Report: BMVA Technical Meeting 'Vision for language and manipulation'

The BMVA Technical Meeting 'Vision for language and manipulation' was held at the BCS headquarters in London. It was organised by Andrew Gilbert (University of Surrey), and chaired by Nick Hockings (University of Bath) & Walterio Mayol-Cuevas (University of Bristol). This proved to be a very substantial affair, organised into 4 sessions each commencing with a 30 minute presentation from a keynote speaker followed by a number of briefer 15 minute presentations. In all 20 talks were presented - almost a "mini conference"!

The speakers explored state-of-the art developments related to sensing and manipulation, grasping, human-robot interaction, pose estimation and learning. There were presentations from academics from around the UK, as well as some key figures in the field from across Europe and the audience was similarly representative of the vision and robotics communities.

Session 1:
The day began with a keynote presentation by **Angelo Cangelosi** exploring embodiment. He showed the use of an iCub to demonstrate association learning between sensing modalities to learn language primitives and counting. Of particular interest was the topology of the classification hierarchies for representing numbers learnt by the robots if they had not other input compared to the hierarchy learnt if the robot visualised counting fingers.

Robert Haschke then discussed a real-time, model-free scene segmentation approach solely based on depth information captured by means of a Kinect camera. A normals-based homogeneity criterion is used to find smooth patches which are then grouped on order to establish object hypotheses. A set of heuristic rules is used to parse a connectivity graph into putative objects, which are labelled by means of an NN classifier trained on colour histograms and shape features. These labelled object classification hypotheses can then be used to mediate tasks based on object shape, colour and location and finally for the execution of the request in terms of manipulation, e.g. "Put this green apple into the left, big basket!"

Aude Billard presented work on planning fine manipulation of objects based on vision and haptic guided human demonstration. In this work bimanual compliant tactile exploration is presented "where a tactile exploration strategy is proposed to guide the motion of the two arms and fingers along the object"

Gerardo Aragon-Camarasa gave a talk on cloth perception and manipulation using a dynamically actuated binocular robot head. Using high-resolution imaging for 2.5D range mapping it is possible to characterise and describe the topology of garments for grasping and flattening manipulation tasks. The talk also over-viewed work on "real-time" GPU stereo matching, feature extraction for gaze control and camera vergence. An overview of real-life robotic scenarios as part of the CloPeMa european project was presented, demonstrating the versatility and robustness of the methods.

Nicola Notcetti described the results of an investigation into discriminating between biological motion versus non-biological motion for the purpose of understanding human interaction. This approach is based on the specific relation between shape and motion "2/3 law" for biological motion. The task of estimating the weight of an object in from visual observation served as the focus of the investigation.

Session 2:
The opening keyonte by **Sinan Kalkan** addressed the key theme of the workshop: Vision, Language and Manipulation, addressing "Learning and conceptualizing word categories in language such as verbs, nouns and adjectives from and manipulation..." To this end a good working framework was presented, based on a classifier to learn and predict object affordances.

Norbert Kruger discussed "the problem of how to bridge from low-level sensory data to symbolic representations." He proceeded to then give a whirl-wind overview of biological vision based on his August 2013 PAMI paper. His talk concluded with an overview of his work addressing "...learning associations between low-level motor-sensory information and symbolic representations."

Frank Foerster presented what can only be described as an uncanny (no pun intended...) demonstration of symbol ground through verbal interaction with the iCub robot, in order to ground the meaning of the word "no", amongst others. Taken at face value, the interaction of the human with the machine, based on unconstrained natural language communication, attempted object interaction/manipulation and negation from the human trainer was immensely impressive. This demonstration was the stuff of sci-fi and I was quite shocked that the field had advanced to this degree of human-machine interaction, which at least superficially, mimicked a young child interacting with an adult.

To end the second session **Waltero Mayol Cuevas** gave a talk on human machine interaction using a very cute hand-held cognitive robot. This tenticle-like robot essentially serves as an intelligent tool that can communicate with the user by pointing, not pointing or refusing to undertake an inappropriate action. A user evaluation of the utility of the robot was then over-viewed.

Session 3:
The third session opened with a fascinating keynote talk by **Marco Davare** who describes experiments that use trans-cranial magnetic stimulation to induce "virtual lesions" in the brains of healthy volunteers. Using this methodology it becomes possible to probe "which parts of the brain are causally involved in integrating visual and tactile cues during action." Using these techniques within a VR environment, where vision and touch can be controlled, a specific network associated with grasping actions has been defined that appears to be responsible for "...the integration between a sensiomotor memory, the object properties and online visual cues..."

Giorgio Metta presented work on mapping sensory features to motor invariants for automated action recognition. These motor invariants are used to improve action discrimination while replicating human movements. By combining a reduced set of vision and auditory features, discrimination of actions become more effective than single-large databases of visual or auditory features. The talk concluded by presenting experimental experiments of these motor invariants in the iCub humanoid robot.

The session then continued with **Giovanni Saponaro** talking about a probabilistic framework to learn object affordances while executing actions. This framework builds on the idea that a robot should explore its environment by interacting with objects and, consequently, assessing object's functionality against "affected objects", e.g. the utility of a hammer while hammering nails. In this case, objects are not treated as abstract concepts with labels but simple low-level features are used to generalise previous knowledge for object affordances.

John Darby talked about a framework to track objects while human-users interact with them. A generative and discriminative tracking framework is developed in order to embed object-pose and body-poses into a mutual and relative coordinate frames. This allows computational systems to reason about "human-object interactions and their outcomes" based on action events over three different contexts, i.e. object-pose, human action and scenes.

Before going to coffee break and biscuits, **Vincenzo Lippiello** discussed a control system for aerial "manipulation" based on vision. In this work, image moments are used to control visual servoing actions while aligning the targeted object and a manipulator mounted on unmanned aerial vehicle.

Session 4:

The keynote talk for the final session of the day was given by **Jan Peters** on the subject of learning motor skills. Jan outlined a general framework for learning robot motor skills applied to manipulation of both static and dynamic objects perceived by means of computer vision. This approach was convincingly demonstrated in video footage showing robots which had learned a number of tasks by imitation including, playing table-tennis and also the “cup-and-ball” game.

Following Jan Peter's inspiring talk, **Jeremy L Wyatt** talked about methods for manipulation using grasp adaptations, active vision and active touch. This work is part of the PaCMan european project in which manipulation interactions are explored in order to reduce uncertainties while interacting with objects. The take-home message was that “...vision for manipulations should ideally be tackled as part of a coherent theory of a larger system...”.

Patricia Shaw presented a biologically and psychologically inspired system that is able to learn gaze and reach control. From observation of the developmental stages in infants, the iCub robot is able to learn to control its eyes, then its gaze space and finally, hand and reach control behaviours. Preliminary work on hand-orientation for pre-reaches towards objects was overviewed.

Panos Trahanias discussed a markless articulated upper human body tracking system for multiple users in RGB-D sequences. This system is able to track different users in real-life settings containing severe intra- and inter-personal occlusions. An experimental comparison to well-known standard methods such as NiTE demonstrated the robustness of the tracking system.

Pietro Falco presented results from the EU FP7 DEXMART project. The results he presented included a description of a Bayesian sensor fusion approach for perception and interpretation of human manipulation, used to map human behaviours into a robotic system. By combining visual information and angular-motion information over different levels of complexity of the described sensor fusion system, a robot is able to grasping and manipulate objects with a high accuracy and precision, e.g. “unscrewing a bottle top”.

The final presentation of the day was given by **Antonis Argyros** who overviewed a framework for hand-tracking. The presented approach centred on first detecting and tracking the hands an fingers. Particle Swarm Optimisation is then used to obtain a consistent interpretation of the 3D position orientation and full articulation of the hand. When the hand is considered in isolation, the optimisation attempts to register the known 3D structure of the hand with its observed appearance. When the hand is holding an object, the process optimises the interpretation of a joint object-hand model. A number of applications of this approach to humanoid robot learning by demonstration and interactive exhibits in smart environments concluded this presentation.

Conclusions:
Overall, the organisers must be commended for the resounding success of their meeting. The talks were of high quality, thought inspiring, and yet they remained accessible to junior academics. A number of attendees openly expressed their praise for the organisation and execution of the meeting.

The meeting appeared to be successful at sparking conversation, with a healthy amount of audience participation, which was always constructive, maintaining a positive atmosphere throughout. The conversations carried on over the lunch (which was prompt, if a little short). The venue for dinner was conveniently located, and service was prompt thanks to pre ordering arrangements, and whilst billing was slow, the abundant and vivacious conversation eclipsed most of the extended wait. A particular highlight of the dinner was Nick Hockings removing a bio-mimetic model of a dog's leg from his rucksack and offering it round the table “for a feel” of this uncomfortably realistic robo-dog paw (fortunately most folks at the table already eaten their meal by this point :)

My overriding impression (PS) was that this meeting had effectively revealed the state-of-the-art in cognitive robotic learning and that the state of play is much more advanced than I had anticipated. I truly envy young researchers entering this field, for this is surely going to be ***the*** most exciting and adventurous field to work in over the next 10-20 years!

Gerardo Aragon-Camarasa, Tadeo Corradi and Paul Siebert, 11 September 2014.