Inventory of available datasets:

Dataset name	*Domain*	*Application*	*Type of sensors*	*Conditions and size*	*Link*
UTD Multimodal Human Action Dataset (UTD-MHAD)	Machine learning	Human activity recognition	• RGB camera providing video. • Kinect sensor, providing depth video and skeleton data. • Inertial data from right wrist or the right thigh.	27 different actions executed by 8 subjects (4 males, 4 females), repeated 4 times, for a total of 861 data sequences.	Paper: http://www.utdallas.edu/~cxc123730/ICIP2015-Chen-Final.pdf Website: http://www.utdallas.edu/~cxc123730/UTD-MHAD.html
A multimodal corpus for gesture expressivity analysis	Machine learning (computer interfacing)	Gesture expressivity (w focus on hands)	- HumanWare data glove - Wii remote - Microphone - Camera	Given 3 (positive, neural and negative) phrases to express. Participants from 3 different countries were invited (total 51)	Paper : http://www.image.ece.ntua.gr/papers/629.pdf
Berkeley MHAD	Machine learning	Human activity recognition	- 2 Kinect sensors - 6 wireless accelerometters - 4microphones - Impulse optical motion capture system (LED markers) - 4 multi-view stereo vision camera arrays	11 actions performed by 7 male and 5 female subjects. 5 repetition, for total of 660 action sequences. T-pose for each subject.	http://tele-immersion.citris-uc.org/berkeley_mhad
CITEC	Robotics (mobile)	Urban Search and Rescue (USAR)	- IMU - Odometer - Omnidirectional camera - rotating laser rangefinder	- Novel fusion scheme based on extended Kalman filter for 6 degree of freedom - 4.4 Km under standard USAR conditions - Indoors and outdoors - valided with ground truths (with motion capture devices, etc)	Paper: http://cmp.felk.cvut.cz/ftp/articles/svoboda/Kubelka-JFR2015.pdf Website: https://www.cit-ec.de/en/content/multimodal-data-fusion-mobile-robots-usar-environment-0 https://sites.google.com/site/kubelvla/public-datasets/nifti-zurich-2013
MobBIO	Machine learning	Person authentification (biometrics)	Asus Transformer Pad TF 300T (back 8MP camera used). Face and iris images. Voice recording as well.	- 105 Volunteers. Mainly Portuguese with few U.K., Romania and Iran. 29% females, 71% males	Paper : http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7295072

				- Voices samples with 16 sentences in Portuguese - Iris images 8 per eye per volunteer. 2 differents lighting condtions. - Faces, 16 images per volunteer with 2 lighting conditions.	
Audio/Video Fusion for Objects recognition	Machine learning	Object recognition	Camera and built in microphone	28 toys with unique sounds move from left to right. 3 different recordings with different conditions(normal, visual occlusion and audio occlusion)	Paper: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5354442
CUAVE	Machine learning	Face tracking, audio localization, speaker detection	720X480 camera with on camera microphone	Part 1 with 36 solo speakers. Under 4 different conditions. Part 2 with 20 pairs of speakers under 3 different condtions green screen	Paper : http://www.clemson.edu/ces/speech/papers/cuave.pdf
SKIG	Machine learning/ AI	Hand gesture recognition	Kinect (RGB and Depth	6 subjects 10 categories of gestures 3 hand poses per gestures 3 different backgrounds 2 types of illumination 2 types of visual (RGB and depth)	http://lshao.staff.shef.ac.uk/data/SheffieldKinectGesture.htm
AVEC2014	Machine Learning	Emotion/depression detection	Camera, microphone	150 audio/video recordings for 2 tasks(Northwind, and Freeform) (total 300) Northwind : reading aloud a fable in german Freeform : answer to common questions in german split into 3 (training, developpemnt,test)	Data description paper: http://www.cs.nott.ac.uk/~pszmv/Documents/avec2014_preprint.pdf Event page : http://avec2013-db.sspnet.eu/ Related paper : http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7344573

				ground truths labels for Valence, arousal, and dominance given every 1/30th second	

Notable mentions:

Survey of video datasets	Machine learning	Various	Various	Multiple camera setups (surveillance, street, mall, dynamic background, static background, etc)	http://www.sciencedirect.com/science/article/pii/S1077314213000295

Domain: robotique (fixed or mobile), artificial intelligence, machine learning, etc.

Application: speech recognition with lips features, knowledge base, human action/activity recognition, human-machine interaction, etc.

Type of sensors: video camera (frame rate, color or black/white), audio (mono or multichannel), depth sensor, joint angle sensors, etc.

Conditions and size: total length of recordings (in minutes), number of experiments and trials, etc.

Link: link to the related conference/journal paper, and website that hosts the dataset if available.