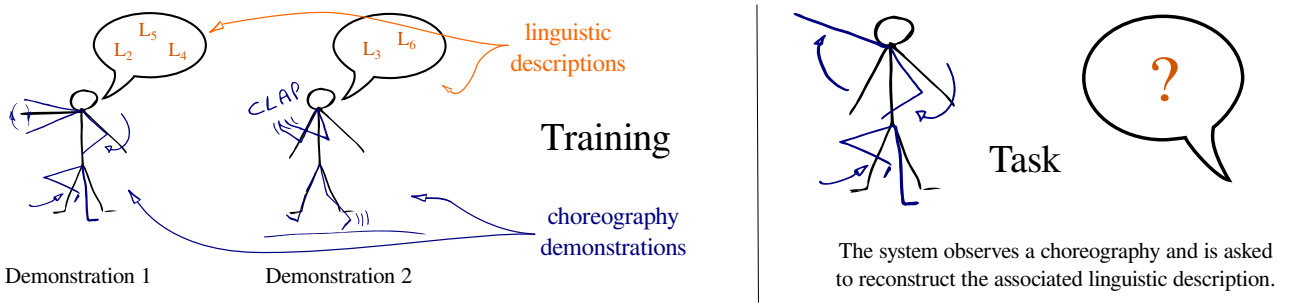

Fig. 1. In the training phase, the learner observes demonstrations of choreographies composed of several elementary gestures, and the associated set of linguistic labels (left). After that the learner has to reconstruct the set of labels associated to a demonstrated choreography (right). (Best seen in color)

constraint relevant in cases such as event or histogram based representations: each atom represent the joint occurrence of some events and things such that an event occurring a negative number of times does not make sense.

Efficient algorithms have been developed, based either on multiplicative updates or alternate gradient descent, for reconstruction errors between the original data V and the reconstructed $W \cdot H$ based on Frobenius norm, Kullback-Leibler divergence and more generally for the class of β-divergences, [2], [3], [4].

In this article we uses the plain NMF algorithm based on multiplicative updates, for both errors based on Frobenius norm and Kullback-Leibler divergence, as described in [3].

A. NMF for multi-modal learning and classification

The NMF algorithm presented in previous section only learns, in an unsupervised manner, a transformation of the original data V into an internal representation H. However, following Bosh, Driesen et al. [16], [5], NMF can be used in a multi-modal framework.

In the problem under consideration each training example contains both a demonstration of the choreography and a linguistic description. Let's consider that the first dimensions of the vectors v^i are storing the motion part of the demonstration and the last ones are used to represent the linguistic part: $(v^i)^T = ({v^i_{motion}}^T \ {v^i_{language}}^T)$.

Both matrices V and W are thus composed of a motion and a language part:

$$V = \begin{pmatrix} V_{motion} \\ V_{language} \end{pmatrix} \qquad W = \begin{pmatrix} W_{motion} \\ W_{language} \end{pmatrix}$$

The inner representation H of the examples is not associated with a specific modality.

In this article NMF algorithm is used in two different ways to first learn the transformation from multi-modal examples to an internal representation, and then use this transformation to reconstruct one modality from another.

1) NMF to learn internal representation from multi-modal demonstrations: In the learning part NMF is applied to a V^{train} data matrix and both W^{train} and H^{train} matrices are learned. The W^{train} matrix is the matrix of most interest since it encodes the structure that has been learned on the data, when H^{train} only encodes the representation of training examples.

2) NMF to reconstruct linguistic descriptions from choreographies demonstrations: Reconstructing the linguistic parts associated to demonstrations of motions, i.e. classifying motions, corresponds to finding the missing $V^{test}_{language}$ given a V^{test}_{motion}.

This operation is performed through two steps:

- reconstructing internal states of the system from demonstrations, which means finding the best matrix H^{test} for the approximation: $V^{test}_{motion} \simeq W^{train}_{motion} \cdot H^{test}$. This step can be performed through NMF algorithms by constraining the W matrix to be constant.
- once H^{test} has been found, the associated linguistic part can be computed through matrix product: $V^{test}_{language} \simeq W^{train}_{language} \cdot H^{test}$

It should be noted here that the reconstructed matrix $V^{test}_{language}$ is not constrained to take only $0, 1$ values like the provided linguistic matrix. This issue is addressed by using a thresholding mechanism (where the threshold is learned by cross-validation on training data), as detailed in Section V-B.

The value of k is a parameter of the system that is fixed to 150 for the experiments presented in this paper. The number of atoms used by the system to represent observed data is quite important to achieve the desired objective. If k is too big, the algorithm does not have to compress data and can use motion only and language only atoms. On the other hand, if k is too small, the system cannot represent the complexity of the data and may focus on representing components that have bigger effects but less multimodal meaning.

B. Histogram representation of choreographies

An important requirement to use NMF techniques is to be able to represent data with vectors of non-negative coefficients which can be combined through non-negative weighted sums. Such a representation of motion data is presented in this section.

In this experiment motions are captured as trajectories in angle and angle velocity spaces of several articulations of the human body. Each trajectory on a specific articulation (or degree of freedom) is considered separately and the entire sequence of angles and velocities is transformed into a histogram, represented by a fixed length non-negative vector. Vectors obtained for each degree of freedom are then concatenated into a larger vector.

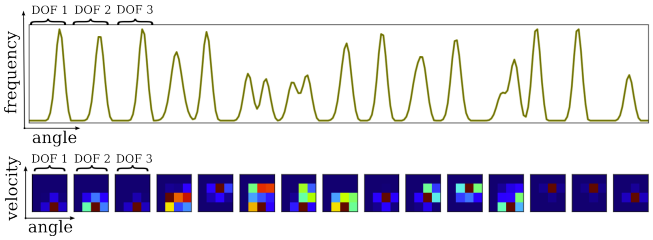

Fig. 2. Illustrations of concatenated histograms on positions (top) and joint position and velocities (bottom). For the first one, x axis is associated to different possible values for each angles and y axis to frequencies. On the second one frequencies are represented through colors, x and y axis correspond respectively to values of angles and velocities. (Best seen in color)

We explore different approaches for the transformation of angles and velocities sequences into histograms. They differ on two modelling choices: 1) Which data is used to build histograms? 2) Which method is used to build histograms?

Answers to the first question are related to the use of angles and velocities values. While velocities can bring precious information, there are several ways of integrating this information in the histogram representation: 1) consider **only angles**, 2) consider **only velocities**, 3) treat **angles and velocities as separate degrees of freedom**, 4) or use the two-dimensional angle-velocity vectors that is to say build histograms on the **joint angle-velocity** space (see Figure 2).

We study two methods for building histograms. 1) Smoothed histograms can be built on **regularly distributed bins**. More precisely we split the angle, velocity or joint angle velocity space into a regular grid of bins. Histograms are built by counting the number of samples from the trajectory falling into each bin and dividing by the length of the trajectory. A Gaussian smoothing kernel is used to make point by point comparison of histograms more robust to perturbations [17]. These methods are referred to as Kernel Density Estimation (KDE) in the following. 2) An alternative approach is to build histograms over a **vector quantization**, which is a more adaptive binning process. Vector quantization (VQ) is performed through a k-means algorithm. Then a histogram is built by counting the proportion of samples falling into each cluster. We explore the use of both hard (each histogram is only counted in one cluster) and softmax (each sample is counted in each cluster with a weight depending on its distance to the cluster's centroid) centroid associations.

Representing motion data by separate histograms on each degree of freedom leads to two approximations: 1) for a given measurement in the trajectory, information about dependency between different degrees of freedom is dropped, 2) the sequential information between measures for a given degree of freedom is dropped.

Similar simplification have however been successfully used in two other fields. Bosh et al. [16] have demonstrated that, even if sequential information may appear necessary in language, and especially in speech utterances, very good word discovery can be achieved without considering this sequential information. Both in text classification and in computer vision *bag-of-words* techniques also achieve good performances by dropping positional information of extracted local features [18], [19].

Furthermore using histograms built on joint angle positions and velocities is similar to representing transitions in angle space. By representing the sequence through its transition we approximate it by a Markovian process. Such an approximation is quite common in the gesture recognition and motion planning literature [20], [21].

V. Experimental setup

In this article a learner is trained on a set of complex full body human motions associated with linguistic descriptions and then asked to produce the linguistic description associated to given test demonstrations of the motion, including demonstrations built as previously unseen combinations of primitives (see Figures 1 and 3).

In order to demonstrate these capabilities we perform two kinds of experiment.

- First the system is tested on simple human motions, each containing only one primitive dance gesture. These experiments demonstrate that the motion representation we use is sufficient to perform motion classification, which corresponds to a simple case of multi-modal learning. We also compare different aspects of the representation.
- Then the system is tested on complex full body human motions to demonstrate its ability to capture the combinatorial structure of the choreographies by exploiting ambiguous linguistic labels.

A. Data acquisition and transformation

The data used in the present paper has been acquired from a single human dancer through a Kinect™ device and the OpenNI™ software[1] that enables direct capture of the subject skeleton.

The device and its associated software provides an approximate 3D position of a set of skeleton points. These points are then converted into 12 angle values representing the dancer position at a specific time. This conversion is achieved through a simple geometrical model of human limbs.

At this step each demonstration is a sequence of angle vectors. In order to capture more information angle velocities are also extracted: a delayed velocity is used to achieve better robustness to noise in the angle sequences. More precisely $\dot{x}_t = x_t - x_{t-d}$ is used to compute the velocities, instead of being restrained to the case where $d = 1$. It is not necessary to divide by the fixed time step since the histogram representation described in Section IV-B is invariant to scaling all the data by the same amount.

The primitive dance motions used in our gesture datasets and illustrated in Figure 3 and Table I, are either associated to legs as for example *squat* and *walk* movements, to both arms e.g. *clap hands* and *paddle*, or to left or right arm, e.g. *punch*, *wave hand*. Yet this structure is not known by

[1] http://www.openni.org

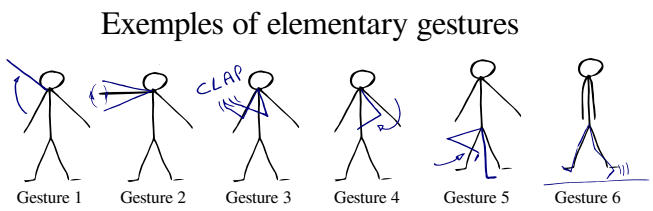

Fig. 3. Illustration of the primitive dance movements that compose demonstrated choreographies.

TABLE I

PRIMITIVE DANCE MOTIONS FROM THE *small mixed dataset*

Id	Limb(s)	Description
1	right arm	hold horizontal
5	right arm	raise from horizontal to vertical
6	right arm	lower from horizontal to vertical
10	right arm	hold horizontal and bring from side to front
19	both arms	clap hands (at varying positions)
20	both arms	mimic paddling on the left
21	both arms	mimic paddling on the right
22	both arms	mimic pushing on ski sticks
23	legs	un-squat
24	legs	mimic walking
25	legs	stay still
28	right leg	raise and bend leg to form a flag (or "P") shape
30	left arm	hold horizontal
38	left arm	mimic punching
40	left arm	lower forearm from horizontal position
43	left arm	swing forearm downside with horizontal upper arm

the system initially. They correspond to both discrete and rhythmic movements.

Three motion datasets are considered in this article. A first dataset is used to separately study the efficiency of the various representations. In this dataset each demonstration only includes one primitive dance motion. There are 47 different dance primitive and the set contains 326 demonstrations. This dataset is referenced as **single primitive dataset**.

Two other datasets are composed of demonstrations of complex choreographies, composed of two or three randomly chosen compatible (i.e. spanned over separate degrees of freedom) primitive motions. The first one contains 137 examples of combinations of 17 distinct primitive dance motions. The second one, contains 277 examples with 47 primitive dance motions (the same as in single primitive dataset). These datasets are referenced as **small** and **full mixed dataset**.[2]

Since the datasets only contain a relatively small number of examples we used *leave-one-out* cross validation to build test and train sets. Presented results are averaged over all possible test sets. With the *full mixed dataset* examples presented for testing contain a combination of primitive movements that in 60% of the cases have not been observed during training.

B. Evaluation

In each experiment the method based on NMF described in Section IV yields a vector of scores over keywords, which forms the linguistic reconstruction. The quality of this reconstruction is evaluated by comparison between the reconstructed $\hat{y}$ (with continuous values) and y from ground truth (with $0, 1$ values) through the following score functions:

1) Score function for single gesture experiment: In that case the good linguistic representation only contains a 1 at the position of the label associated to the demonstrated gesture and 0 elsewhere. The score function is defined as:

$$l_{\text{single}}(\hat{y}, y) = \begin{cases} 1 & \text{if } \underset{i}{\text{argmax}}\ \hat{y}_i = \underset{i}{\text{argmax}}\ y_i \\ 0 & \text{else} \end{cases}$$

2) Score function for mixed gesture with given number of gestures: In that case several elementary gestures are present in each example. The reconstructed vector is tested on the fact that gestures that are actually present in the experiment have the best scores.

It can be described by the following equation, where $\#(y)$ denotes the number of gestures present in the demonstration and $\text{best}(n, \hat{y})$ is defined as the set containing the n gestures having the best scores in $\hat{y}$.

$$l_{\text{given number}}(\hat{y}, y) = \begin{cases} 1 & \text{if } \text{best}(\#(y), y) = \text{best}(\#(y), \hat{y}) \\ 0 & \text{else} \end{cases}$$

In other words the system is given the number of elementary gestures present in the example and asked which are those gestures.

3) Score function for mixed gestures, exact reconstruction: This score function evaluates the exact reconstruction of the linguistic description. It requires the reconstructed vector to be converted to a discrete one before comparison.

For that purpose an additional thresholding mechanism is added to the system: after reconstruction through NMF, all values from $\hat{y}$ above a threshold are put to 1, and others are put to 0. The threshold η is evaluated through cross-validation on the training data.

The score function is then simply defined as:

$$l_{\text{full}}(\hat{y}, y) = \begin{cases} 1 & \text{if } y = \text{threshold}(\hat{y}, \eta) \\ 0 & \text{else} \end{cases}$$

In each case the score function defined above for one example is averaged over all examples from the test set to yield a final score in $[0, 1]$.

VI. RESULTS

A. Demonstrations with a single primitive

We performed a first set of experiments on the *single primitive* dataset in order to evaluate our learning system on a simple multi-modal learning task.

In this section primitive dance movements are presented to the learning system with unambiguous labels and the recognition performances are evaluated with the l_{single} score function. We focus on comparisons of the various parameters of the motion representation.

The first experiment compares the use of regular binning with Gaussian smoothing (KDE) and adaptive binning (VQ)

[2]Dataset and examples available at `http://flowers.inria.fr/choreography_database.html`

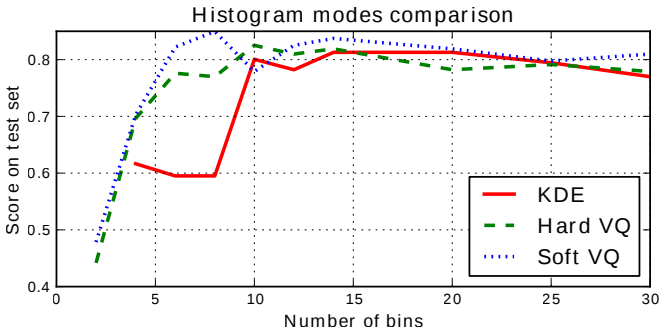

Fig. 4. Comparison of mechanisms used to build the 2D joint angle-velocity histograms, for different number of bins. KDE refers to regular binning, hard and soft VQ to the vector quantization modes described in Section IV-B. (Best seen in color)

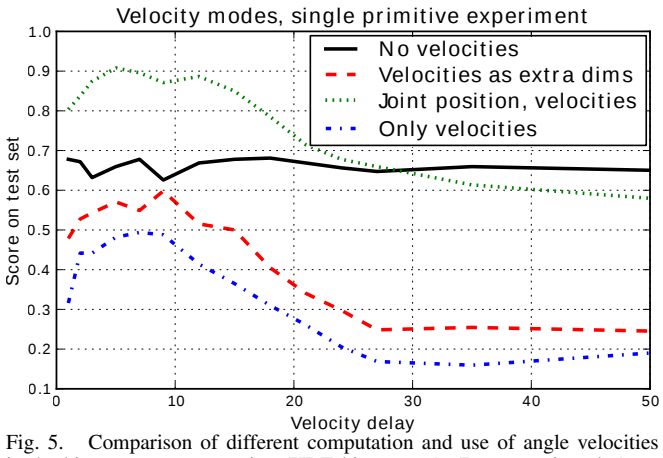

Fig. 5. Comparison of different computation and use of angle velocities in the histogram representation (KDE histograms). (Best seen in color)

with both hard and softmax associations to build the histograms. The comparisons are performed over a range of total number of bins, for **joint angle-velocity** 2D histograms.

Results from this experiment in Figure 4 outline the advantage of using vector quantization over regular binning (KDE) for small numbers of bins, which corresponds to low resolution of the input. This difference is however less sensitive for larger numbers of bins. A typical issue of regular binning, that can explain the better results with adaptive binning, is that for the same grid step (i.e. resolution), the number of required bins grows exponentially with dimension. Even with two dimensional histograms, a maximum number of ten bins would lead to a three-by-three (thus low resolution) regular binning. In the same situation adaptive binning can identify ten relevant clusters.

A second experiment is performed to compare the efficiency of histograms built either only on **angles**, only on **velocities**, on **angles and velocities as separate degrees of freedom** and on the **joint angle-velocity** space. We compared these representations of the motion over a range of values for the delay used in velocity computation, and using KDE histograms with a fixed total number of 15 bins by degree of freedom. The results of the second experiment, presented in Figure 5, demonstrate that histograms on joint angle and velocities values capture the most information from the original motions.

TABLE II

RESULTS ON THE MIXED DATASETS

	l_{full}	$l_{\text{given number}}$
17 labels (SVM, linear)	0.818	-
17 labels (NMF, Frobenius)	0.854	0.971
17 labels (NMF, DKL)	0.789	0.905
47 labels (SVM, linear)	0.422	-
47 labels (NMF, Frobenius)	0.625	0.755
47 labels (NMF, DKL)	0.574	0.679

B. Demonstrations with complex choreographies

In this section we evaluate the performance of our learning system on the full choreographies with ambiguous labels. We used regular binning for the building of 2D histograms of joint angles and velocities, with 15 bins.

Table II presents results obtained on the two mixed datasets for both Kullback-Leibler (DKL) and Frobenius versions of NMF algorithm. The reconstructed label vectors are evaluated by $l_{\text{given number}}$ and l_{full} score functions which enables to understand which part of the error is due to the thresholding mechanism.

For comparison purposes we also tested a method based on support vector machines (SVM) on our dataset. More precisely we trained one SVM for the recognition of each label. The SVM method directly yields a set of recognized labels, with no need for thresholding. However this method relies entirely on the symbolic form of the labels and won't generalize to other multi-modal settings with continuous linguistic modalities. There is no such theoretical limitation on our NMF setting (see discussion in Section VII).

The results in Table II demonstrates that after being exposed to demonstrations of mixed primitive dance motions associated with ambiguous labels, the presented system is able to successfully produce linguistic representations of newly demonstrated choreographies. The second dataset can be considered as difficult since each one of the 47 primitive dance motions only appears in an average of 14 demonstrations which labels are ambiguous.

C. Handling unknown combinations of primitives

The ability of our system to capture the combinatorial structure of the data is illustrated by its behaviour on unknown combinations of motion primitives. For instance in the *full mixed dataset* more than 60% of the examples demonstrates a combination that is not observed in other examples.

In order to get more precise results for this behaviour we set up a slightly different experiment where test sets are only composed of combinations of motion primitives that were not observed during training. The results of this experiment are reported in Table III.

VII. DISCUSSION AND FURTHER WORK

In a first experiment we demonstrated on a dance motion recognition task the efficiency of our method based on a histogram representation of motion and a NMF algorithm. This contribution extends the scope of NMF applications to learning motion from prediction tasks, as shown by

TABLE III

RESULTS ON MIXED DATASET WITH TESTING ON COMBINATIONS NEVER OBSERVED IN TRAINING

	l_{full}	$l_{\text{given number}}$
17 labels (NMF, Frobenius)	0.568	0.800
17 labels (SVM, linear)	0.667	-
47 labels (NMF, Frobenius)	0.406	0.653
47 labels (SVM, linear)	0.206	-

Hellbach et al. [6], to classification. The motion representation presented in this paper enable the application of the architecture developed by Bosh, Driesen et al. [16], [5] for speech learning to motion learning. This constitutes a useful step toward comparison of structural similarities between language and motion learning.

In a second experiment we showed that the architecture presented in this paper is capable of learning to recognize complex gestures, composed of simultaneous motion primitives, while only observing ambiguous symbolic labels. It is demonstrated in a third experiment that the system has captured the combinatorial structure of the observed gestures and is capable of generalization by recognizing combinations that were never observed in training.

We presented a learning system that is capable, after learning from demonstrations of complex gestures and linguistic descriptions, to re-construct the linguistic modality from an example involving only the motion modality. The experiments that we performed only use a symbolic representation of speech labels. It is however possible to replace this symbolic representation by real acoustic data (for example represented in the same way than in [5]) without changing the learning and reproduction algorithms. However, evaluation of such an extension remains to be done and in such a setting evaluating the reconstruction by comparison to ground truth labels would not be direct any more.

The method presented in this article also makes it possible to reconstruct the motion representation associated with a given linguistic description, although it is not evaluated in this paper. It is however not possible to produce the actual motion from the motion representation used in this article. Hellbach et al. [6] have given an example of motion representation that allows such reproduction. This work could thus be extended, by changing the motion representation, to an imitation learning setting, in which the system could be evaluated on producing gestures on a real robot, corresponding to a given linguistic description.

While in this article we focused on primitive motions active at the same time, it is possible to use the same setting to recognize choreographies where motions are composed in sequence with eventual overlaps. A direct application of our method, would however only enable reconstructing the set of active motions and not their order.

VIII. ACKNOWLEDGMENTS

This work was partly funded by ERC EXPLORERS grant 240 007. The authors would like to thank Louis ten Bosh for his help on NMF algorithms, and Haylee Fogg for her help in the data acquisition process.

REFERENCES

[1] A. A. Salah, T. Gevers, N. Sebe, and A. Vinciarelli, "Challenges of Human Behavior Understanding," *Human Behavior Understanding*, pp. 1–12, 2010.

[2] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[3] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization.," *Nature*, vol. 401, pp. 788–91, Oct. 1999.

[4] C. Févotte and A. Ozerov, "Notes on Nonnegative Tensor Factorization of the Spectrogram for Audio Source Separation : Statistical Insights and Towards Self-Clustering," in *International Symposium on Computer Music Modeling and Retrieval* (K. Jensen, S. Ystad, M. Aramaki, and R. Kronland-Martinet, eds.), no. 7, (Málaga, Spain), pp. 102–115, Springer Berlin / Heidelberg, 2011.

[5] J. Driesen, L. ten Bosch, and H. Van Hamme, "Adaptive Non-negative Matrix Factorization in a Computational Model of Language Acquisition," in *Interspeech*, pp. 1–4, 2009.

[6] S. Hellbach, J. P. Eggert, E. Körner, and H.-m. Gross, "Basis Decomposition of Motion Trajectories using Spatio-Temporal NMF," in *Int. Conf. on Artificial Neural Networks (ICANN)*, (Limassol, Cyprus), pp. 597–606, Springer, 2009.

[7] D. Kulic and Y. Nakamura, "Incremental Learning of Human Behaviors using Hierarchical Hidden Markov Models," in *IEEE International Conference on Intelligent Robots and Systems*, pp. 4649–4655, IEEE Comput. Soc. Press, 2010.

[8] V. Kruger, D. Herzog, S. Baby, A. Ude, and D. Kragic, "Learning actions from observations," *Robotics and Automation Magazine*, vol. 17, no. 2, pp. 30–43, 2010.

[9] S. Calinon, F. D'Halluin, E. L. Sauser, D. G. Caldwell, and A. G. Billard, "An approach based on Hidden Markov Model and Gaussian Mixture Regression," *IEEE Robotics and Automation Magazine*, vol. 17, no. 2, pp. 44–54, 2010.

[10] J. Butterfield, S. Osentoski, G. Jay, and O. C. Jenkins, "Learning from Demonstration using a Multi-valued Function Regressor for Time-series Data," in *International Conference on Humanoid Robots*, no. 10, (Nashville), IEEE Comput. Soc. Press, 2010.

[11] Y. Li, C. Fermuller, Y. Aloimonos, and H. Ji, "Learning shift-invariant sparse representation of actions," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (San-Francisco), pp. 2630–2637, IEEE, June 2010.

[12] E. Tuci, T. Ferrauto, A. Zeschel, G. Massera, and S. Nolfi, "An Experiment on Behaviour Generalisation and the Emergence of Linguistic Compositionality in Evolving Robots," *IEEE Transactions on Autonomous Mental Development*, vol. 3, no. 2, pp. 1–14, 2011.

[13] G. Massera, E. Tuci, T. Ferrauto, and S. Nolfi, "The Facilitatory Role of Linguistic Instructions on Developing Manipulation Skills," *IEEE Computational Intelligence Magazine*, vol. 5, no. 3, pp. 33–42, 2010.

[14] Y. Sugita and J. Tani, "Learning semantic combinatoriality from the interaction between linguistic and behavioral processes," *Adaptive Behavior*, vol. 13, no. 1, p. 33, 2005.

[15] T. Cederborg and P.-Y. Oudeyer, "Imitating Operations on Internal Cognitive Structures for Language Aquisition," in *International Conference on Humanoid Robots*, no. 11, (Bled, Slovenia), IEEE/RAS, 2011.

[16] L. F. M. ten Bosch, H. Van Hamme, and L. W. J. Boves, "Unsupervised detection of words questioning the relevance of segmentation," in *Speech Analysis and Processing for Knowledge Discovery*, ITRW ISCA, Bonn, Germany : ISCA, 2008.

[17] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover ' s Distance as a Metric for Image Retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[18] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," Tech. Rep. 23, UniversUniversität Dortmund, LS VIII-Reportität Do, 1997.

[19] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," *International Conference on Computer Vision*, p. 1470, 2003.

[20] S. Calinon and A. G. Billard, "Statistical Learning by Imitation of Competing Constraints in Joint Space and Task Space," *Advanced Robotics*, vol. 23, pp. 2059–2076, 2009.

[21] D. Kulic, H. Imagawa, and Y. Nakamura, "Online acquisition and visualization of motion primitives for humanoid robots," *Symposium on Robot and Human*, pp. 1210–1215, 2009.