Team name	LIRIS
Team leader name	Natalia Neverova
Team leader address, phone number and email	35 rue Alexandre Boutin, 69100 Villeurbanne, France +33 6 58 10 27 87 natalia.neverova@liris.cnrs.fr
Rest of team members	Christian Wolf Graham Taylor
Team website URL (if any)	http://liris.cnrs.fr/natalia.neverova

Title of the contribution	Multi-scale spatial and temporal integration for multi-modal gesture recognition
General method description	In the context of multi-modal gesture detection and recognition, we propose a deep neural architecture that iteratively learns and integrates discriminative data representations from individual channels, modeling cross-modality correlations and temporal dependencies. Our framework integrates three data modalities: depth video, intensity channel and articulated pose. We propose a novel algorithm for pre-training the architecture on individual modalities followed by iterative representation fusing scheme. In our system, each gesture is decomposed into large-scale body motion and local subtle movements such as hand articulation. The idea of learning at multiple scales is also applied to the temporal dimension.
References	N. Neverova, C. Wolf, G. Paci, G. Sommavilla, G. W. Taylor, F. Nebout, A multi-scale approach to gesture detection and recognition. ICCV Workshop on Understanding Human Activities: Context and Interactions (HACI), 2013. (preliminary work)

Describe data preprocessing techniques applied (if any)	Local contrast normalization of depth and intensity channels
Describe features used or data representation model (if any)	Depth and intensity channels: raw data (from bounding boxes surrounding hands). Articulated pose: a skeleton descriptor based on characteristic angles and distances between upper body joints as well as their speeds and accelerations.
Data modalities used, i.e. depth, rgb, skeleton… (if any)	Skeleton, depth, intensity.
Fusion strategy applied (if any)	Early fusion within a neural structure.
Dimensionality reduction technique applied (if any)	-

Temporal clustering approach (if any)	Concatenating of features from sequences of frames with different steps
Temporal segmentation approach (if any)	Motion detection based on skeleton stream.
Gesture representation approach (if any)	Each gesture is represented as a combination of large scale body motion (derived from the skeleton stream) and hand articulation (based on depth and intensity channels).
Classifier used (if any)	Deep neural architecture.
Large scale strategy (if any)	

Transfer learning strategy (if any)	
Temporal coherence and/or tracking approach considered (if any)	Hand tracking based on skeleton stream.
Other technique/strategy used not included in previous items (if any)	
Method complexity analysis	

Qualitative advantages of the proposed solution	Fusing multiple modalities at several spatial and temporal scales leads to a significant increase in recognition rates, allowing the model to compensate for errors of the individual classifiers as well as noise in the separate channels.
Results of the comparison to other approaches (if any)	
Novelty degree of the solution and if is has been previously published	Novel multi-scale approach to feature extraction and fusion scheme, an article is submitted for a journal publication.

Language and implementation details (including platform, memory, parallelization requirements)	Python (with Theano library), tested on Ubuntu 12.04 and GeForce GTX580 graphics card. Using GPU is beneficial for training, but is not mandatory.
Human effort required for implementation, training and validation?	Significant.
Training/testing expended time?	Training: 3-4 days depending on the hardware, testing: 1.5 days for the test set (including data extraction, reformatting and preprocessing).
General comments and impressions of the challenge	