Team name	Telepoints
Team leader name	Karthik Nandakumar
Team leader address, phone number and email	Institute for Infocomm Research 1 Fusionopolis Way, Singapore 138632 +65 64082483 knandakumar@i2r.a-star.edu.sg
Rest of team members	Wang Jiangang Wan Kong Wah Alice Chan Siu Man Manoj Ramanathan Yau Wei Yun
Team website URL (if any)	

Title of the contribution	A Multimodal Gesture Recognition System Based on 3D Video and Skeletal Joint Locations
General method description	We first perform temporal segmentation of the video into independent gesture segments using the skeletal joint locations. In particular, the segmentation is based on changes in the y-coordinates of the six hand joints (marked as elbow, wrist, and hand). We then attempt to predict the gesture contained in each gesture segment by utilizing three modalities, namely, RGB, depth, and skeletal joints. While Space-Time Interest Points (STIP) are used to represent the RGB and depth modalities, a covariance descriptor is extracted from the skeletal joint data. For all three modalities, Support Vector Machines (SVM) are used for gesture classification. A Gaussian classifier is also constructed based on the raw joint location data. In the case of RGB and depth modalities, we also include a rejection class (unrecognized gesture) using non-labelled gesture segments in the development and validation datasets. Finally , we fuse the classification results from the following three classifiers: STIP-SVM from the RGB modality, covariance-based joint descriptor+SVM, and skeletal joint locations + Gaussian classifier. Fusion is based on the based on the weighted sum rule, where the weights are optimized based on a subset of the validation dataset that is not used for classifier training.
References	I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64:107–123, September 2005. A. Sanin, C. Sanderson, M. Harandi, and B. Lovell. Spatio-temporal covariance descriptors for action and gesture recognition. In *Proceedings of Workshop on the Applications of Computer Vision, Clearwater,* USA, January 2013. K. Nandakumar et al., “A Multi-modal Gesture Recognition System Using Audio, Video, and Skeletal Joint Data”, Proc. of ICMI '13, 475-482.

Describe data preprocessing techniques applied (if any)	
Describe features used or data representation model (if any)	**RGB & Depth Videos**: Bag of visual features: Spatio-temporal interest points (harris-3D + histogram of optical flow) – the STIP features are quantized into a lexicon of size 14000 **Skeletal Joint Locations**: Covariance descriptor based on 3D (Wx, Wy, Wz) and 2D (Px, Py) values of six hand joints (marked as elbow, wrist, and hand joint of the right and left hands). Relative 3D locations of the six hand joints with respect to the HipCenter
Data modalities used, i.e. depth, rgb, skeleton… (if any)	RGB, depth, skeletal joints
Fusion strategy applied (if any)	Weighted sum of likelihoods from the following three classifiers: STIP-SVM from the RGB modality, covariance-based joint descriptor+SVM, and skeletal joint locations + Gaussian classifier.
Dimensionality reduction technique applied (if any)	

Temporal clustering approach (if any)	
Temporal segmentation approach (if any)	Temporal segmentation is based on changes in the y-coordinates of the six hand joints. For each hand, the y-locations (relative to the y-coordinate of the HipCenter) of the three hand joints are added together to create a 1-D signal. This signal is differentiated and smoothed to obtain a change signal. The positive and negative peaks of the change signal are detected. The gesture is assumed to start when the change signal begins to increase from zero towards the positive peak. Similarly, the gesture is assumed to end, when the change signal decreases to zero after the negative peak. Thresholds are applied to filter out spurious peaks. Ad-hoc rules are also designed to combine neighboring peaks if the detected gestures are too short or split a gesture segment that is too long.
Gesture representation approach (if any)	
Classifier used (if any)	Support Vector Machines (SVM), Gaussian Classifier
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	STIP feature extraction for RGB and depth modalities is computationally very extensive. For the development and validation datasets, extraction took 5 days to complete over a cluster of 8 Linux VM machines. The greatest bottleneck is feature clustering, where we explore various cluster sizes to empirically optimize the classification performance over the Validation dataset.

Qualitative advantages of the proposed solution	**Our segmentation algorithm is computationally very efficient** **In the RGB domain, our bag of visual features approach is robust to occlusions.** **For skeletal joints, our covariance descriptor is viewpoint-invariant and compact**
Results of the comparison to other approaches (if any)	
Novelty degree of the solution and if is has been previously published	While the individual components used in our proposed system (e.g., STIP features, covariance descriptor, SVM and Gaussian classifiers) cannot be considered as very novel, we believe that the key to achieve a good performance on this challenging data set is to leverage of the complementary strengths of the different modalities. The proposed solution is very similar to the Telepoints submission for the2013 ICMI CHALEARN Multi-modal Gesture Recognition Challenge, the details of which were published in ICMI 2013. The key differences include the use of depth modality, the temporal segmentation algorithm, and the use of Gaussian classifiers for the skeletal joint data.

Language and implementation details (including platform, memory, parallelization requirements)	**RGB and Depth modalities (STIP-SVM): C/C++ on Linux platform** **Skeletal Joint Covariance + SVM: C/C++/Matlab on Windows platform** **Temporal segmentation, Guassian Classifier, and Fusion: Matlab**
Human effort required for implementation, training and validation?	
Training/testing expended time?	The most computational intensive part is STIP feature extraction for RGB and depth modalities, which takes a few hours. SVM training also takes a few hours All other processing can be completed in a few minutes
General comments and impressions of the challenge	Very interesting and challenging. The challenges include: (a) short & sometimes contiguous gestures, (b) similar gestures (especially those around the face), and (c) gestures with large intra-class variations. In our view, the use of Jaccard score was a bit extreme. Even when a gesture is correctly detected and recognized, the Jaccard score may be only around 0.9 if the start and end frames are off by 1 or 2 frames from the ground truth labels. It is very difficult to detect the start and end frames accurately to match the ground truth labels to the exact frame number.