Team name	CSU-SCM
Team leader name	Bin Liang
Team leader address, phone number and email	Address: Charles Sturt University, Wagga Wagga, NSW, Australia Phone number: +61426966508 Email: bliang@csu.edu.au
Rest of team members	Bin Liang
Team website URL (if any)	No

Title of the contribution	Multi-modal gesture recognition using video and skeleton data
General method description	1. We firstly segment each video sample into several single gestures based on the skeleton joints differences within sliding windows. 2. Upper-body skeleton data in each frame are used as time-domain features for GMM-HMM. The normalized probability scores compose the time-domain features for RBF-Kernel SVM. 3. Summary statistics of time-domain features are used for the input for another RBF-Kernel SVM. 4. 2DMTM-PHOG features are used for linear-SVM. 5. Finally, we use late fusion of three SVM probabilities to combine three types of features.
References	1. Bin Liang and Lihong Zheng. “Three Dimensional Motion Trail Model for Gesture Recognition.” In *Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops* (ICCVW '13), 2013. 2. Nandakumar, Karthik, et al. “A multi-modal gesture recognition system using audio, video, and skeletal joint data.” *Proceedings of the 15th ACM on International conference on multimodal interaction* (ICMI '13), 2013. 3. Xi Chen and Markus Koskela. “Online RGB-D gesture recognition with extreme learning machines.” In *Proceedings of the 15th ACM on International conference on multimodal interaction* (ICMI '13), 2013. 4. Immanuel Bayer and Thierry Silbermann. “A multi modal approach to gesture recognition from audio and video data.” In *Proceedings of the 15th ACM on International conference on multimodal interaction* (ICMI '13), 2013

Describe data preprocessing techniques applied (if any)	Skeleton featrures normalization
Describe features used or data representation model (if any)	2DMTM-PHOG Skeleton features
Data modalities used, i.e. depth, rgb, skeleton… (if any)	Depth video Skeleton data Mask video
Fusion strategy applied (if any)	Late fusion with weighted SVM probabilities
Dimensionality reduction technique applied (if any)	No

Temporal clustering approach (if any)	**No**
Temporal segmentation approach (if any)	Segmentation based on skeleton joints differences within sliding windows.
Gesture representation approach (if any)	2D-MTM
Classifier used (if any)	RBF-kernel SVM Linear SVM RBF-kernel SVM with GMM-HMM scores
Large scale strategy (if any)	No

Transfer learning strategy (if any)	No
Temporal coherence and/or tracking approach considered (if any)	No
Other technique/strategy used not included in previous items (if any)	No
Method complexity analysis	Training: about 35 hours on PC with CPU (Intel Core i5-3320M) and RAM (8G) Testing: about 10 hours on PC with CPU (Intel Core i5-3320M) and RAM (8G)

Qualitative advantages of the proposed solution	**Easy to implementation** **Effective** **Efficient**
Results of the comparison to other approaches (if any)	No
Novelty degree of the solution and if is has been previously published	This is a novel method which combines three types of features, *i.e.* time-domain features, statistic skeleton features and 2DMTM-PHOG features.

Language and implementation details (including platform, memory, parallelization requirements)	Language: Python 2.7.6 (64 bit), Modules: OpenCV 2.4.8, libsvm-3.1.8, numpy, scipy-0.14.0 scikit-image-0.9.3, scikit-learn-0.14.1 Platform: Windows 7 64 bit Memory: 8 GB
Human effort required for implementation, training and validation?	No
Training/testing expended time?	Training: 25 hours Testing: 10 hours
General comments and impressions of the challenge	This competition provides a large dataset for gesture recognition using multi-modal, which can help researchers evaluate their methods efficiently. If the submission system works well, it will be better. Thanks for the efforts of the organizers.