Real-time recognition of human gestures for collaborative robots on assembly-line

E. COUPETÉ*, S. MANITSARIS and F. MOUTARDE

Robotics Lab, Mines ParisTech, 60 Bd St Michel, 75006 Paris, FRANCE

Abstract

We present a framework and preliminary experimental results for real-time recognition of human operator actions. The goal is, for a collaborative industrial robot operating on same assembly-line as workers, to allow adaptation of its behavior and speed for smooth human-robot cooperation. To this end, it is necessary for the robot to monitor and understand behavior of humans around it. The real-time motion capture is performed using a "MoCap suit" of 12 inertial sensors estimating joint angles of upper-half of human body (neck, wrists, elbows, shoulders, etc…). In our experiment, we consider one particular assembly operation on car doors, which we have further subdivided into 4 successive steps: removing the adhesive protection from the waterproofing sheet, positioning the waterproofing sheet on the door, pre-sticking the sheet on the door, and finally installing the window "sealing strip". The gesture recognition is achieved *continuously* in real-time, using a technique combining an automatic time-rescaling similar to Dynamic Time Warp (DTW), and Hidden Markov Model (HMM) for estimating respective probabilities of the 4 learnt actions. Preliminary evaluation, conducted in real-world on an experimental assembly cell of car manufacturer PSA, shows a very promising action correct recognition rate of 96% on several repetitions of the same assembly operation by a single operator. Ongoing work aims at evaluating our framework for same actions recognition but on more executions by a larger pool of different human operators, and also to estimate false recognition rates on unrelated gestures. Another interesting potential perspective is the use of workers' motion capture in order to estimate effort and stress, for helping prevention of physical causes of some musculoskeletal disorders.

Keywords: Technical gestures recognition, collaborative robotics (cobotics), factory assembly-line.

1. Introduction

Recent advances in motion capture technologies make it now easier to perform real-time monitoring of human activities, for various purposes. Meanwhile, a current trend in manufacturing robotics is the development of collaborative robotics (cobotics), in which robots perform activities jointly, or at least side-by-side, with human operators. As already highlighted long ago by (Inagaki et al., 1995), for humans and robots to have a common goal and work cooperatively, human intention inference by robots is required. This in turn necessitates robots to be able to recognize human actions.

Prototyping a framework (from sensors to data-processing) for real-time human technical gestures recognition in the context of automotive assembly line is precisely one of our research goals within the "*Chaire PSA Peugeot-Citroën on Robotics and Virtual Reality*". Technical gestures are invented by a specialist minority for use strictly within the limits of their particular activity. These gestures are meaningless to anyone outside the specialization, and make sense only in their narrow operating field. In automotive assembly lines and more precisely in the experimental cell of PSA Peugeot Citroen, technical gestures can be "to remove the sealing sheet", "to fit the windows sealing sheet", "to screw up without tightening", "to hammer" etc. Recognition of human gestures and activities has become an important research area with numerous potential applications including sign language interpretation, automated surveillance and human robot collaboration. In order to capture human motions, several systems can be used, as illustrated on Figure 1. The oldest and most commonly used is the video from RGB cameras. More recently the apparition of real-time depth cameras, like the Microsoft Kinect™, brought new possibilities for human motion capture and monitoring. Thirdly, inertial motion sensors (IMS) like MotionPod™ of Movea (see http://www.movea.com/), or the

*Corresponding author. Email: Eva.Coupete@mines-paristech.fr

motion capture suit developed by AnimaZoo™ (see http://www.animazoo.com/), enable the direct acquisition of motion information in real time.

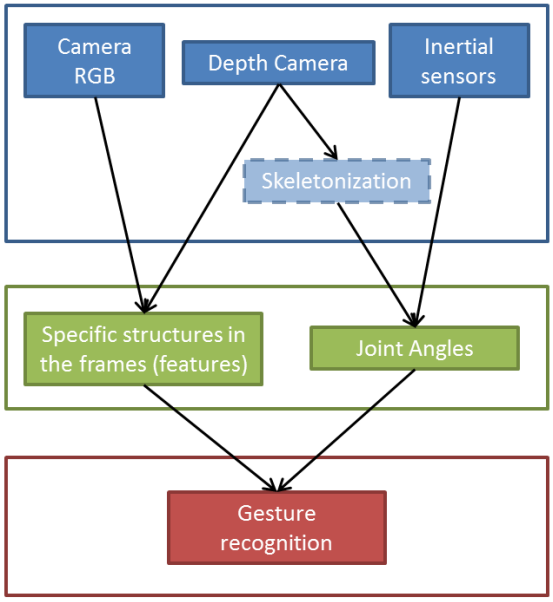

Figure 1 : Framework for human motion capture and gesture recognition.

When using information provided by RGB camera, in order to be able to perform gesture recognition, some features have to be extracted. In (Laptev and Lindeberg, 2003) and (Schuldt et al., 2004), authors introduce space time interest points, features computed in a spatio-temporal domain. More specifically to human, (Gorelick et al., 2007) use body motion as a space time feature. In (Laptev and Perez, 2007) authors use optical flow histograms and spatial gradient histograms. In (Lv et al., 2007) authors compare silhouettes extracted in the frames with key poses, using an extension of the PMK (Pyramid Match Kernel) described in (Grauman and Darrell, 2005). Similarly in (Sullivan and Carlsson, 2002) authors compare key frames to the frames computing shape equivalences.

With inertial motion sensors, the output is the angles of joints, which can directly be used as features for gesture recognition. However, there are several possible choices for rotational representation: for instance Euler angles as in (Piovan and Bullo, 2012), quaternions as in (Bachmann et al., 1999) or Direction Cosine Matrix (DCM) as in (Bar-Itzhack et al., 2010).

As for depth cameras, a skeleton posture has to be estimated first, before extracting the joint angles. For example, in case of the Microsoft Kinect, a default skeletonization is performed by the provided SDK. It enables real-time pose estimation, restricted however to human standing up and facing the sensor. This limitation is because this skeletonization relies on a randomized decision forest (Lepetit et al., 2005) trained on a database of depth images of human poses, very large but limited to standing-up and facing postures (Shotton et al., 2011).

Various approaches have been proposed to handle dynamic gestures recognition by exploiting the features previously calculated. Most of the problems have been solved with statistical techniques: Hidden Markov Models (HMM) as in (Lovell et al., 2004), (Kellokumpu et al., 2005) or (Oka et al., 2002) and/or Principal Component Analysis (PCA) as in (Kim and Song, 2008). SVM (Support Vector Machine) are also becoming a popular way for visual spatio-temporal pattern recognition as in (Schuldt et al., 2004). Template-matching is also sometimes used for matching with temporal templates, as in (Bobick and Davis, 2001).

This article presents our research work on development of a technical gesture recognition system for human-robot collaboration in an experimental cell of PSA Peugeot Citroen. Section 2 presents our approach, sensors and algorithms used, and our experimental protocol. Section 3 then provides and analyzes results of our experiments. Finally, section 4 contains our conclusions and perspectives for this work.

2. Methodology

Technical gestures are strongly related to the movement of parts of the body, what is called “motion”. But, what we mean by the word “motion” and how can we analyze it? Motion analysis intends to describe the body movements, but in most applications not all the details of the human body is required. The representation and structure of human motion should be simplified according to the application.

In the case study of the Human-Robot Collaboration (HRC), a good simplification of the complexity of the human body is to estimate only position of ***joints***, or angle of ***segments***, since they can provide a sufficient representation of the human posture. Other information, such as clothing, tendons, muscles etc, are absolutely unnecessary for the technical gesture recognition for HRC in general, and in particular for our specific case study of PSA. In order to avoid self-occlusions and scene occlusions of the gestures, we have chosen to capture rotational information of body segments using inertial sensors.

2.1. Inertial motion sensors for technical gesture recognition

Two different types of IMS have been tested with virtually performed gestures: a) 3 individual and non-hierarchical inertial sensors from Movea™ (MotionPod); b) 12 hierarchical inertial sensors for the upper part of the body from AnimaZoo™ (IGS120+). Both of these technologies are

unaffected by occlusions (contrary to vision-based techniques), and they provide a rotational representation of the gestures.

For using Movea™ inertial sensors, two of them have been mounted on the wrists of the worker and the third one on his torso. The latter has been used as a reference point for the other two sensors. The acquired data are angular accelerations for the X, Y, Z axes for the wrists, and they have been normalised a posteriori to Euler angles in order to train the HMMs.

Figure 2 : On the left, a worker fitting the windows sealing-sheet on a door in the experimental cell of PSA, while he wears the AnimaZoo™ IGS120+; on the right 3 inertial sensors from IGS120+.

The IGS120+ is a product designed for an industrial use, with robustly packaged inertial sensors and specific sensor data correction to reduce potential disruptions due to possible magnetic fields that may exist in the industrial environment. On the other hand, the price of a single sensor of the IGS120+ is twice the price of a single Movea™ sensor. Additionally, the IGS120+ is more appropriate for an industrial use for security reasons. More precisely, the IGS120+ gives a complete representation of the upper part of the worker's body, which is an important information required in order to be able to detect body postures in case of accidents, or other abnormal situations. The IGS120+ provides exactly the same type of data as those from Movea™, but with 12 sensors placed on different body segments (see Figure 2).

2.2. Gestures recognition technique

We perform recognition of technical gestures using the processing pipeline illustrated on Figure 3. Gestures execution speed, as well as details of movements can vary significantly from one realization to another. We therefore use, for learning and recognition, a hybrid approach combining Hidden Markov Model (HMM) with a time-rescaling similar to Dynamic Time Warping (DTW); this technique was designed and developed by (Bevilacqua et al., 2007, 2010), and we use the implementation they provide in the "Gesture Follower" tool[1]. This is a template-based method which allows us to use a single gesture to define a gesture class, by training a HMM. One of the advantages of this method is the DTW-like online time-alignment, illustrated on Figure 4, which permits good robustness of recognition to variation in execution of same technical gesture.

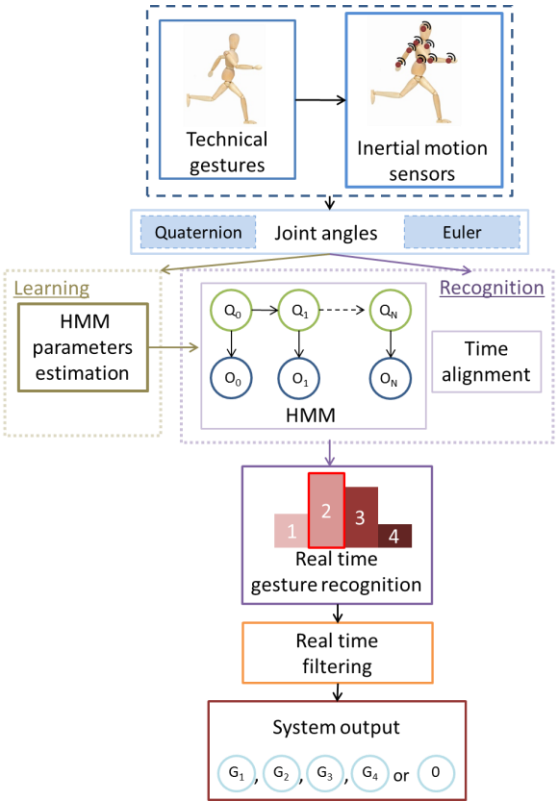

Figure 3 : Gesture recognition pipeline based on the AnimaZoo™ suit of wireless motion sensors for the upper-part of the body, and one-shot learning of HMM + time-alignment for robust recognition.

A HMM is a statistical modeling method which can be used to establish a model for a time-series with spatial and temporal variability. As described by (Rabiner,1989), a HMM is composed of N hidden states $\{S_1, S_2, \ldots, S_N\}$, each associated with M possible observations. We denote the actual state at time t as q_t and the observation at time t as O_t. A HMM is defined by its structure and its parameters. The whole is denoted by λ. The structure of a HMM is the number of states and the non-zero transition probabilities. The parameters are:

- The starting state probabilities:
  $\boldsymbol{\Pi} = \{\boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \ldots, \boldsymbol{\pi_N}\}$, with
  $\boldsymbol{\pi_i} = \boldsymbol{P(q_1 = S_i \mid \lambda)}$.
- The state transition probabilities:
  $\boldsymbol{A} = \{\boldsymbol{a_{ij}}, \boldsymbol{1 \le i,j \le N}\}$ with
  $\boldsymbol{a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i, \lambda)}$
- The observation probabilities:
  $\boldsymbol{B} = \{\boldsymbol{b_j(k)}, \boldsymbol{1 \le j \le N, 1 \le k \le M}\}$
  with $\boldsymbol{b_j(k) = P(O_t = k \mid q_t = S_j, \lambda)}$.

To simplify the HMM learning procedure, each learning example is defined by a left-to-right

[1] http://imtr.ircam.fr/imtr/Gesture_Follower

Markov chain. The left-to-right model does not have backward paths, the state transition probabilities allow to stay in the same state or to go in the upper state. This model is appropriate to represent a temporal system. The learning example is down-sampled, with a constant sampling rate, and, as in DTW, each sample is associated to a state. The observation probability for each state is a Gaussian model centered on the sample value. The recognition procedure can be done in real time using the well-known forward procedure which computes the likelihoods for the different models (Bevilacqua et al., 2010).

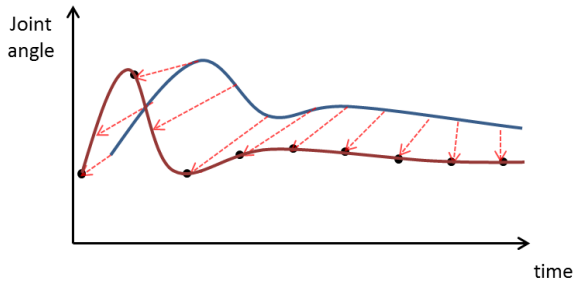

Figure 4 : The red curve represents the learning example and the blue curve an observation. The red arrows illustrate the time alignment between the reference gesture and the performed gesture. Each sample of the reference gesture is used to generate a state of a left-to-right HMM.

2.3. Experimental protocol

The gestures that have been analysed and recognized with our system are the following:

- G1: "**to remove protective paper**"
- G2: "**to fit on door the waterproofing-sheet**"
- G3: "**pre-stick the waterproofing-sheet**"
- G4: "**to fit the window sealing strip**".

The durations of all 4 gestures are about 8-10 seconds. It has been asked to the worker to perform 5 times these 4 gestures. Recognition performances of the system have been evaluated offline on pre-recorded isolated gestures.

In our use case, a dataset contains observations of all the 4 gestures (*co-articulated, i.e. one gesture after the other, without interruption*). In total, 5 observations for each gesture have been recorded and different learning and recognition databases have been used in 5 iterations. We have first evaluated our gesture recognition approach using the « jackknife » method. Here, jackknifing means estimation of the precision of the recognition accuracies for isolated gestures by using subsets of the available gestural data. The basic idea behind the jackknife variance estimator lies in systematically recomputing the statistical estimate leaving out one or more observations at a time from the sample set. For each iteration, one dataset (4 observations) is left out to be used as the learning database and train the HMM of each gesture, until all the datasets are used once. The rest of the datasets are used as a test database.

3. Results

3.1. Gesture recognition performances

The usual precision and recall metrics have been used to evaluate gesture recognition performances of our approach.

- **Precision** is defined by:

$$\frac{\#True\_Maximum\_Likelihood}{\#True\_Maximum\_Likelihood + \#False\_Maximum\_Likelihood}$$

where $\#True\_Maximum\_Likelihood$ is the number of test sequences for which the maximum likelihood from the HMM corresponds to the correct actually performed gesture, while $\#False\_Maximum\_Likelihood$ is the number of test sequences for which HMM output is wrong. Precision thus measures the proportion of ***correct*** results among all examples for which output by the system corresponds to the considered gesture.

- **Recall** is defined by:

$$\frac{\#True\_Maximum\_Likelihood}{\#True\_Maximum\_Likelihood + \#False\_Non\_Maximum\_Likelihood}$$

where $\#False\_Non\_Maximum\_Likelihood$ is the number of test sequences for which the HMM does not output maximum likelihood for the actual gesture. Recall therefore estimates the proportion of ***correct*** results among all examples of considered class in the recognition database.

Table 1: Precision and Recall per gesture obtained with motionPods capture, estimated by jackknifing.

		Output (maximum likelihood)				
		G1	G2	G3	G4	**Recall**
Observation (Gesture)	G1	**18**	1	1	-	90%
	G2	1	**16**	1	2	80%
	G3	-	-	**20**	-	100%
	G4	1	3	1	**15**	75%
	Precision	90%	80%	87%	88%	**86%**

Table 1 shows the results for the 5 iterations of the jackknifing for motionPods data, as well as the Precision and Recall per gesture. The average recognition performances obtained are:

- **precision ≈ 86%**
- **recall ≈ 86%**

Gesture G3 ("pre-sticking the waterproofing-sheet") is perfectly recognized, with 100% recall. Conversely, recognition rate of gesture G4 ("to fit the window sealing strip") is the worst one, with only 75% recall. One reason that may explain the very good results for gesture G3 can be the fact that the worker makes circular movements with his/her right hand without walking in his/her workspace, in contrast with the other 3 gestures. In other words, from a stochastic point of view, G3 is easier to

discriminate, and present less variability. Regarding precision, gesture G2 ("to fit on door the waterproofing-sheet") is the one for which performance is lowest (80%); the significant rate of confusion between gestures G4 and G2 can be explained by actual similarity of postures in those two gestures.

Table 2: Precision and Recall per gesture obtained with AnimaZoo™ inertial "MoCap suit" IGS-120+, estimated with jackknifing.

		Output (maximum likelihood)				
		G1	G2	G3	G4	**Recall**
Observation (Gesture)	G1	**20**	-	-	-	100%
	G2	-	**20**	-	-	100%
	G3	-	-	**20**	-	100%
	G4	-	3	-	**17**	85%
	Precision	100%	87%	100%	100%	**96%**

Table 2 shows the recognition results using AnimaZoo™ data. The average recognition performances are therefore the following:

- **precision ≈ 96%**
- **recall ≈ 96%**.

These rates are significantly higher than the 86% obtained with 3 MotionPods, which is not surprising, as the 12 inertial sensors of AnimaZoo™ half-suit provide much more information on upper-body movements. Also note that per-gesture results are coherent with those obtained with 3 motionPods: gesture G4 ("to fit the window sealing strip") has the lowest recall (85%), and gesture G2 has lowest precision (87%), with significant confusion of gesture G4 with G2.

3.2. Test of real-time early recognition

While the above performance evaluation has been done offline on isolated pre-recorded gestures, it is essential for our application to be able to recognize gestures online in real-time, as early as possible. For testing this, we analyzed the probabilities returned on every timestep for each learnt gesture, on a continuous sequence of actions. We used the AnimaZoo™ for the learning and the testing.

As we can see on top of Figure 5, there is almost always a gesture with an estimated probability around 1 issued by one of the trained HMM. However, if we look at the gesture with the highest probability in real-time (see blue curve on bottom of Figure 5), there are some fluctuations at each gesture transition due to overlapping movements that occur during the gesture sequence (gesture co-articulation). For short time periods, the gesture model with the highest probability may not correspond to the actually performed gesture. Indeed, the system needs a minimum amount of data to establish which gesture is currently being executed. These fluctuations are longer for the fourth gesture G4 because the beginning of the gesture is similar with the G2 gesture, and only the end of G4 is really characteristic of this gesture.

To prevent false recognition and add robustness to the output of our system, we process data in real-time to establish which gesture is being performed. For each timestep, we analyze instantaneous likelihoods in a time-window with a length equal to 1/3 of duration of the shortest gesture. In our experiment the duration of this time-window is 1.64 second.

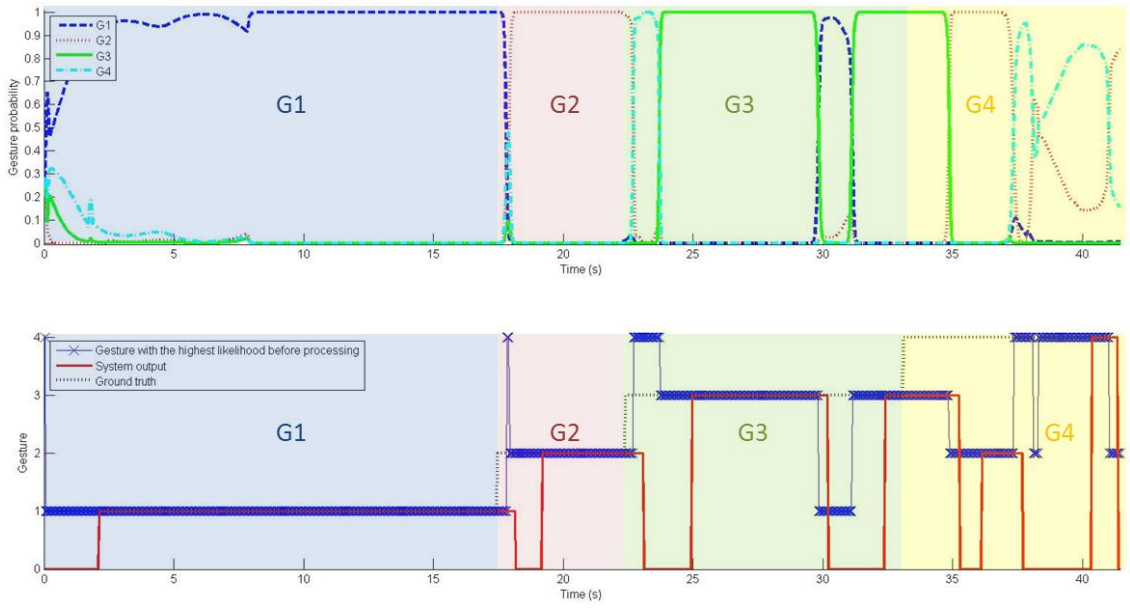

Figure 5: On top, gesture probabilities output in real-time by our recognition system. On bottom, the gesture with the highest probability in real-time (blue), the final result (estimate of performed gesture) obtained after time-window filtering (red), and the ground truth (black). The four performed gestures are successively G1, G2, G3 and G4. Their time intervals are represented with color background: blue for G1, red for G2, green for G3 and yellow for G4.

We calculate which gesture has the largest number of highest probabilities in this window, taking into account only the probabilities above 0.7. If this gesture has the highest probabilities for at least 75% of the timesteps in the window, the system returns this gesture, otherwise the system returns 0. The '0' output means that no gesture is recognized with enough certainty. As we can see on bottom of Figure 5 (red curve), our time-window filtering prevents most of the false recognitions. At the beginning of gesture G2 and gesture G3 and in the middle of gesture G3, the system returns 0. There is still an error at the beginning of the fourth gesture. The confusion between gesture G2 and gesture G4 is too long to be ignored. A solution could be to extend the window but we will take the risk to not recognize gesture G4. Note that this is totally coherent with the lower recall and precision observed for gesture G4 on tables 1 and 2. Finally, it can be noted that the average time between the actual gesture beginning and its recognition by our system is 3.4 seconds, which is an acceptable early-recognition for our application.

4. Conclusions & perspectives

In this paper, we have presented a framework and preliminary experimental results for real-time recognition of human operator actions on factory assembly-line. The goal is to enable a collaborative industrial robot, to understand behavior of humans around it, so that it can adapt its own actions and execution speed accordingly.

Our experiments shows that it is possible to obtain very good gesture recognition rates (96% precision and recall on several repetitions of the same assembly operations by a single operator), using real-time motion capture with a "MoCap suit" of 12 inertial sensors (AnimaZoo™) estimating joint angles of upper-half of human body (neck, wrists, elbows, shoulders, etc…). The great advantage of the chosen motion capture technology is total avoidance of any occlusion problems, contrary to visual-based motion capture. Also, we find that somewhat lower but nevertheless rather good recognition rates (86% precision and recall) can be obtained with only 3 inertial sensors placed respectively on the torso and the two wrists of the operator. Finally, we have tested real-time recognition capabilities of our framework using MoCap suit, during ***continuous*** sequences of actions performed by a human operator. These tests show that our system is able to perform online early recognition, after an average delay of less than 4 seconds after gesture beginning, with very little erroneous output.

Ongoing work aims at evaluating our framework for same actions recognition but on more executions by a larger pool of different human operators, and also to estimate false recognition rates on unrelated gestures. Another interesting potential perspective is the use of workers' motion capture in order to estimate effort and stress, which could be helpful for prevention of physical causes contributing to initiate or worsen some musculoskeletal disorders.

Acknowledgements

This study was carried out thanks to the financial support of "*Chaire PSA Peugeot-Citroën on Robotics and Virtual Reality*".

Authors wish to thank Pascal Ligot, from PSA, for his important contribution to the experimentations conducted in PSA experimental cell.

References

Bachmann E.R., Duman I., Usta U.Y., Mcghee R.B., Yun X.P. & Zyda M.J., *Orientation Tracking for Humans and Robots Using Inertial Sensors A Quaternion Attitude Filter*, Proc. IEEE International Symposium on Computational Intelligence in Robotics and Automation, pp 187–194, 1999.

Bar-Itzhack I., Oshman Y., Choukroun D. and Weiss H., *Direction cosine matrix estimation from vector observations using a matrix Kalman filter*, IEEE Transactions on Aerospace and Electronic Systems, pp 61–79, 2010.

Bevilacqua F., Guédy F., Schnell N., Fléty E., and N. Leroy, *Wireless sensor interface and gesture-follower for music pedagogy*, Proc. 7th international conference on New interfaces for musical expression - NIME '07, p.124, 2007.

Bevilacqua F., Zamborlin B., Sypniewski A., Schnell N., Guédy F., and Rasamimanana N., *Gesture in Embodied Communication and Human-Computer Interaction*, vol. 5934. Springer Berlin Heidelberg, pp.73–84, 2010.

Bobick A.F. and Davis J.W., *The Recognition of Human Movement Using Temporal Templates*, IEEE transactions on pattern analysis and machine intelligence (PAMI), *23*(3), pp 257–267, 2001.

Gorelick L., Blank M., Shechtman E., Irani M. and Basri R., *Actions as space-time shapes*, IEEE transactions on pattern analysis and machine intelligence (TPAMI), 29(12), pp 2247–53, 2007.

Grauman K. and Darrell T., *The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features*, proc. 10th IEEE International Conference on Computer Vision (ICCV'2005), vol.2, pp. 1458-1465, 17-21 Oct. 2005.

Inagaki Y., Sugie H., Aisu H., Ono S., Unemi T., *Behavior-based intention inference for intelligent robots cooperating with human*, proc. 4th IEEE International Joint Conference on Fuzzy Systems and 2nd International Fuzzy Engineering Symposium, vol.3, pp.1695-1700, 20-24 Mar 1995.

Kellokumpu V. and Pietikäinen M., *Human Activity Recognition Using Sequences of Postures*, Proceedings of the IAPR Conference on Machine Vision Applications (IAPR MVA), 2005.

Kim J.-M. and Song M.-K., *Three Dimensional Gesture Recognition Using PCA of Stereo Images and Modified matching Algorithm*, proc. International Conference on Fuzzy Systems and Knowledge Discovery, pp 116–120, 2008.

Laptev I. and Lindeberg T., *Space-time interest points*, proc. 9th IEEE International Conference on Computer Vision (ICCV'2003), pp 432–439, 2003.

Laptev I. and Perez P., *Retrieving actions in movies*, proc 11th IEEE International Conference on Computer Vision (ICCV'2007), pp 1–8, 2007.

Lepetit V., Lagger P. and Fua, P., *Randomized Trees for Real-Time Keypoint Recognition*, proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2005), pp. 775–781, 2005.

Lovell B.C., Kootsookos P.J., and Davis R.I., *Model Structure Selection & Training Algorithms for an HMM Gesture Recognition System*, proc. 9th International Workshop on Frontiers in Handwriting Recognition, 2004.

Lv F. and Nevatia R., *Single View Human Action Recognition using Key Pose Matching and Viterbi Path Searching*, proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2007), pp 1–8, 17-22 June 2007.

Oka K., Sato Y. and Koike H., *Real-time tracking of multiple fingertips and gesture recognition for augmented desk interface systems*, proc. 5th *IEEE* International Conference on Automatic Face and Gesture Recognition, pp.429-434, 21 May 2002.

Piovan G. and Bullo F., *On Coordinate-Free Rotation Decomposition: Euler Angles About Arbitrary Axes*, IEEE Transactions on Robotics, 28(3), pp 728–733, 2012.

Rabiner L.R., *A tutorial on hidden Markov models and selected applications in speech recognition*, Proc. IEEE, vol. 77, no. 2, pp. 257–286, 1989.

Schuldt C., Laptev I. and Caputo B., *Recognizing Human Actions: A Local SVM Approach*, proc. IEEE International Conference on Pattern Recognition (ICPR'2004), pp 3–7, 2004.

Shotton J., Fitzgibbon A., Cook M., Sharp T., Finocchio M., Moore R., Kipman A. and Blake A., *Real-time human pose recognition in parts from single depth images*, proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR), 2011.

Sullivan J. and Carlsson S., *Recognizing and Tracking Human Action*, proc. European Conference on Computer Vision (ECCV), 2002.