Inverse reinforcement learning to control a robotic arm using a Brain-Computer Interface

Laurent Bougrain, Matthieu Duvinage, Edouard Klein

Inverse reinforcement learning to control a robotic arm using a Brain-Computer Interface

Laurent Bougrain[1,2], Matthieu Duvinage[3], *Research Fellow, FNRS,* Edouard Klein[1,4]

***Abstract*—The goal of this project is to use inverse reinforcement learning to better control a JACO robotic arm developed by Kinova in a Brain-Computer Interface (BCI). A self-paced BCI such as a motor imagery based-BCI allows the subject to give orders at any time to freely control a device. But using this paradigm, even after a long training, the accuracy of the classifier used to recognize the order is not 100%. While a lot of studies try to improve the accuracy using a preprocessing stage that improves the feature extraction, we work on a post-processing solution. The classifier used to recognize the mental commands will provide as outputs a value for each command such as the posterior probability. But the executed action will not only depend on this information. A decision process will also take into account the position of the robotic arm and previous trajectories. More precisely, the decision process will be obtained applying an inverse reinforcement learning (IRL) on a subset of trajectories specified by an expert. At the end of the workshop, the convergence of the inverse reinforcement algorithm has not been achieved. Nevertheless, we developed a whole processing chain based on OpenViBE for controlling 2D-movements and we present how to deal with this high dimensional time series problem with a lot of noise which is unusual for the IRL community.**

***Index Terms*—Inverse reinforcement learning, Brain-Computer Interfaces, Motor imaginery, Robotic arm**

I. Introduction

Brain-Computer interfaces (BCI) [1] interpret brain activity to produce commands on a computer or other devices like a robotic arm (see figure 1). A BCI therefore allows its user, and especially a person with high mobility impairment, to interact with its environment only using its brain activity.

A major difficulty to properly interpret the mental command lies in the fact that brain activity is very variable even if a particular task is reproduced identically. Beyond the noise acquired by the recording system, background brain activity, concentration, fatigue or medication of the subject are the source of this variability. This variability makes it difficult for the classifier to recognize the different mental commands. Specific preprocessings such as common spatial pattern filter [2] are useful to help distinguish the mental command. However, this effort is not always sufficient. It therefore becomes necessary to explore new solutions to address this variability.

[1]Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
[2]Inria, Villers-lès-Nancy, F-54600, France
[3]TCTS Lab, University of Mons, Belgium
[4] Supelec, Metz, France

Thus, it is now necessary to make decision systems able to deal with this variability. This is why some projects introduce a reinforcement learning in their BCI system such modifying the classifier [3]. We propose to use reinforcement learning in a broader context.

In this project we studied how a reinforcement learning can improve the control of a robotic arm. More precisely, the decision process will take into account a subset of trajectories specified by an expert and the position of the robotic arm in addition to the usual outputs of the mental commands classifier.

II. Methods

The goal of this study is to present the possible improvement on command recognition obtained by a post-processing performed by an inverse reinforcement learning algorithm. In this section, we first present the almost standard processing chain we used to obtain four different commands using motor imagery. Then we present how inverse reinforcement learning can help to better identify the mental order provided by the user.

A. A BCI system based on motor imagery

For controlling a neuroprosthesis of the upper limb several options are available nowadays. Firstly, the neural activity in the arm/hand area of the motor cortex can be directly recorded and decoded using invasive [4] or noninvasive electrodes ([5], [6]). But it is also possible using noninvasive electrodes to exploit various physiological phenomena such as sensorimotor rhythms, event-related desynchronization/event-related synchronization, event-related potential or steady-state visual evoked potentials. In particular, motor imaginary [11] can be used to control a 2D cursor ([7], [8]) or perform a 3D control [9]. They can even be combined in a hybrid BCI [10]. We selected motor imagery for several reasons: i) intending to produce a real movement is more natural for controlling a neuroprosthesis, ii) no additional device is needed to produce stimulations if used in a self-paced mode [12] iii) it has been already used successfully with healthy people [13] and patients [14] and iv) it can be used for rehabilitation [15]. Nevertheless, the number of commands is small (two or three usually); the information transfer rate is slow (1 action per 8s); and the accuracy is not very high (80 %).

We used motor imagery (MI) in a system-paced BCI. Having a self-paced BCI is not essential for this study and it is technically easy to switch from one mode to the other.

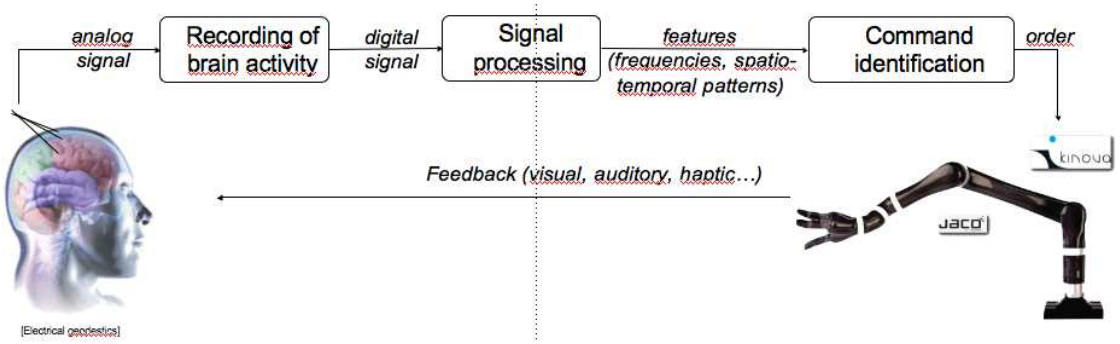

Fig. 1. The Brain-Computer Interface loop : from electroencephagraphic signals acquisition, feature extraction and classification to feedback. Our project will add a decision process based on an inverse reinforcement learning in the command identification module.

We defined a standard processing chain for motor imagery based on the parameters used for the Graz paradigm. We want to identify four commands corresponding to four motor imageries: left hand, right hand, both hands and feet. These four MI will allow us to control a robotic hand in a 2D horizontal space using respectively left, right, forward and backward commands [16].

We used a conventional montage for MI when applying a preprocessing based on common-spatial filters [17], [18], [2]. Then, among various possible classifiers to detect the MI [19], we selected linear discriminant analysis for its stability. More details are presented in the following sections.

1) Signal acquisition: We used a TMSi Refa amplifier with 32 EEG channels. We only selected 13 electrodes : Fz, FC5, FC1, FC2, FC6, C3, Cz, C4, CP5, CP1, CP2, CP6, Pz (see Fig. 2) located according to a layout 10/10 on a WaveGuard 32 channel sintered Ag/AgCl. This system use a AFz ground and a common average. We used a sampling rate of 512 Hz.

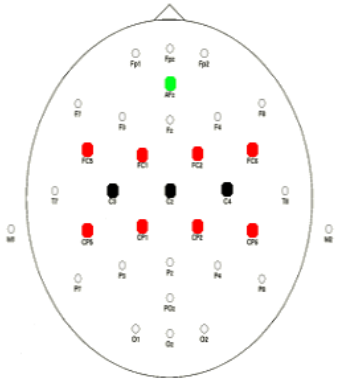

Fig. 2. Position of the selected electrodes for motor imagery of left hand, right hand, both hands and feet. The green electrode corresponds to the ground, the black ones are the main locations of our motor imageries and the red ones are useful for common spatial patterns.

2) Pre-processing: We used a 4th order Butterworth band-pass filter 8-30 Hz to only keep *mu* and *beta* bands.

Then we applied a Common Spatial Pattern (CSP). This filter takes into account the distribution of each class of a two-classes classification. The variance of the filtered signal is maximal for one class and minimal for the other class. Thus, we want to extremize using generalized eigen value decomposition:

$$J(w) = \frac{wX_1X_1^Tw^T}{wX_2X_2^Tw^T} = \frac{wC_1w^T}{wC_2w^T}$$

where X_i is the multichannel EEG signals from class i, C_i is the EEG spatial covariance matrix for class i and w is the spatial filter to optimize.

We obtained features $f = \log(wCw^T)$.

3) Motor imagery paradigm: Figure 3 presents our timing for motor imagery. Each session contains 20 trials per class. After the presentation of the cue, we analysis the signal for 3,5 seconds. The features are extracted for a 1-s period. We use a sliding window of 100 ms to repeat the analysis and confirm the decision of the classifier using a vote.

4) Classifier: For discriminating four motor imageries, we combined one-versus-all linear discriminant classifiers (one per class). In case of ambiguity, the longest distance to the separation plane shows the winner class.

5) Device: By default, the JACO arm can be controlled using a joystick. An API by Kinova is available to read sensors and send commands of movement for a specific direction and a specific duration. This API provides a virtual joystick. This mode of operation does not make it possible to specify the final position of the arm. Thus, interacting with the JACO arm via the API necessitates the definition of elementary movements (right, left, forward, backward, up.). The VRPN protocol already implemented in OpenViBE is a natural candidate to control the arm. Thus, we used a VRPN client/server using predefined action IDs which can be interpreted by the JACO arm as virtual joystick commands but sent through our application. The recording features also supports the recording of VRPN clients' commands.

B. Inverse Reinforcement Learning

Inverse Reinforcement Learning (IRL) is the problem of eliciting a succinct description of a task from demonstrations by an expert [20]. This succinct description of the task can then be used to train an agent in order to make it imitate the expert.

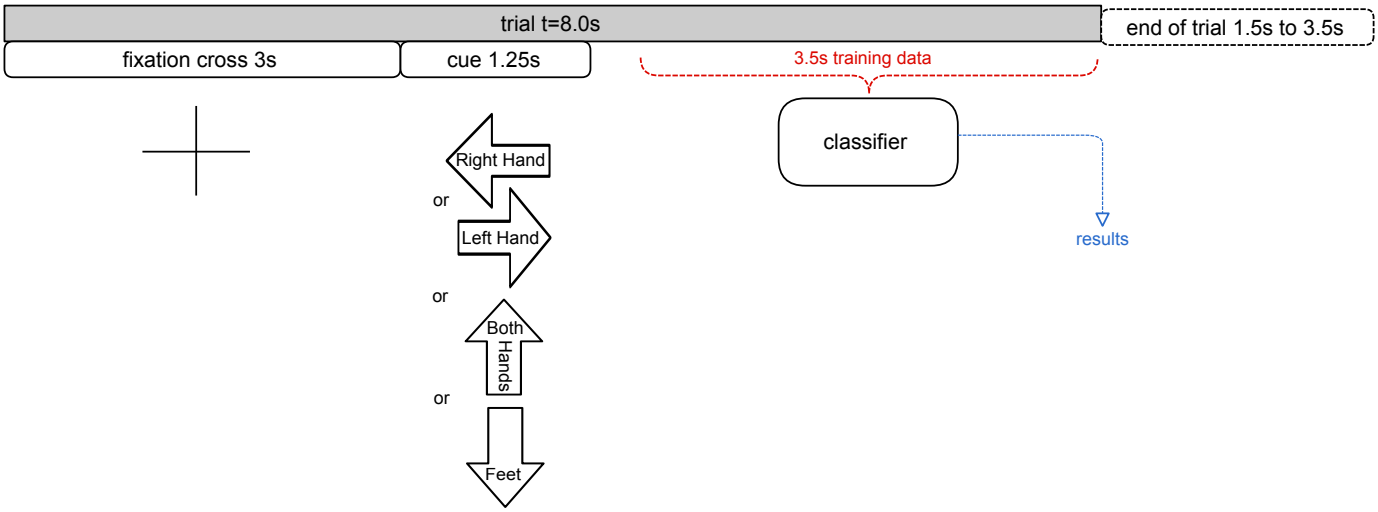

Fig. 3. Timing used for the motor imagery paradigm.

More formally, IRL assumes that an expert is acting optimally in an Markov Decision Process (MDP)[21] and seeks the reward function for which this expert is optimal. As noted in the existing literature, this is an ill-posed problem in the Hadamard sense. However, recent advances [22] in the domain may make solving the IRL problem on large or complex tasks feasible.

In our setting, we would like to use IRL to alleviate the problem of accuracy in order recognition from BCI signals. By using information about previously recognized commands and learning from human-labelled movement sequences, it should be feasible to gain a certain consistency in the overall arm movement. To put it in another way, after seeing a few examples of the arm moving in a direct manner from point A to point B, one is unlikely to admit a command that make the arm flail in seemingly random directions.

Using hand-labelled arm trajectories as expert demonstrations, we wish to extract a reward function that could be used to train an agent to recognize commands from BCI signals. The main challenges behind this task are the difficulty of finding a suitable MDP setting for casting the problem, the high dimensionality of BCI signals, the sparsity of data for both reward function inference and its optimization by an agent once the expert's actions have been analyzed by the IRL algorithm.

One of the main assumption of IRL is that the expert is acting optimally in an MDP with respect to an unknown reward function. Our goal when choosing a MDP setting for our experiment is to try to make that assumption hold. In previous test for the algorithm we used, the expert was explicitly created from a reward function and an MDP. Although the reward function was unknown to the IRL algorithm, it existed. Sharing the same MDP as the expert is one of the basic assumption made by the analysis of our algorithm. In this setting, however, the so-called expert is an omniscient agent as the path the arm followed was fixed in advance and the operator only had to follow it. There may or may not exist a MDP describing the process. We tried more than one characterization of the problem, discovering various flaws, and understanding better and better the subtleties of the exercise as we went on. This is described in the next section.

BCI signals typically are high dimensional time series with a lot of noise. From a signal processing perspective they are quite a challenge. This is very unusual for the IRL community who is more used to toy problems (although impressive applications have been published [23]) where the dimension is low and the observation perfect. IRL can be applied to partially observable environments, although it is not the direction we wish to take here as it has its own set of challenges, mainly related to computation cost explosion. The high dimensionality problem has been circumvented by the use of SCIRL, a new IRL algorithm that among other advantages is quite fast to run. The low signal to noise ratio, however, is at the heart of our problem and the very reason for the existence of this project. It raised its lot of problems when trying to come up with a reasonable MDP setting.

The model for our system being unknown yet (although modeling the brain have been promised over and over by sci-fi authors, it is not yet within the reach of a one-month project) we had to rely on sampling to make things work. This means that we had to rely on expert demonstration only to retrieve a reward function and to optimize it. Reward inference from expert data only is one of the marketed features of SCIRL. Having access to samples drawn by a random policy is one of the many ways to run a Reinforcement Learning (RL) algorithm, and the most accessible to us. The high practical cost of generating samples with a BCI prevented us from getting even that in the allocated timeframe.

To wrap up, although IRL may alleviate the accuracy problem in BCI driven settings, the many challenges this approach implies are far outside the comfort zone of the community.

III. Results

We installed OpenViBE, a user-friendly open-source tool for BCIs, on a windows XP system. This system supports both a

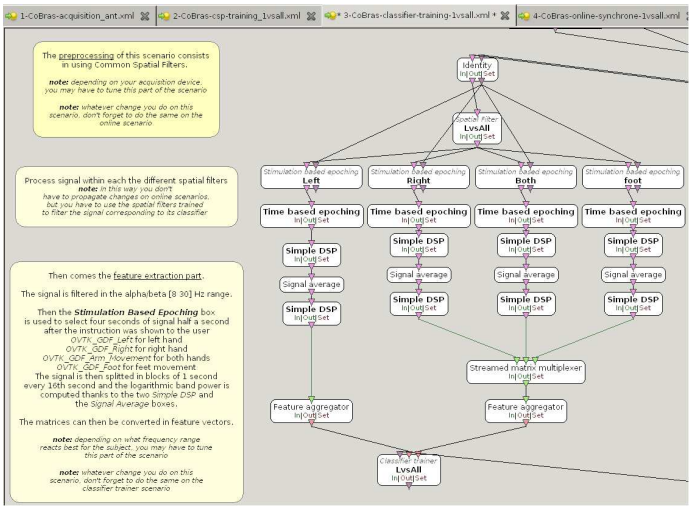

Fig. 4. OpenViBE scenario designed for training the one-versus-all classifiers.

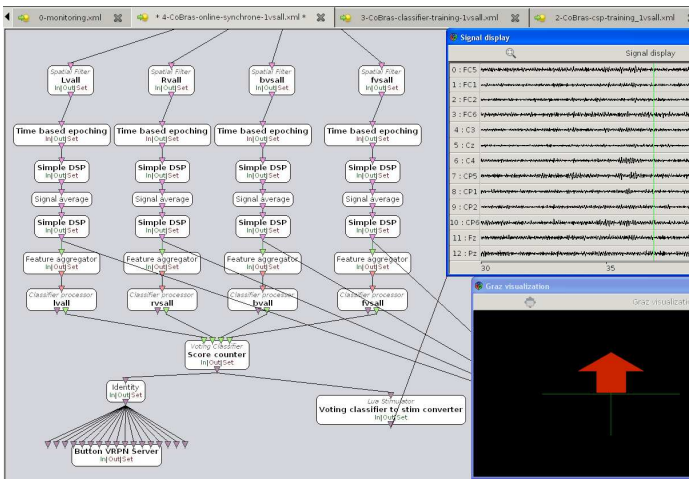

Fig. 5. OpenViBE scenario designed for on-line use. EEG signals are recorded, filtered, classified and one movement is sent to the robotic arm via the API.

JACO robotic arm driver and a Refa32 amplifier driver.

We built OpenViBE scenarii for i) signal acquisition ii) common spatial pattern filter training iii) classifier training and iv) offline use (see Fig. 4 and Fig. 5).

A. Standard motor imagery recognition

State-of-the-art similar results were obtained with imaginary and actual movements. The best combination strategy was the one-vs-all combined with a voting classifier. There were much less confusion and thus, better overall performance. It often happened that some classifier outputs had a very high confident level while the correct class was not represented. Confusion matrices were similar in both conditions (see Table I).

B. IRL

Let us disclose the end story immediately: not all challenges exposed earlier were overcomed.

TABLE I
CONFUSION MATRICE OBTAINED ON A TESTING SESSION.

Correct classes	*Predicted classes*			
	left hand	right hand t	both hands	feet
left hand	0.9	0	0.05	0.05
right hand	0.1	0.8	0.1	0
both hands	0.1	0.05	0.85	0.05
feet	0.1	0.1	0	0.8

The most hacky topic in the whole ordeal clearly was the composition of the state and action space of the MDP. Encouraging results were obtained on a simulation built to validate an initial approach. Sadly, this failed to generalize to the real thing as the real noise was much higher than modelized. A second, more sound approach was built, in which the state space directly consists in the output of the spatial filters and the last decision taken by the agent. This parametrization did not show any deep flaw and would be our goto parametrization if we are given the opportunity to work on this problem again.

The high dimensionality of the MDP was not a problem for our IRL algorithm, which was indeed able to infer a reward only from a few expert demonstration (corresponding to less than an hour of work for the operator).

Sadly, and this is the blocking point of the experiment so far, we were not able to train an agent on this reward. We need more data, specifically data sampled with a policy different from the expert's, in order to use the basic RL algorithm we tried to use [24]. We were thus not able to assess the quality of the found reward, although the fast convergence of SCIRL let us hope that it was good. We hope to solve this problem by either using less data greedy algorithm [25] or brutally generating more data (cumbersome for the operator). Another solution would be to use spatial filters able to deal with a displacement of the BCI, in order to allow the use of data from different sessions.

IV. CONCLUSION

We developed a whole processing chain using OpenViBE for controlling a robotic arm. According to the literature, we designed OpenViBE's scenarii (acquisition, filtering, classification and on-line use) based on a classic motor imagery paradigm. We selected four motor imageries (left hand, right hand, both hands and feet). They are respectively associates with 2D-movement (left move, right move, forward, backward). We used common spatial filters and one-versus-therest (linear discriminant analysis) classifiers. Our goal was not to improve the paradigm parameters but study how inverse reinforcement learning can help to select the right movement according to the classifier outputs and stored trajectories. Our classifier accuracy corresponds to the state-of-the-art. Thus, it is possible to control the Jaco to press a button. But, up to now, the IRL algorithm is not converging so cannot help to perform the right movement. Nevertheless, a significant analysis of the difficulties to apply IRL on high dimensional and noisy problem and ways to overcome them has been done.

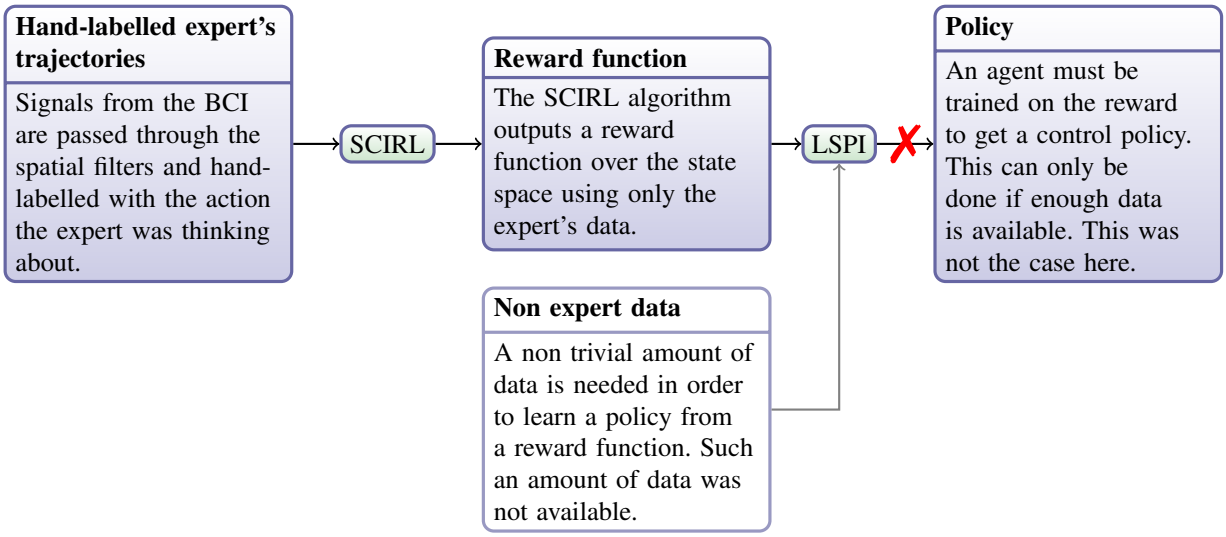

Fig. 6. Visual explanation of the IRL pipeline.

V. Perspectives

A deeper study is necessary for understanding the non-convergence of the IRL algorithm. If the IRL algorithm is robust enough, we will modify the processing chain to have a self-paced BCI. In the future, we also would like to use multiclass classifiers to avoid ambiguities due to the one-versus-therest approach. We would like to explore the tongue motor imagery to replace the both hands one. This choice avoids overlapped locations with the other motor imageries. We need to assess performance in offline and online conditions with a large population.

References

[1] J. R. Wolpaw, *et al.*, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/B6VNP-45HFKTC-2/2/070472147433d00168e8d54909b982d2

[2] F. Lotte and C. Guan, "Spatially regularized common spatial patterns for EEG classification," in *International Conference on Pattern Recognition (ICPR)*, 2010.

[3] J. Fruitet, *et al.*, "Automatic motor task selection via a bandit algorithm for a brain-controlled button," INRIA, Research Report 7721, 2011. [Online]. Available: http://hal.inria.fr/inria-00624686/fr/

[4] L. R. Hochberg, *et al.*, "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm," *Nature*, vol. 485, no. 7398, pp. 372–375, 05 2012. [Online]. Available: http://dx.doi.org/10.1038/nature11076

[5] T. J. Bradberry, R. J. Gentili, and J. L. Contreras-Vidal, "Reconstructing three-dimensional hand movements from noninvasive electroencephalographic signals." *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 30, no. 9, pp. 3432–3437, Mar. 2010. [Online]. Available: http://dx.doi.org/10.1523/JNEUROSCI.6107-09.2010

[6] G. M.-P. P. Ofner, "Decoding of hand movement velocities in three dimensions from the eeg during continuous movement of the arm," TOBI Workshop III, 2012. [Online]. Available: http://www.tobi-project.org/sites/default/files/public/Publications/TOBI-247.pdf

[7] J. R. Wolpaw and D. J. McFarland, "Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 51, pp. 17 849–17 854, Dec. 2004. [Online]. Available: http://dx.doi.org/10.1073/pnas.0403504101

[8] H. Yuan, C. Perdoni, and B. He, "Relationship between speed and eeg activity during imagined and executed hand movements," *Journal of Neural Engineering*, vol. 7, no. 2, p. 026001, 2010. [Online]. Available: http://stacks.iop.org/1741-2552/7/i=2/a=026001

[9] A. S. Royer, *et al.*, "EEG control of a virtual helicopter in 3-dimensional space using intelligent control strategies." *IEEE Trans Neural Syst Rehabil Eng*, vol. 18, no. 6, pp. 581–9, 2010.

[10] P. Horki, *et al.*, "Combined motor imagery and ssvep based bci control of a 2 dof artificial upper limb." *Med. Biol. Engineering and Computing*, vol. 49, no. 5, pp. 567–577, 2011.

[11] G. Pfurtscheller, *et al.*, "Current trends in graz brain-computer interface (bci) research," *Rehabilitation Engineering, IEEE Transactions on*, vol. 8, no. 2, pp. 216 –219, June 2000.

[12] G. Townsend, B. Graimann, and G. Pfurtscheller, "Continuous eeg classification during motor imagery-simulation of an asynchronous bci," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 12, no. 2, pp. 258–265, June 2004. [Online]. Available: http://dx.doi.org/10.1109/TNSRE.2004.827220

[13] C. Guger, *et al.*, "How many people are able to operate an eeg-based brain-computer interface (bci)?" *IEEE Trans Neural Syst Rehabil Eng*, vol. 11, no. 2, pp. 145–7, 2003. [Online]. Available: http://www.biomedsearch.com/nih/How-many-people-are-able/12899258.html

[14] K. K. Ang, *et al.*, "Clinical study of neurorehabilitation in stroke using eeg-based motor imagery brain-computer interface with robotic feedback." *Conf Proc IEEE Eng Med Biol Soc*, vol. 1, pp. 5549–52, 2010. [Online]. Available: http://www.biomedsearch.com/nih/Clinical-study-neurorehabilitation-in-stroke/21096475.html

[15] V. Kaiser, *et al.*, "First steps towards a motor-imagery based stroke bci: New strategy to set up a classifier," *Frontiers in Neuroprosthetics*, vol. Special Topic "Future invasive and non-invasive Brain-Machine-Interfaces(BMI) for acute and chronic stroke", 2011. [Online]. Available: http://www.tobi-project.org/sites/default/files/public/Publications/TOBI-136.pdf

[16] M. Naeem, *et al.*, "Seperability of four-class motor imagery data using independent components analysis," *Journal of Neural Engineering*, vol. 3, no. 3, p. 208, 2006. [Online]. Available: http://stacks.iop.org/1741-2552/3/i=3/a=003

[17] T. Wang, J. Deng, and B. He, "Classifying eeg-based motor imagery tasks by means of time?frequency synthesized spatial patterns," *Clinical Neurophysiology*, vol. 115, no. 12, pp. 2744–2753, 2004. [Online]. Available: http://dx.doi.org/10.1016/j.clinph.2004.06.022

[18] A. Bashashati, *et al.*, "A survey of signal processing algorithms in bci based on electrical brain signal," *J Neural Eng.*, vol. 4, no. 2, pp. 32–57, 2007.

[19] F. Lotte, *et al.*, "A review of classification algorithms for eeg-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 4, no. 2, 2007. [Online]. Available: http://stacks.iop.org/1741-2552/4/i=2/a=R01

[20] A. Ng and S. Russell, "Algorithms for inverse reinforcement learning," in *Proc. ICML*, 2000, pp. 663–670.

[21] M. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc. New York, NY, USA, 1994.

[22] E. Klein, *et al.*, "Structured Classification for Inverse Reinforcement Learning," in *European Workshop on Reinforcement Learning (EWRL 2012)*, Edinburgh (UK), 2012.

[23] P. Abbeel, A. Coates, and A. Ng, "Autonomous helicopter aerobatics through apprenticeship learning," *International Journal of Robotics Research*, vol. 29, no. 13, pp. 1608–1639, 2010.
[24] M. Lagoudakis and R. Parr, "Least-squares policy iteration," *The Journal of Machine Learning Research*, vol. 4, pp. 1107–1149, 2003.
[25] M. Geist and O. Pietquin, "Kalman Temporal Differences," *Journal of Artificial Intelligence Research (JAIR)*, vol. 39, pp. 483–532, October 2010. [Online]. Available: http://www.metz.supelec.fr/metz/personnel/geist_mat/pdfs/Supelec632.pdf

Laurent Bougrain is an associate professor at the university of Lorraine (France). He is a member of the Inria team NeuroSys dedicated to computational neuroscience at LORIA/Inria Nancy grand Est. He has been working for more than a decade on time series analysis with a focus on experimental data obtained during neuroscientific experiments. In recent years, he has dedicated his research to Brain-Computer Interfaces (BCI). He is working on template-based classifier for single trial detection using multichannel denoising techniques. He is the winner of the international BCI competition IV of the challenge about predicting the finger flexion from ECoG in 2008. He is currently working on a project on reinforcement learning to control a robotic arm and a wheelchair from EEG. He also collaborates to the worldwide BCI-software OpenVibe (http://openvibe.inria.fr). E-mail: bougrain@loria.fr, http://www.loria.fr/~bougrain.

Matthieu Divinage As a TIME student, Matthieu Duvinage holds an Electrical Engineering degree from the Facult Polytechnique of Mons (UMons, Belgium, 2009) and one degree from SUPELEC (France, 2009). He also holds a degree of fundamental and applied physics from Paris Sud XI Orsay (France, 2009) and a degree of management science from the School of Management at the University of Louvain (UCLouvain, 2011). His master thesis was performed at the Multitel research center (Mons, Belgium) and dealt with robust low complexity speech recognition using frame dropping based on voicing information and clustering techniques. He obtained an F.R.S-FNRS grant for pursuing a PhD thesis about the development of a lower limb prosthesis driven by a neural command in close partnership with the Universit Libre de Bruxelles (ULB).

Edouard Klein is a PhD student co-supervised by Yann Guermeur (ABC team, CNRS), Matthieu Geist (IMS research group, Suplec) and Olivier Pietquin (IMS research Group, Suplec and UMI 2958 (GeorgiaTech - CNRS)). The topic of his PhD is automatic feature selection in inverse reinforcement learning. He received the Electrical Engineering degree of PHELMA (Grenoble, France) in 2010. Since that, he has worked on reinforcement learning and learning from demonstration in the IMS research group of Suplec.