Learning Semantic Components from Subsymbolic Multimodal Perception

Olivier Mangin
Flowers Team, INRIA / ENSTA-Paristech, France
Université de Bordeaux, France
olivier.mangin@inria.fr

Pierre-Yves Oudeyer
Flowers Team, INRIA / ENSTA-Paristech, France
pierre-yves.oudeyer@inria.fr

***Abstract*—Perceptual systems often include sensors from several modalities. However, existing robots do not yet sufficiently discover patterns that are spread over the flow of multimodal data they receive. In this paper we present a framework that learns a dictionary of words from full spoken utterances, together with a set of gestures from human demonstrations and the semantic connection between words and gestures. We explain how to use a nonnegative matrix factorization algorithm to learn a dictionary of components that represent meaningful elements present in the multimodal perception, without providing the system with a symbolic representation of the semantics. We illustrate this framework by showing how a learner discovers word-like components from observation of gestures made by a human together with spoken descriptions of the gestures, and how it captures the semantic association between the two.**

I. Introduction

Applications targeted for personal and social robots require them to make significant improvements in terms of human behavior understanding and communication with humans. For that they typically have to deal with language and the conceptual objects it describes, that often are spread over the several modalities (e.g. sound, vision) spanned by the robot sensors.

The recent advances in the fields of visual object recognition, automatic speech recognition, and motion recognition have brought several new techniques that enable robots to autonomously discover words from speech, gestures from motion, or objects from visual scenes. However, in order to learn natural communication with humans, a robot also needs the additional capability to relate words that appear, for example, in their acoustic sensory input, to meanings, that can be concepts emerging from perception in one or several other modalities. This problem is known as the *symbol grounding* problem [1], [2].

In order to address this problem, some approaches first learn the acoustic words and then try to associate them with meanings. Conversely, others first learn visual clusters and then use classification algorithms to learn their names. What we propose is that adequate learning of concepts should not proceed by first identifying primitives or clusters in each modality, and second learn their associations. One should rather learn primitives with respect to their correlations across modalities. For instance, visual clusters obtained by performing unsupervised clustering only in the vision modality may be very different from the visual concepts corresponding to the meaning of words. Furthermore, we put forward the idea that primitives in each modality can be discovered more easily by mutually benefiting from the context of other modalities. Also, many concepts cannot be completely characterized without grounding them on several modalities: the concept 'metallic' cannot be characterized without taking into account its perceptual expression on several modalities (e.g. visual aspect, sound, touch, or taste), together with the recognition of the spoken or written word.

Building techniques for robots and intelligent systems to discover meaningful elements from multimodal flows of data is currently an active field of research. In their seminal work, Roy and Pentland [3] present a cognitive architecture that learns concepts as pairs of clusters in the audio space (where words appear) and in the visual space (where objects appear). Yu and Ballard [4] demonstrated how a system can learn from an automatic and noisy transcript of a speech signal together with objects and actions perception. This is obtained by first clustering objects and actions into meanings; then, using cross-modal information, the system discovers word elements from its acoustic perception.

According to Roy and Arbib [5], the evolution of syntax might be grounded in the structure of actions; a few experiments therefore explore the relations between structure in motions and language. For example, Sugita and Tani [6] have built a recurrent neural networks which is capable of discovering a joint grammar based on simple object-action pairs. Similarly, Tuci et al. [7] trained artificial agents on executing actions corresponding to linguistic instructions that are object-action pairs. They show how the joint language-action representation allows the agents to generalize action knowledge to unknown object-action pairs. Massera et al. [8], in an other experiment, demonstrate that a robot can reach better performance when a linguistic guidance is present, even if it does not initially know the meaning of the symbols composing the guidance.

In our previous work [9], we presented an experiment in which a learner observes demonstrations of short dance arrangements (that we call choreographies), each composed of several simple gestures taken from a fixed repertoire and executed simultaneously. Together with each demonstration, the learner also observes a symbolic linguistic description of the choreography. After learning, the system is capable to recognize gestures and associate them to symbols, so that the learner generates its own descriptions of choreographies. Furthermore, we demonstrated that the robot learns the combinatorial structure of choreographies: after the learning phase, it can successfully generate linguistic descriptions of new

unknown choreographies composed from the same repertoire of gestures. In [9], we used a method inspired from work by Driesen, Bosch et al. [10], [11]. In [12], Driesen et al. built an system that learns in a multimodal setting: spoken utterances are observed together with an object in an image and a symbol associated with the object-word pair. Their system is then asked to reconstruct symbols associated with a new utterance, a new image, or both. Such setting is very similar to previous work from Saenko and Darell [13]. Interestingly, Lienhart et al. [14], Akata et al. [15], and BenAbdallah et al. [16] all have used similar techniques to learn from images together with a keyword or a set of keywords. Ngiam et al. [17] have used deep networks to address a similar problem: learn features on joint datasets that contain both acoustic records and records of the corresponding lip motions. They demonstrated how taking into account all the modalities can improve the quality of the learned features with respect to a supervised classification task.

In this paper, we present a framework together with experiments in which a learner simultaneously learns gestures, associated words, and the semantic connection between them. More precisely, the learner is trained by observing examples each of which is composed of a demonstration of a gesture and a sentence describing the gesture. From these observations, it learns a dictionary of multimodal components, denoted by W, that captures part of the structure of the data. Through experiments, we evaluate the information encoded in such a dictionary and how the learner can 1) recognize gestures in new motion demonstrations, 2) recognize words from new spoken utterances, and 3) relate one to the other.

In opposition to most of the related works presented above, the work presented in this paper do not use symbols, neither at some stage of the learning phase nor to test the learner. In [3], [4], no symbolic information is given to the learner but an off-line phoneme recognizer is trained and encoded in the system. In this paper, we do not use such a phoneme recognizer and we present a system that learns semantic classes only from raw multimodal information. Furthermore, in [4], objects, actions, and word clusters are computed only from respectively visual, motion, and speech modalities; then semantic classes are formed on top of these clusters. In this paper, we instead propose an approach in which linguistic elements, motion primitives, and their associations are learned in a single process.

Our approach shares similarities with [17]: we propose not to explicitly learn the association probability between models of words and gestures, but instead to learn a dictionary W of multimodal components that can be combined to explain the observed data. We demonstrate how such a dictionary of components contains information about the semantic connection between words and gestures.

II. Using nonnegative matrix factorization to learn from several modalities

In this section we present how we built a system that learns a dictionary of multimodal components to represent the observed data. Then we explain how the learned dictionary provides a representation of data that is not bound to any modality. We call it the learner's internal representation of data. Finally we explain how the learner can transform data from one or several modalities to an internal representation or to an expected representation in unobserved modalities. The setting we present was originally used by Driesen et al. in [10].

A. Representation of multimodal data

We consider a setting in which the learner observes examples in several modalities. For example, the system observes visually a demonstration of a choreography while hearing a spoken description of this choreography. We represent the perception of the example in each modality by a vector v_a, where a denotes the modality (e.g. the system observes the motions as v_m and the sound description as v_s). We detail in Sections III-A and III-B how recording of spoken utterances and choreographies can be represented in such a way. Finally the example is represented as the concatenation of its representation in each modality. If motions and sound are perceived, the example is represented by a vector v defined as:

$$v = \begin{pmatrix} v_m \\ v_s \end{pmatrix}.$$

In the following we denote by d the dimension of the vectors v, which is the sum of the dimensions of the representation in each modality. We denote by n the number of examples. In this paper representations of data in each modalities only contain nonnegative values.

B. Learning a dictionary of multimodal components

We call *components* primitive elements that are mixed together into observations, in the same way phonemes can be seen as mixed together into a word or a sentence. Compared to the common context of clustering, This notion of component is more general than the one of centroids: observations are mixtures of several components at the same time instead of being just a noisy observation of one centroid.

The learner presented in this paper builds a dictionary of multimodal components according to the following model: it searches k components, each represented by a vector w_j (j from 1 to k), such that each observed example v^i verifies:

$$v^i \simeq \sum_{j=1}^{k} h_i^j w^j \tag{1}$$

where h_i^j are coefficients and $\simeq$ denotes a notion of similarity between matrices that is defined below. This is equivalent to clustering when the w_j are the centroids and for each i only one h_i^j is nonzero and equals 1. We consider a more general case where w_j and h_i^j are only constrained to be nonnegative.

In the following, the set of n examples is represented by a matrix v of shape $d \times n$ (each example is a column of V), the set of components by a matrix W of shape $d \times k$, called dictionary, and the coefficients by a matrix H of shape $k \times n$. The previous equation which models the objective of our learner can thus be re-written as:

$$V \simeq W \cdot H \tag{2}$$

Matrix factorization [18], [19] is a class of machine learning techniques that can be used to learn a dictionary represented by a matrix W together with a coefficient matrix H

such that the error between V and its reconstruction as $W \cdot H$ is minimized. More precisely nonnegative matrix factorization (NMF) is a set of algorithms specific to the case we consider in this paper, where data is represented by nonnegative values and matrices W and H are constrained to have nonnegative values.

In order to fully define the reconstruction error between V and $W \cdot H$, we use a variant of the Kullback-Leibler divergence often called generalized Kullback-Leibler or I-divergence. The Kullback-Leibler divergence is originally a information theoretic measure of similarity between probability distributions. The I-divergence is defined, for two matrices A and B of same shape, as $D_I(A\|B)$ given by equation (3).

$$D_I(A\|B) = \sum_{i=1}^{d}\sum_{j=1}^{n}\left(A_{i,j}\ln\left(\frac{A_{i,j}}{B_{i,j}}\right) - A_{i,j} + B_{i,j}\right) \quad (3)$$

In this paper in order to minimize $D_I(V||W \cdot H)$, we use the algorithm, based on multiplicative updates of W and H, that was originally presented in Lee and Seung's paper [19]. This algorithm consist in alternating the two update steps from equation (4) where $\circledast$ and $/$ denote Hadamard's (coefficient-wise) product and division on matrices. More details can be found in [19].

$$H \leftarrow H \circledast \frac{W^T \dfrac{V}{W \cdot H}}{W^T \cdot \mathbf{1}} \qquad V \leftarrow V \circledast \frac{\dfrac{V}{W \cdot H} H^T}{\mathbf{1} \cdot H^T} \quad (4)$$

The I divergence that is minimized in the factorization induces a trade-off between error in one modality relatively to others. In order for the error in each modality to be treated on a fair level by the algorithm it is important that the average values in the representations are of similar magnitude. It can be easily obtained by normalizing data in each modality. In the experiment we normalize data in each modality according to its average L_1 norm.

C. NMF to learn mappings between modalities

For a given set of observations from several modalities that is represented by a matrix V, the NMF algorithm can learn a dictionary W and a coefficient H matrices such that training examples are well approximated by the product $W \cdot H$.

Since the examples (i.e. the columns of V) are composed of several modalities, the dictionary W can also be split into several parts each corresponding to one modality. That is to say each components can be seen as the concatenation of several parts: one for each modality. For example if the data is composed of a motion and a sound part, there exist matrices W_m and W_s such that:

$$W = \begin{pmatrix} W_m \\ W_s \end{pmatrix}.$$

In the following we interpret the columns of the matrix H, as an internal representation of the data by the learner. For example, an internal representation h is induced by the observation of a motion such that $v_m = W_m h$ or a sound such that $v_s = W_s h$ or a multimodal example such that $v = Wh$. Also, for a given internal representation h we say that our learner expects the observations given by the previous formulae.

Interestingly, it is possible to use the learned dictionary to compute an internal representation of an example, even if the example is only observed in a subset of the modalities. Given an example observed only in one modality, for example v_s representing sound, one can search for an h such that v_s is well approximated as $W_s h$. More precisely this is equivalent to finding an h solution of:

$$\arg\min_h \; D_I(v_s, W_s h) \quad (5)$$

The NMF algorithms we use actually alternates steps minimizing $D_I(V||W \cdot H)$ with respect to W and H. Solving equation (5) is equivalent to the NMF problem with respect to H only; therefore, it can be obtained with the same algorithm, but only using the steps that update H. This approach also scales to the situation where several modalities are observed: the dictionary to be used is obtained by stacking vertically all the dictionaries corresponding to the observed modalities.

Finally it is also possible to reconstruct a representation of the data that the system would expect in a modality, given observations in other modalities. For that, from an observation featuring a subset of the modalities, the system fits an internal representation h using the method described previously. Then it can reconstruct the expected representation in an unobserved modality (e.g. motion) by computing the product $W_m h$. This forms a framework that uses a learned multimodal dictionary to transform data from modalities to internal representations or expected data in other modalities. It enables a large set of experiments as illustrated in Section IV.

III. Data and representation

The motion dataset was recorded from a single human dancer with 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 together with angle values representing the dancer position at a specific time.

We recorded a motion dataset composed of a thousand examples of ten dance gestures, illustrated in Figure 1 and Table I. The gestures 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 the learner initially. They correspond to both discrete and rhythmic movements.

We also use a dataset of sounds from the Acorns project, called *Caregiver dataset* [20][2]. It is a set of sentences in infant directed speech in which each sentence actually contains a keyword taken from a set of ten keywords. For example the sentence "Are you reading a book?" includes the keyword "book" and "Daddy reads on his vacation" contains the keyword "Daddy".

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

[2] In the presented experiments we use the thousand English records from the first speaker among the year one (Y1) records.

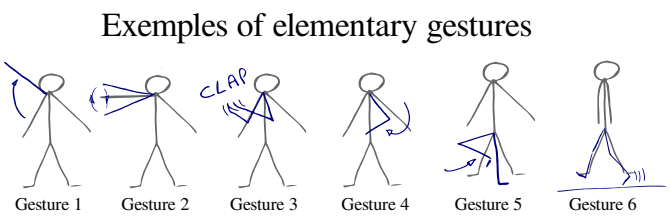

Fig. 1: Illustration of the demonstrated gestures (these are examples, not all of them are used in the following experiments).

TABLE I: List of associations between keywords from the acoustic dataset (names) and gestures from the motion dataset. The limbs on which the motions occur are also mentioned.

Name	Limb(s)	Motion
shoe	both legs	squat
nappy	both legs	walk
book	right leg	make a flag/P on right leg
daddy	both arms	clap
mummy	both arms	mimic paddling left
Angus	right arm	mimic punching with right arm
bath	right arm	right arm horizontal goes from side to front
bottle	left arm	horizontal left arm, forearm goes down to form a square angle
telephone	left arm	make waves on left arm
car	left arm	say hello with left arm

To investigate the learning of semantic connections between the language and motion modalities, we use an artificial mapping between acoustic words and gestures. The mapping is given in Table I. Each demonstration of gesture is paired to exactly one record of a sentence and conversely.

A. *Representation of choreographies*

One modality used in the following experiments is composed of observations of choreographies through a motion capture system.

We assume that motions are captured as trajectories in angle space, for each body joint. In order to capture more information angular velocities are also extracted: a delayed velocity is used to achieve better robustness to noise in the angle sequences. More precisely $\dot{x}_t = x_t - x_{t-\delta}$ is used to compute the velocities, instead of being restrained to the case where $\delta = 1$. It is not necessary to divide by the fixed time step since the histogram representation described below is invariant to scaling all the data by the same amount.

Then 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 nonnegative vector. Vectors obtained for each degree of freedom are then concatenated into a larger vector. The following explains how to build the histogram representation.

The transformation of angles and velocities sequences into histograms is done by first dropping the sequential information and then performing vector quantization over the joint position-velocity samples. A two-dimensional histograms is then built on the joint angle-velocity space (see Figure 2).

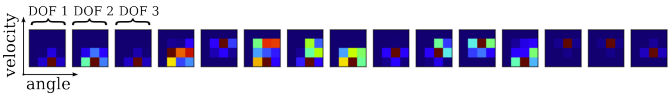

Fig. 2: Illustrations of histograms on joint positions and velocities. Frequencies are represented through colors, x and y axis correspond respectively to values of angles and velocities. (Best seen in color)

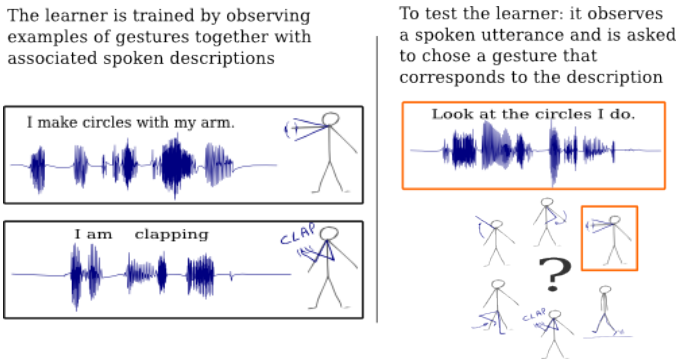

Fig. 3: Illustration of the cross-modal classification task on which the learner presented in this paper is tested

More precisely, vector quantization (VQ) is performed through a k-means algorithm. Finally a histogram is built by counting the proportion of samples falling into each cluster. The choice of this representation is discussed in more details in [9].

B. *Representation of sound*

We use the *Histogram of Acoustic Co-occurence* (HAC) representation presented by Van Hamme in [21]. This representation is based on Mel-Frequency Cepstral coefficients (MFCC), computed from the spoken utterances. From the MFCC, first and second order time derivative are also computed. In a second step the MFCC and its derivative are split at multiple time scales into small chunks and a vector quantization process is applied to the chunks so that each utterance is represented by a sequence of discrete *events* each corresponding to an occurrence of a chunk from a cluster. Finally the successive co-occurrences of pairs of events are counted and the final representation of the utterance is an histogram over the occurrences and co-occurrences of events. A more detailed presentation of the process can be found in [21].

Representations of utterances obtained from this process have two important properties. First, they are very high dimensional and very sparse. Also, their construction gives them the property that if two words were represented by w_1 and w_2, the sentence formed by concatenating these two words would have a representation close to $c_1 w_1 + c_2 w_2$, c_1 and c_2 being two coefficients.

IV. Experiments

In this section we present experiments that test the ability of our system to recognize gestures, words and to capture their

semantic association. We also show how this system can be used to solve a more traditional classification task and what is the impact of the various modalities in solving such a task. Finally we illustrate how some coefficients of the internal representation specialize in one semantic concept. The data is represented by vectors of dimension 270 for the motion part and 110002 for the sound part. When not mentioned otherwise, the learning system uses a value of $k = 50$ components. The training sets contains 890 examples and the testing sets 100.

A. Cross-modal classification

The learner does not have access to the semantic classes of motions and sentences as symbols. Therefore it cannot be directly evaluated on a regular classification task, as usually considered in machine learning. However, it can be evaluated on a classification task similar to the ones faced by children that typically do not directly produce a symbol out of their brain but are rather asked to choose a symbol from a set of examples or to produce a vocal or written instantiation of such a symbol. Similarly our learning system does not have direct means for producing symbols, however it can choose an example in one modality that corresponds to an observation in an other modality. For example, the system can be asked, given an acoustic description, to choose a gesture from a set of examples, that best matches the description. Figure 3 illustrates this task.

We train a learner on joint examples of motions and acoustic descriptions, a process which results in a multimodal dictionary. We evaluate the learner in two different manners. The first one consists in presenting a motion demonstration that we call the *test example* to the learner and then ask it to choose the best fit for the observed motion between a set of linguistic descriptions, called *reference examples*. The evaluation is considered successful if the chosen reference example (a linguistic description) is of the same semantic class than the test example (which is the demonstrated motion). We also consider the symmetric setting where the test example is an acoustic description and the reference examples are motion demonstrations. In both cases, neither the test nor the reference examples are encountered by the system during the training phase. We consider experiments in which one reference example for each class is provided.

We use the approach presented in Section II-C as a basis to implement a classification behavior for the learner. For a given example the system uses the learned multimodal dictionary to produce an internal representation of the example (coefficients h) and eventually also an expected transcription of this example in an other modality. It then compares an example from the test modality to those of the reference modality by either:

- compute an internal representation of the test example, compute internal representations of the reference examples, and then compare these internal representations.
- compute an internal representation of the test example, use it to generate an expected representation in the reference modality, and compare it to the reference examples.
- compute internal representations of reference examples, for each of them compute an expected representation in the test modality, and compare then the test example.

TABLE II: Performance on the multimodal matching task

Modality			Score		
Test	Reference	Comparison	KL	Euclidean	Cosine
Sound	Motion	Internal	0.608	0.612	0.646
Sound	Motion	Motion	0.552	0.379	0.444
Sound	Motion	Sound	0.238	0.126	0.208
Motion	Sound	Internal	0.610	0.704	0.830
Motion	Sound	Sound	0.106	0.090	0.186
Motion	Sound	Motion	0.676	0.642	0.749

(a) Scores of recognition of the right reference example from a test example. The values are given for many choices of the reference test and comparison modalities and various measures of similarity.

Modality			Score		
Test	Reference	Comparison	KL	Euclidean	Cosine
Sound	Motion	Internal	0.387	0.699	0.721
Sound	Motion	Motion	0.543	0.261	0.424
Sound	Motion	Sound	0.136	0.089	0.131
Motion	Sound	Internal	0.573	0.620	0.702
Motion	Sound	Sound	0.114	0.090	0.122
Motion	Sound	Motion	0.519	0.469	0.552

(b) Same scores as previously but with a learner that observed symbolic labels representing the semantic classes during training.

The choice of one of these method is referred as the modality of comparison.

Finally, best matching reference example can be chosen according to various metrics. In the following results we consider comparison with respect to Kullback-Leibler divergence, Euclidean norm, and cosine similarity. Table IIa presents results for this experiment. A system trained on examples built by random association of gesture demonstrations and acoustic descriptions, that is to say on a dataset where no semantic semantic association exists between the two modalities, generally scores around 0.11[3]. The results demonstrate that the system has captured semantic associations between gestures and spoken descriptions. They outline big differences between the modality where the comparison occurs. This is not surprising since the representation on the various modalities have very different structures and dimensionality (110002 for sound, 270 for motions and 50 for the internal coefficients).

B. Supervised classification

In [9], [10], [12] the NMF algorithm is used to perform supervised classification. This is possible by adding a symbolic modality to the setting. The symbols are actually the labels of the semantic classes. They are represented as vectors of binary values, which length are the number of classes. A gesture or a sentence of semantic class i would be represented by the vectors of zeros with a one at position i. We added such a symbolic modality to the training data to form the two following experiments. The goal of these experiments is to 1) compare the multimodal learner to one that observes symbolic information, 2) study whether observing several modalities during training is beneficial for the classification task.

First we reproduced the experiment from previous section but added a symbolic modality to the training data, in addition

[3]This is not 0.1 because the distribution of sound examples from the Caregiver dataset is not exactly uniform.

TABLE III: Performance on reconstructing symbolic labels from various modalities

Training	Testing	Score	Standard deviation
$S+L$	$S \rightarrow L$	0.916	0.034
$M+L$	$M \rightarrow L$	0.906	0.052
$S+M+L$	$S \rightarrow L$	0.896	0.043
$S+M+L$	$M \rightarrow L$	0.910	0.054
$S+M+L$	$S+M \rightarrow L$	0.917	0.055

to sound and motion. The results are presented in Table IIb. That the learner can observe the semantic classes in the symbols during training suggests that performance should improve. Surprisingly this modification has mitigated effects and no really significant increase in performances is observed.

We also compared the various combinations of modality observed during training and testing in a supervised classification task. The learner is either trained with sound, denoted by S, and symbols, denoted by L as labels, or with motion (M) and labels, or with the three modalities. The three learning situations are denoted by $S+L$, $M+L$ and $M+S+L$.

Once trained the learner is given an example from a modality and computes the expected label for this example by using the technique presented in Section II-C. We then test the accuracy of the reconstruction of the symbolic modality. The system does not directly yield a label but instead an expected value of the symbolic modality, that is to say a vector of coefficients. To evaluate such a vector, the position of its maximum coefficient is compared with the ground truth label. For example, the $S+L$ learner is tested on reconstruction of labels from sound examples. The task is then denoted by $S \rightarrow L$. The $M+L$ learner is tested on the $M \rightarrow L$ task. Finally, the $M+S+L$ learner is tested on three tasks: $S \rightarrow L$, $M \rightarrow L$ and $M+S \rightarrow L$. The results are presented in Table III where the score corresponds to the ratio of correctly reconstructed labels.

Interestingly training with the two modalities (sound and motion) does not significantly change the performance of the learner, and that when tested on sound, motion or both. In that case the benefit of having two non-symbolic modalities is not an increase in performance, but rather that the same learner can use either acoustic perception and / or motion perception to classify an example.

C. Emergence of concepts

In the beginning of this section we evaluated the learner on concrete tasks that emphasis its ability to relate information from one modality to an other. A natural question that follows is whether the learner develops an internal representation of concepts associated to the semantic classes from the data, although it never observes the symbolic information.

The question is actually non-trivial since it is not immediate to interpret the internal representation that the system builds, that is to say, the role of the various components of the dictionary matrix. However some insight can be gained that suggests that at least some components are more specialized into some of the semantic classes.

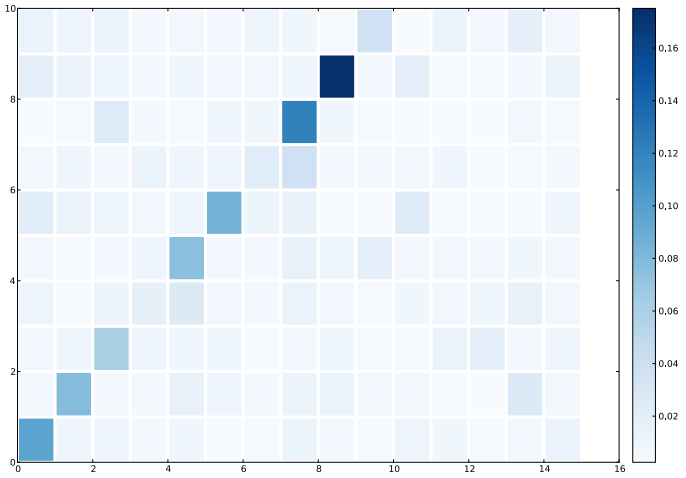

Fig. 4: Illustration of the specialization of some components with respect to some semantic labels. The figure represents the mutual information between (vertically) semantic classes (that are not observed by the learner) and (horizontally) each internal coefficient used by the learner to represent pairs of motion demonstration and acoustic descriptions from the training set. A value of $k = 15$ was used in this experiment.

Figure 4 illustrates this effect by representing, for each semantic class and each coefficient of the internal representation, the mutual information between the belonging of examples to that class and the value of a given coefficient of the internal representations of these examples. To emphasis the specialization of some internal coefficients we re-ordered internal coefficients so that classes and coefficients that have high mutual information are aligned.[4]

V. Discussion and perspectives

We presented a framework that enables a system to learn a joint representation over data from several modalities without observing symbolic information. We showed through experiments how the system is capable to recognize gestures together with words from full spoken utterances and to learn the semantic associations between gestures and words. Finally we illustrated how information about a semantic classes emerges from the structure of the learned representation. The experiments presented in this paper demonstrate that the system can learn to relate information from two very different and complex modalities, speech and gestures without the need to introduce symbolic information neither during training nor testing. Interestingly, our system does not directly optimize an explicit criterion representing the word-gestures associations, as in [3], [4]. Instead the system optimizes a dictionary of multimodal components that lead to the best reconstruction of training data. We show that such criterion is sufficient to learn semantic associations between words and gestures and to lead to the emergence of a representation of the semantic classes. However an interesting direction to explore is to introduce explicit criteria, for example to force the system

[4]The best alignment was computed by a Kuhn-Munkres algorithm and the mutual information are computed from a discretized representation of the distribution over coefficient values.

to find components that maximize the encoded dependency between modalities, and evaluate their impact.

Our work can also easily be extended to more complex settings with other modalities, as vision of objects. Such a setting is presented by [12], [17] but only tested with symbolic labels. Also, in this paper we only considered demonstrations of single gestures but the acoustic descriptions are full sentences containing several words. We have shown in our previous work [9] that our framework can be used to learn gestures from complex choreographies where the gestures are only observed mixed together. A natural extension consists in testing the ability of the learner to discover the semantic associations between gestures and words that are observed in complex examples, that is to say several primitive gestures are mixed together in the motion demonstrations and must be related to different words from a spoken description.

Finally several ways can be explored to further study the emergence of representations encoding the semantics of the data. For example one could enforce stronger structural properties on the matrix factorization (e.g. sparseness of the representation) or combine several layers of learning.

ACKNOWLEDGEMENT

The authors would like to specially thank Louis ten Bosch for providing detailed explanation on his previous work on NMF with language and direct help for producing HAC features from the ACORNS Caregiver dataset.

REFERENCES

[1] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1, pp. 335–346, 1990.

[2] A. M. Glenberg and M. P. Kaschak, "Grounding language in action." *Psychonomic bulletin & review*, vol. 9, no. 3, pp. 558–65, September 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12412897

[3] D. K. Roy and A. P. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive science*, vol. 26, no. 1, pp. 113–146, January 2002. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0364021301000611

[4] C. Yu and D. H. Ballard, "A Multimodal Learning Interface for Grounding Spoken Language in Sensory Perceptions," *Transactions on Applied Perception*, no. 1, pp. 57–80, 2004.

[5] A. C. Roy and M. A. Arbib, "The syntactic motor system," *Gesture*, vol. 5, no. 1, pp. 7–37, January 2005.

[6] 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. [Online]. Available: http://adb.sagepub.com/cgi/content/abstract/13/1/33

[7] 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.

[8] 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.

[9] O. Mangin and P.-Y. Oudeyer, "Learning to recognize parallel combinations of human motion primitives with linguistic descriptions using non-negative matrix factorization," in *International Conference on Intelligent Robots and Systems (IROS 2012)*. Vilamoura, Algarve (Portugal): IEEE/RSJ, 2012.

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

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

[12] J. Driesen, H. Van Hamme, and B. W. Kleijn, "Learning from images and speech with Non-negative Matrix Factorization enhanced by input space scaling," in *IEEE Spoken Language Technology Workshop (SLT)*. Berkeley, California, USA: IEEE, 2010, pp. 1–6.

[13] K. Saenko and T. Darrell, "Object Category Recognition Using Probabilistic Fusion of Speech and Image Classifiers," in *Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms - MLMI*, no. 4, Brno, Czech Republic, 2007.

[14] R. Lienhart, S. Romberg, and E. Hörster, "Multilayer pLSA for multimodal image retrieval," in *International Conference on Image and Video Retrieval -CIVR*, no. April. New York, New York, USA: ACM Press, 2009, p. 1.

[15] Z. Akata, C. Thurau, and C. Bauckhage, "Non-negative Matrix Factorization in Multimodality Data for Segmentation and Label Prediction," in *Computer Vision Winter Workshop*, no. 16, Mitterberg, Autriche, 2011. [Online]. Available: http://hal.inria.fr/hal-00652879

[16] J. BenAbdallah, J. C. Caicedo, F. A. Gonzalez, and O. Nasraoui, "Multimodal Image Annotation Using Non-negative Matrix Factorization," in *International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE / WIC / ACM, August 2010, pp. 128–135. [Online]. Available: http://www.informed.unal.edu.co/jccaicedo/papers/wic10.pdf

[17] J. Ngiam, A. Khosla, M. Kim, J. Nam, and A. Y. Ng, "Multimodal deep learning," in *International Conference on Machine Learning*, no. 28, Bellevue, Washington, USA, 2011. [Online]. Available: http://ai.stanford.edu/~ang/papers/icml11-MultimodalDeepLearning.pdf

[18] 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.

[19] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization." *Nature*, vol. 401, no. 6755, pp. 788–91, October 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10548103

[20] T. Altosaar, L. ten Bosch, G. Aimetti, C. Koniaris, K. Demuynck, H. van den Heuvel, S. Proc, P. O. Box, and F. Tkk, "A Speech Corpus for Modeling Language Acquisition: CAREGIVER," in *Language Resources and Evaluation - LREC*, 2008, pp. 1062–1068.

[21] H. Van Hamme, "HAC-models: a Novel Approach to Continuous Speech Recognition," in *Interspeech ISCA*, 2008, pp. 2554–2557.