Team name	Charles River Analytics
Team leader name	Camille Monnier
Team leader address, phone number and email	625 Mount Auburn St Cambridge, MA 02138 USA 617.416.5132 cmonnier@cra.com
Rest of team members	Stan German Andrey Ost
Team website URL (if any)	www.cra.com

Title of the contribution	A Multi-scale Sliding Window Detector for Efficient and Robust Gesture Detection
General method description	Our method combines a collection of individual sliding-window gesture detectors with multi-modal features. A set of boosted classifiers is trained on labeled gestures and evaluated in a one-vs-all manner. Our features include skeleton and image descriptors, which are extracted at each frame and summarized over a temporal window to produce a fixed-length feature vector. Our system may process over a minute of data per second once features have been extracted – we achieve this same runtime at slightly reduced accuracy (0.79 vs. 0.82 on the validation data) if we use only Kinect pose estimates.
References	

Describe data preprocessing techniques applied (if any)	Simple normalization of the skeleton coordinates relative to each individual’s torso length. We masked the hands using depth-based segmentation prior to computing HOG features.
Describe features used or data representation model (if any)	Skeletal pose, including joint positions, angles, and derivatives. Depth-masked hand HOG bag-of-words descriptors.
Data modalities used, i.e. depth, rgb, skeleton… (if any)	Skeleton, RGB, depth
Fusion strategy applied (if any)	We let the classifier decide which features mattered
Dimensionality reduction technique applied (if any)	None

Temporal clustering approach (if any)	None
Temporal segmentation approach (if any)	None
Gesture representation approach (if any)	Fixed-length feature descriptor describing temporal sequence of image+pose data
Classifier used (if any)	Adaboost
Large scale strategy (if any)	Bootstrapping for efficiently collecting hard examples

Transfer learning strategy (if any)	None
Temporal coherence and/or tracking approach considered (if any)	None
Other technique/strategy used not included in previous items (if any)	None
Method complexity analysis	Linear in the number of gestures

Qualitative advantages of the proposed solution	**The method is straightforward and rooted in well-understood and well-established detection algorithms. Errors are generally easy to visualize and understand.**
Results of the comparison to other approaches (if any)	None
Novelty degree of the solution and if is has been previously published	The method has not been published, although many components have been. The novelty lies primarily in the application of a general approach that has been successfully applied to many object detection/recognition problems, but has not (to our knowledge) been as widely or successfully applied to the problem of gesture recognition.

Language and implementation details (including platform, memory, parallelization requirements)	**Matlab was used to rapidly develop basic feature extraction code. The detection framework is implemented in C++. Parallelization is enabled for Matlab feature extraction.**
Human effort required for implementation, training and validation?	Implementation: This depends on the tools available and familiarity with the domain. On the order of months for somebody who is generally familiar with the domain.
Training/testing expended time?	Skeletal data only: training achieved in ~30 minutes, testing requires ~1 second/minute of Kinect data (unparallelized). Extracting image features with Matlab is quite slow, and adds about ~2-4 hours of preprocessing time (although this is still much faster than actual frame rate).
General comments and impressions of the challenge	Fun challenge! We had recently developed a gesture recognition system for a completely different purpose and joined at the last minute. It quickly became apparently that the skeleton data was more than adequate to do quite well (0.79 on validation), with the exception of gestures that differ only in hand pose. Hand pose information was useful (0.83 on validation), but the image quality seemed to pose a challenge. Out-of-sample gestures pose a challenge. The out-of-sample gestures raise an interesting philosophical question. How does a human interpret a small variation (e.g., different hand pose), if they've never seen it before? Usually by being taught, or inferring a new meaning from context. It seems likely that some out-of-sample