Learning Semantic Components from Subsymbolic Multimodal Perception

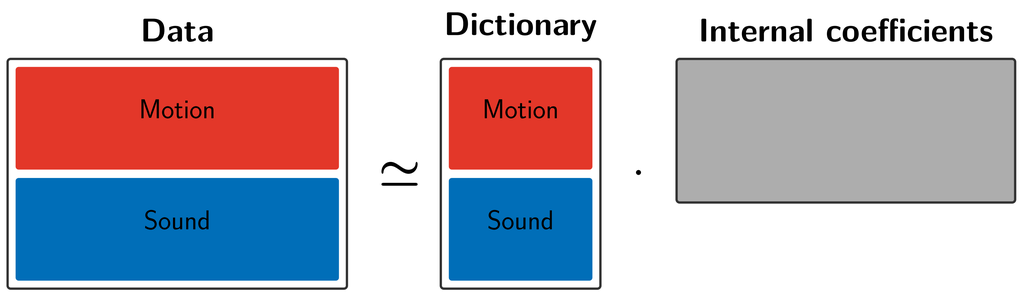
Olivier Mangin and Pierre-Yves Oudeyer

Flowers Team — INRIA / ENSTA-ParisTech / Université de Bordeaux

1. Introduction

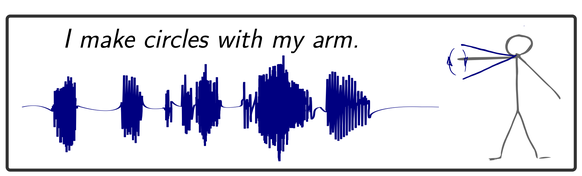
Perceptual systems often include sensors from several modalities. However, existing robots do not yet sufficiently discover patterns that are spread over the flow of multimodal data they receive. We present a framework based on nonnegative matrix factorization and demonstrate that it enables a system to learn to recognize multimodal semantic concepts without observing symbolic labels. More precisely, the system 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.

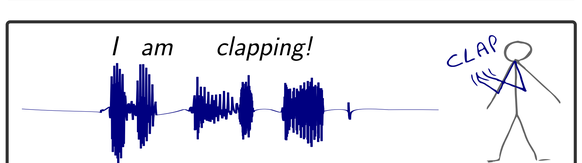
2. Multimodal learning with NMF



3. Experimental setup

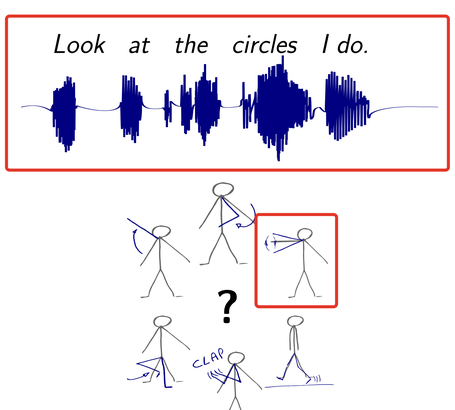
The learner is trained by observing a set of examples of gestures each of which is paired with a spoken descriptions of the gesture.





The dataset contains 10 semantic classes, that is to say 10 distinct gestures and 10 keywords.

During testing, it hears a new spoken utterance and is asked to chose a gesture from a new small set of examples that best fit to the description.



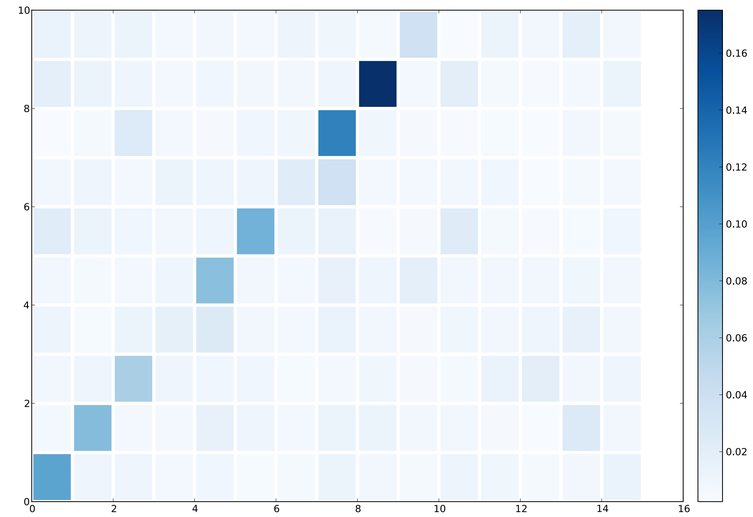
4. Results

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

Modality			Score Trained without symbol			Score Trained with symbols		
Test	Reference	Comparison	KL	Euclidean	Cosine	KL	Euclidean	Cosine
Sound	Motion	Internal	0.608	0.612	0.646	0.387	0.699	0.721
Sound	Motion	Motion	0.552	0.379	0.444	0.543	0.261	0.424
Sound	Motion	Sound	0.238	0.126	0.208	0.136	0.089	0.131
Motion	Sound	Internal	0.610	0.704	0.830	0.573	0.620	0.702
Motion	Sound	Sound	0.106	0.090	0.186	0.114	0.090	0.122
Motion	Sound	Motion	0.676	0.642	0.749	0.519	0.469	0.552

5. Emergence of semantic concept representation

Some components features a specialisation with respect to some semantic labels. 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 is displayed here. A value of $k = 15$ was used in this experiment. (Internal coefficients and semantic classes were aligned to display the best associations on the diagonal)



6. Conclusion

We present a framework that enables a system to learn a joint representation over data from several modalities without observing symbolic information. Interestingly, our system does not directly optimize an explicit criterion representing the word-gestures associations, but only the signal reconstruction.

This work can easily be extended to more complex settings with other modalities, as vision of objects or testing the ability of the learner to discover the semantic associations between gestures and words that are observed in complex examples (several primitive gestures are mixed together in the motion demonstrations).