Team name	Lionel Pigou
Team leader name	Lionel Pigou
Team leader address, phone number and email	lionelpigou@gmail.com
Rest of team members	
Team website URL (if any)	

Title of the contribution	Gesture recognition with convolutional neural networks (CNNs)
General method description	The architecture of the model consists of two CNNs (pooling-scheme = max-pooling), one for extracting hand features and one for extracting upper bodyfeatures. Each CNN is three layers deep. Classification is done with a classical artificial neural network (ANN) with one hidden layer. Also, local contrast normalisation (LCN) is applied in the first two layers and all artificial neurons are rectified linear units (ReLUs). During training, dropout and data augmentation are used to generalise the model. The data augmentation is performed in realtime on the CPUs during the training phase while the model trains on the GPU. This consists of spatial - and temporal translations, zooming, and rotations. I used Nesterov's accelerated gradient descent (NAG). (See next slide for preprocessing)
References	• V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 807–814. • X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Networks," Proceedings of the 14th International Con- ference on Artificial Intelligence and Statistics, vol. 15, pp. 315–323, 2011. • G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv preprint arXiv:1207.0580, 2012. • A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in Neural Information, pp. 1–9, 2012. • I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in Proceedings of the 30th International Conference on Machine Learning (ICML-13), 2013, pp. 1139–1147.

Describe data preprocessing techniques applied (if any)	In the preprocessing, the highest hand and the upper body are cropped using the given joint information. This results in four video samples (hand and body with depth and grayscale) of resolution 64x64x32. Furthermore, the noise in the depth maps is reduced with thresholding, background removal using the user index, and median filtering.
Describe features used or data representation model (if any)	CNN. (Instead of constructing complex handcrafted features, CNNs are able to automate the process of feature construction.)
Data modalities used, i.e. depth, rgb, skeleton… (if any)	Everything except skeleton world positions and orientations.
Fusion strategy applied (if any)	Early fusion
Dimensionality reduction technique applied (if any)	CNN: max-pooling

Temporal clustering approach (if any)	
Temporal segmentation approach (if any)	Once the prediction probability is high enough (thesholding), the model predicts the starting frame. If it decreases again or it switches class, the model predicts the endframe. Also, extra class added to distinguish sequences not in the 20 gestures. There was more focus on classification than segmentation.
Gesture representation approach (if any)	CNN
Classifier used (if any)	Artificial neural network (ANN)
Large scale strategy (if any)	

Transfer learning strategy (if any)	
Temporal coherence and/or tracking approach considered (if any)	
Other technique/strategy used not included in previous items (if any)	
Method complexity analysis	Training is computationally intensive (GPU speeds up things), evaluation is relatively efficient.

Qualitative advantages of the proposed solution	**Instead of constructing complex handcrafted features, CNNs are able to automate the process of feature construction.** **The more data available for training the model, the robuster it will get. Often, these deep learning methods outperform feature engineering methods in terms of classification, if provided with enough data.**
Results of the comparison to other approaches (if any)	
Novelty degree of the solution and if is has been previously published	CNNs are not new, but the way I used them as described earlier is my own solution. The solution is not published.

Language and implementation details (including platform, memory, parallelization requirements)	**Python, Theano, PyLearn2.** **Ubuntu 12.04LTS, 32GB mem, GPU NVIDIA GeForce GTX 680, hexacore (i7-3930K).** **Used multiple cores for data-augmentation while training the model on GPU.**
Human effort required for implementation, training and validation?	I did this project in context of my master thesis for graduating in Master of Science (MSc), Computer Science Engineering: Information and Communication Technology. This took me 10 months while finishing my studies and writing a master thesis.
Training/testing expended time?	One or two days.
General comments and impressions of the challenge	I like to thank the organizers for the challenge and the data! It is my first challenge, but to me the organization was perfect.