Seeing and Understanding People

Richard Bowden
Centre for Vision Speech and Signal Processing
University of Surrey, Guildford, Surrey, GU2 7XH, UK
r.bowden@surrey.ac.uk
www.ee.surrey.ac.uk/Personal/R.Bowden

ABSTRACT: This manuscript and associated talk gives an historical but not exhaustive overview of work in the Cognitive Vision Lab at the University of Surrey's Centre for Vision Speech and Signal Processing. Work concentrates on people, tracking or identifying their actions and interpreting the meaning of those actions. To do this we employ techniques from a variety of sources which include the use of Mutual Information in Tracking, Data mining in Learning and using linguistics in classification. This Manuscript covers approaches to General Tracking, Multutal Information Estimation, Human Pose Estimation, Head and Hand Tracking, Expression Recognition, Lip Reading, Non Verbal Communication, Sign Language Recognition and Activity Recognition.

1 INTRODUCTION

Computer vision has its roots in Artificial Intelligence, but over the past two decades has firmly established itself as a research field in its own right. Related areas have their own communities but we all share a substantial body of techniques and terminology. Computer/cognitive vision has moved beyond image processing with classification and regression techniques developed in the machine learning community predominant in current state-of-the-art. Another community which shares many techniques, but typically operating in isolation, is that of data mining.

This manuscript gives a overview of work in the Cognitive Vision Lab within the Centre for Vision Speech and Signal Processing at the University of Surrey and demonstrates how techniques from different disciplines can be used to tackle common problems. The common theme is *seeing and understanding people* which includes head and hand tracking, sign language recognition, expression recognition, non-verbal communication and more general activity recognition. A common approach being weakly supervised learning using many techniques inspired from the data mining community.

2 FROM MUTUAL INFORMATION TO TRACKING

Our interest in mutual information (MI) came from the same properties that have made it an important technique in medical image registration, its ability to register two images from different modalities. In terms of 2D tracking, typically consecutive frames do not come from different imaging modalities but due to lighting variation and the properties of the object surface, the relationship between pixels in two consecutive images can be far from linear. One of the earliest and widely used techniques for matching image patches between frames was proposed by Lucas and Kanade (Lucas and Kanade 1981).

LK matching typically employs simple brightness constancy assumptions and uses Sum of Squared Difference (SSD). We chose to base our tracking on MI because of its robustness to environmental lighting/noise, pronounced maxima and similar computation cost to SSD. Our earliest attempt at using MI in a tracking context was the M^3I tracker (Dowson and Bowden 2004) which developed into the Simultaneous Modelling and Tracking (SMAT) algorithm (Dowson and Bowden 2005) (Dowson and Bowden 2006b). SMAT was an on-line tracking algorithm that, given a single image patch in the first frame, would track and learn a hierarchical constellation model of appearance and structure on the fly. As such, it builds a model of appearance variation as it tracks, becoming more robust over time. Tracking was performed in an optimised LK framework but using MI as the similarity measure.

Work by (Baker and Matthews 2004) revolutionized LK when they proposed the inverse compositional method. The key to the approach was posing the warp function as a function of two warps and inverting the roles of the template and image. This allowed an approximation of the Hessian to be derived

that was solely based on the template. As the template is typically constant, the Hessian can be precomputed and this decreases the complexity of each iterative update. In (Dowson and Bowden 2006a, Dowson and Bowden 2008) we presented a single mathematical framework for an inverse compositional approach to MI for four common variants including Standard Sampling, Partial Volume Estimation, In- and Post-Parzen Windowing. However, our work highlighted problems with PDF estimation due to the discrete nature of the underlying histograms used and the sparsity of samples when applied in 2D.

The histogram accuracy, and hence registration accuracy, is limited by the quantization of intensity and number of samples available. For volumetric data, the number of samples are high, but in 2D, histograms are typically under populated. This is a well understood problem, with a considerable body of work devoted to In Parzen or post Parzen windowing, Partial Volume (PV) Interpolation or PV Estimation (Dowson et al. 2008) all of which attempt to overcome these issues but in some cases actually introduce bias due to the kernels used. In (Dowson et al. 2008) we applied Non-Parametric (NP) Windows to the problem of estimating the joint statistics of images, equivalent to sampling at a high (infinite) resolution for an assumed interpolation model. This overcomes sampling issues and introducing less bias than other approaches.

3 TRACKING VS PREDICTION

One of the fundamental problems with LK is that it relies on the appearance of a template. It is posed as an optimization problem where some metric (e.g. SSD or MI etc.) is used to calculate the warp between a template and image. Models like SMAT allow variation in the template, gradually incorporating change into the model. But there are a whole class of problem where appearance change is so radical, that template based approaches cannot cope. Furthermore, tracking is limited by the basin of convergence of the optimization approach meaning that the motion between frames must be small. Multi-scale approaches can help or we can abandon optimization in favour of treating tracking as an offset prediction problem.

Linear Predictors (Matas et al. 2006) are a simple displacement predictor that maps a sparse set of *support* pixels, to a displacement in the image. The relationship is a simple linear mapping between pixel intensities and translational displacement learnt through synthetically offsetting a tracker during training. In (Ellis et al. 2007) we integrated this predictor approach into the SMAT algorithm, later developing more robust partitioning of the appearance modes and demonstrated how banks of different predictors could be used for different appearances of an object (Ellis et al. 2008) (Ellis et al. 2011).

Perhaps one of the most important aspects of linear predictor tracking is their ability to make predictions

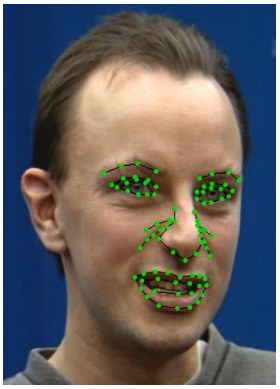

Figure 1: Facial Feature Tracking

from a varied support region. Consider the problem of tracking the face. Figure 1 shows several features that one might want to track. Standard features consist of the corner of the eyes and mouth. The key word being *corners*. Corners are easy to track, they are well localized, robust to scale and remain consistent in terms of appearance. It is perhaps unsurprising then that many approaches to facial feature extraction employ such easily detected landmarks. However, if one considers the contour of the lips, the problem is more complex. Tracking a point on an edge suffers from the aperture problem, where edge points are only well defined in one direction, perpendicular to the line. Points on the inner lip are even more problematic, as the texture can change dramatically as the mouth opens and closes. Perhaps the most challenging task is some arbitrary point on the cheek. Assuming the resolution of the image is insufficient to see micro texture or the pores of the skin, there is no information with which to track. Linear predictors can overcome this problem. If the motion of any given point can be modelled by its relationship to other points in the image that are well localized, then a predictor can be constructed. The key idea here is selection of support: *can we find points in the image which allow a linear displacement predictor to be constructed?*

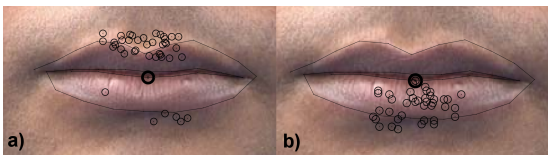

Figure 2: Linear Predictor Support Selection. a) Horizontal Predictor, b) Vertical Predictor

In (Ong and Bowden 2008) we proposed such a selection framework that allows a learning framework to choose the best visual support regions for any specific feature point and motion. Figure 2 shows the selection for a point on the inner, lower lip depicted by the dark circle. Predictors are separated into horizontal (Fig.2a) and vertical prediction (Fig.2b). Lighter circles show the flock of linear predictors selected for motion prediction. Note that although we are tracking the lower lip, the approach selects support from the upper lip to localize in the horizontal direction as the

structure of the upper lip as this is a good feature with which to localize horizontally. In (Fig.2b), the selection procedure chooses support from the lip and chin to localize vertically, away from mouth itself which can change so drastically in appearance.

In (Ong et al. 2009) and (Ong and Bowden 2011c) we developed this approach into a tracker capable of tracking any facial feature, using hierachical predictors to provide robust and accurate tracking. The underlying linear mathematical assumptions in the approach provide an efficient solution.

While this tracking methodology has been used in much of our lip-reading work[1], we have recently proposed a non-linear version based on regression trees (Sheerman-Chase et al. 2013). Replacing the underlying linear assumption with a nonlinear model overcomes some of the limiting assumptions. This newer version is more robust to head pose, requires less training data, is more resilient to lighting while retaining computational efficiency.

While this allows tracking of features with variable or no visual appearance, it is only possible where some mathematical relationship to other features can be established. There is another class of problem where features simply do not exist. To tackle this, our most recent work has developed FLO-track, a featureless tracking algorithm that uses line correspondences within a SMAT like tracking framework (Lebeda et al. 2013). Using low-level line correspondences in tracking allows operation even when there is a lack of texture. While such approaches work well for tracking objects with relatively consistent appearance, such as faces, tracking highly deformable objects such as people or hands requires a different approach.

4 TRACKING AND DETECTING PEOPLE

Tracking people in the context of surveillance is typically done using static camera assumptions (KaewTrakulPong and Bowden 2003, KaewTraKulPong and Bowden 2002). However, such approaches lack the fidelity required to recognize activity and typically concentrate on more general behaviour. Simple approaches to identifying behaviour can be used as priors during tacking when objects are occluded (KaewTraKulPong and Bowden 2004) or moving between cameras (Bowden and Kaewtrakulpong 2005). In (Gilbert and Bowden 2005), (Gilbert and Bowden 2006), (Gilbert and Bowden 2008) we developed approaches to self calibrating distributed camera networks using the people moving between cameras as the calibration targets by looking for statistical trends in weakly correlated motion cues.

Body part detection became popular when Viola and Jones (Viola and Jones 2004) proposed an efficient method for head detection and it is relatively simple to extend the approach to other body parts such as the torso (Micilotta and Bowden 2004) or hands (Ong and Bowden 2004). Detecting parts in isolation has many advantages over full body detection as independance reduces the complexity of the detector. Overall structure can then be applied after detection using probabilistic body part assembly (Micilotta et al. 2005). Such approaches have gained popularity since pictorial structures were reformulated (Felzenszwalb and Huttenlocher 2005) allowing an efficient framework for part based modelling.

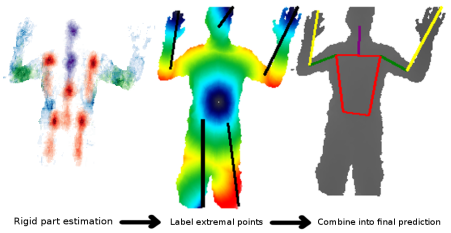

Figure 3: Pose Estimation via Regression and Geodesic Extrema

More recently, the introduction of the Microsoft Kinect™ has resulted in an explosion in approaches that employ depth. Our first work with the Kinect was to apply poselets (Bourdev et al. 2010) in the depth domain (Holt et al. 2011). Although part detection can still be used, as in the seminal work of (Shotton et al. 2011), we chose to adopt direct regression based approaches (Holt and Bowden 2012), (Holt et al. 2013), the later of which combines regression of joints with the identification of geodesic extrema. As seen in Figure 3, regression works well for the torso but degrades as the degrees of freedom of the body parts increase, leading to poor hand prediction. However, the hands form extrema which can be efficiently computed by treating the depth map as a geodesic surface using Dikstra's algorithm.

The concept of geodesic extrema can also be applied to the hands allowing fingertip extraction to be performed (Krejov and Bowden 2013). Figure 4 shows geodesic extrema computed on a depth image of the hand. The advantage of operating in the depth domain being that discontinuities in object segmentation through self occlusion can be identified more easily and corrected for. This approach to tracking fingertips is extremely fast allowing the fingers of up to 4 hands to be tracked in real time and forms the input to our work on MultiTouchless interfaces[2].

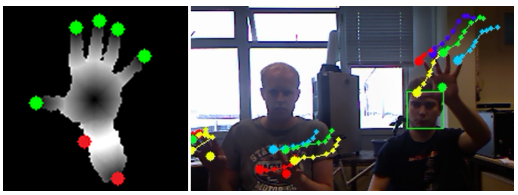

Figure 4: Pose Estimation via Regression and Geodesic Extrema

5 SIGN LANGUAGE RECOGNITION

Although the Kinect™ plays a key role in providing robust real-time Sign Language Recognition (SLR)

[1] http://www.youtube.com/watch?v=Tu2vInqqHX8

[2] www.ee.surrey.ac.uk/Personal/R.Bowden/multitouchless/

demonstration systems, our work in this area predates the sensor considerably.

Sign consists of three main parts: Manual features involving gestures made with the hands, Non-manual features such as facial expressions or body posture, which can form part of a sign or modify the meaning of a sign, and Finger spelling, where words are spelt out in the local verbal language. Naturally this is an over simplification, sign language is as complex as any spoken language and each sign language has many thousands of signs, differing from the next by minor changes in hand shape, motion, position, non-manual features or context. It also has its own grammar.

To date, most work in the literature has concentrated on the manual aspects of sign or the simpler problem of finger spelling (Cooper et al. 2011). Our own work on finger spelling is limited to (Bowden and Sarhadi 2002) and (Pugeault and Bowden 2011) as it is more an artefact of modelling hand shape as part of continuous sign than working on the problem *per se*.

Although we know the importance of non-manual features in communication, this is also something we have yet to integrate successfully into SLR. However, we have investigated facial expression recognition (Moore et al. 2010), (**?**); the effects of pose on expression recognition (Moore and Bowden 2011), (Moore and Bowden 2009); and non verbal communication during speech (Sheerman-Chase et al. 2009), (Sheerman-Chase et al. 2011).

Any SLR system needs to recognise thousands of different signs. As such the simple approach of training a classifier per sign soon becomes intractable especially when one considers the training requirements needed to cope with natural variability between individuals, motion epenthesis and co-articulation. The emergent solution in speech was to recognise the subcomponents (phonemes), then combine them into words using Hidden Markov Models (HMMs). Sub-unit based SLR uses a similar two stage approach, sign linguistic sub-units are identified and sub-units combined together to create a sign level classifier.

Our early work in this area turned to the linguistic annotation used in the British Sign Language (BSL) Dictionary which used a HA (hand arrangement), TAB (hand position), SIG (hand movement), all of which are relative measures and DEZ (hand shape). A set of deterministic rules converted incoming tracking data into a symbol sequence based on these linguistic descriptors (Bowden et al. 2004), (Kadir et al. 2004). A second stage classification was then used to recognise the temporal ordering of the symbols that corresponded to a particular sign. This provided huge advantages. As the initial stage of classification generalise well, models could be trained with as little as 1 example. Despite its simplicity, its legacy remains with us today, however, the evolution of the approach now allows us to tackle far higher lexical sizes with better generalization between people. We have attempted various approaches to overcoming tracking failure and noise in the initial stage of classification which is often a limiting factor for fast and/or subtle hand motion (Cooper and Bowden 2007) (**?**) a good overview is given in (Cooper et al. 2012).

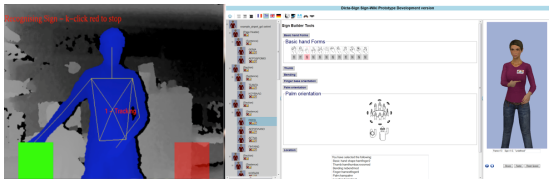

Figure 5: Pose Estimation via Regression and Geodesic Extrema

Within the EU project Dictasign, a Sign Wiki application was developed . The system incorporated a recognition engine based on a kinect sensor, editing software and an avatar for replay (Efthimiou et al. 2012). Maintaining a 2 stage classification architecture, the initial level was based on HamNoSys[3] and second stage classification based on markov chains (Cooper et al. 2011).

In (Ong and Bowden 2011b) and (Ong and Bowden 2011a) we developed Sequential Pattern recognition primarily for lip reading, but employed this classification approach in the final versions of the Sign Wiki recognition engine. The technique identifies patterns by performing spatio temporal feature selection to find minimal signatures that are both distinctive and discriminative. Although initially developed for lip reading as binary classifiers, in (Ong et al. 2012) we developed a multiclass approach, *sequential pattern trees* which provides excellent state-of-the-art performance by combining aspects of classical machine learning with efficient tree pruning strategies taken from data mining.

In (Cooper and Bowden 2009) we identified signs from broadcast footage using the subtitles as weak supervision. To achieve this, we used an adapted version of the *a priori* data mining algorithm to identify co-occurring motions in the sign stream that correspond to possible repetitions of words in the subtitles. The process is weakly supervised as there is no guarantee that a sign will be present and the temporal offset between subtitle and sign is unknown. The approach was able to automatically identify signs without user intervention or ground truth labelling. More recent work attempts to automatically identify sub-units of sign for training using an iterative forced alignment algorithm to transfer the knowledge of a user edited open sign dictionary to the task of annotating a challenging, large vocabulary, multi-signer corpus recorded from public TV (Koller et al. 2013).

A priori mining has become an important tool in our learning frameworks. Commonly know as the *shopping basket algorithm* its ability to process extremely large amounts of data to find co-occurring

[3]The **Ham**berg **No**tation **Sys**tem (HamNoSys) is a "phonetic" transcription system, which has been in widespread use by Sign linguists for over 20 years.

symbols directly lends itself to large scale video learning. We have used this algorithm to identify subtle social signals in videos of people conversing and to identify participant interest in a topic from body motion (Okwechime et al. 2011b). Rules can also be used in animation (Okwechime et al. 2011a). We have also used it to find the relationship between perception and action in the context of learning autonomous control in robotics (Ellis et al. 2011), but one of our largest applications has been in its use in action recognition.

6 ACTION RECOGNITION

A priori is ideally suited to activity/action recognition as datasets typically provide positive and negative examples of the action but do not specify when or where the important information is located. In its native form *a priori* calculates co-occurrence statistics, so we force the algorithm to find items that are both frequent and discriminative. This is achieved by appending features from positive and negative examples with a symbol that delineates its source class and then extracting rules that co-occur with the positive symbol. Our activity recognition approach starts with low level corners in 3 different planes: (x, y), (x, t) and (y, t). This makes features more dense than normal interest point detectors, a single short video can contain millions of features. Each corner is encoded relative to its neighbours and mining performed to find small spatio-temporal structures that are both frequent in the positive example and infrequent in the negative data i.e. discriminative. The process is repeated hierarchically using the features from the last stage in a wider encoding. As the spatiotemporal structures become more complex, they become more accurate in both classification and localisation and because they are based on collections of simple corners, they are extremely quick to compute, see (Gilbert et al. 2008), (Gilbert et al. 2009) and (Gilbert et al. 2011).

While action recognition *in the wild*, involving broadcast footage, has become prevalent in the literature, recognition is still performed in 2D. However, there is a growing source of 3D footage available and our recent dataset Hollywood3D (Hadfield and Bowden 2013) provides an action recognition dataset taken from Hollywood films but with dense stereo depth available. This additional 3D information can be incorporated in classification to improve classification performance. Our current work is to apply our Scene Particles algorithm (Hadfield and Bowden Nov) to this dataset. Scene Particles allows the efficient computation of Scene Flow, the 3D motion field, which will provide richer 3D features for classification and scene understanding.

REFERENCES

Baker, S. & I. Matthews (2004). Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision 56*(3), 221–255.

Bourdev, L., S. Maji, T. Brox, & J. Malik (2010). Detecting people using mutually consistent poselet activations. In *European Conference on Computer Vision (ECCV)*.

Bowden, R. & P. Kaewtrakulpong (2005). Towards automated wide area visual surveillance: tracking objects between spatially-separated, uncalibrated views. *Vision, Image and Signal Processing, IEE Proc. - 152*(2), 213–223.

Bowden, R. & M. Sarhadi (2002). A non-linear model of shape and motion for tracking finger spelt American sign language. *Image and Vision Computing 20*(9-10), 597–607+.

Bowden, R., D. Windridge, T. Kadir, A. Zisserman, & J. M. Brady (2004). A linguistic feature vector for the visual interpretation of sign language. In *Euro Conf. on Comp Vis*.

Cooper, H. & R. Bowden (2007). Large lexicon detection of sign language. In *Human Computer Interaction*, Volume 4796 of *LNCS*, pp. 88–97.

Cooper, H. & R. Bowden (2009). Learning signs from subtitles: A weakly supervised approach to sign language recognition. In *Comp Vis and Pat Rec, 2009. CVPR 2009. IEEE Conf. on*, pp. 2568–2574.

Cooper, H., B. Holt, & R. Bowden (2011). Sign language recognition. In *Visual Analysis of Humans*, pp. 539–562.

Cooper, H., E.-J. Ong, N. Pugeault, & R. Bowden (2012, Jul). Sign language recognition using sub-units. *Journal of Machine Learning Research 13*, 2205–2231.

Cooper, H., N. Pugeault, & R. Bowden (2011). Reading the signs: A video based sign dictionary. In *Comp Vis Workshops (ICCV Workshops), 2011 IEEE Int. Conf. on*, pp. 914–919.

Dowson, N. & R. Bowden (2004). Metric mixtures for mutual information (m3i) tracking. In *Pat Rec, 2004. ICPR 2004. Proc. of the 17th Int. Conf. on*, Volume 2, pp. 752–756 Vol.2.

Dowson, N. & R. Bowden (2006a). A unifying framework for mutual information methods for use in non-linear optimisation. In *Europen Conf. Comp Vis ECCV 2006*, Volume 3951 of *LNCS*, pp. 365–378.

Dowson, N. & R. Bowden (2008). Mutual information for lucas-kanade tracking (milk): An inverse compositional formulation. *PAMI, IEEE Trans. on 30*(1), 180–185.

Dowson, N., T. Kadir, & R. Bowden (2008). Estimating the joint statistics of images using nonparametric windows with application to registration using mutual information. *IEEE Trans. on PAMI 30*(10), 1841–1857.

Dowson, N. D. H. & R. Bowden (2005). Simultaneous modeling and tracking (smat) of feature sets. In *Comp Vis and Pat Rec, 2005. CVPR 2005. IEEE Computer Society Conf. on*, Volume 2, pp. 99–105 vol. 2.

Dowson, N. D. H. & R. Bowden (2006b). N-tier simultaneous modelling and tracking for arbitrary warps. In *BMVC'06*, pp. 569–578.

Efthimiou, E., S.-E. Fotinea, T. Hanke, J. Glauert, R. Bowden, A. Braffort, C. Collet, P. Maragos, & F. Lefebvre-Albaret (2012). The dicta-sign wiki: Enabling web communication for the deaf. In *Computers Helping People with Special Needs*, Volume 7383 of *LNCS*, pp. 205–212.

Ellis, L., N. Dowson, J. Matas, & R. Bowden (2007). Linear predictors for fast simultaneous modeling and tracking. In *In submitted to Workshop on Non-rigid Registration and Tracking through Learning, Eleventh IEEE Intl. Conf. Comp Vis*, pp. 1–8.

Ellis, L., N. Dowson, J. Matas, & R. Bowden (2011). Linear regression and adaptive appearance models for fast simultaneous modelling and tracking. *Int. Journal of Comp Vis 95*, 154–179. 10.1007/s11263-010-0364-4.

Ellis, L., M. Felsberg, & R. Bowden (2011). Affordance mining: Forming perception through action. In *Proc. of the 10th Asian Conf. on Comp Vis - Volume Part IV*, Volume 6495 of *LNCS*, pp. 525–538.

Ellis, L., J. Matas, & R. Bowden (2008). Online learning and partitioning of linear displacement predictors for tracking. In *BMVC08*, Volume 1, pp. 33–43. BMVA.

Felzenszwalb, P. F. & D. P. Huttenlocher (2005, January). Pic-

torial structures for object recognition. *Int. J. Comput. Vision 61*(1), 55–79.

Gilbert, A. & R. Bowden (2005). Incremental modelling of the posterior distribution of objects for inter and intra camera tracking. In *Proc. of BMVC.*, Volume 1, pp. 419–428.

Gilbert, A. & R. Bowden (2006). Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In *Comp Vis ECCV 2006*, Volume 3952 of *LNCS*, pp. 125–136.

Gilbert, A. & R. Bowden (2008). Incremental, scalable tracking of objects inter camera. *Comp Vis and Image Understanding 111*(1), 43 – 58.

Gilbert, A., J. Illingworth, & R. Bowden (2008). Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In *Proc. of the 10th Euro Conf. on Comp Vis: Part I*, ECCV '08, pp. 222–233.

Gilbert, A., J. Illingworth, & R. Bowden (2009). Fast realistic multi-action recognition using mined dense spatio-temporal features. In *Comp Vis, 2009 IEEE 12th Int. Conf. on*, pp. 925–931.

Gilbert, A., J. Illingworth, & R. Bowden (2011). Action recognition using mined hierarchical compound features. *IEEE Trans. on PAMI 33*, 883–897.

Hadfield, S. & R. Bowden (2013). Hollywood 3d: Recognizing actions in 3d natural scenes. In *Proceeedings, Conf. on Comp Vis and Pat Rec*, Portland, Oregon.

Hadfield, S. & R. Bowden (Nov.). Kinecting the dots: Particle based scene flow from depth sensors. In *Comp Vis (ICCV), 2011 IEEE Int. Conf. on*, pp. 2290–2295.

Holt, B. & R. Bowden (2012). Static pose estimation from depth images using random regression forests and hough voting. In *VISAPP (1)*, pp. 557–564.

Holt, B., E.-J. Ong, & R. Bowden (2013). Accurate static pose estimation combining direct regression and geodesic extrema. In *10th IEEE Int. Conf on Face and Gesture Recognition FG2013*.

Holt, B., E.-J. Ong, H. Cooper, & R. Bowden (2011). Putting the pieces together: Connected poselets for human pose estimation. In *Comp Vis Workshops (ICCV Workshops), 2011 IEEE Int. Conf. on*, pp. 1196–1201.

Kadir, T., R. Bowden, E. J. Ong, & A. Zisserman (2004). Minimal training, large lexicon, unconstrained sign language recognition. In *BMVC*.

KaewTraKulPong, P. & R. Bowden (2002). An improved adaptive background mixture model for real-time tracking with shadow detection. In *Video-Based Surveillance Systems*, pp. 135–144.

KaewTrakulPong, P. & R. Bowden (2003). A real time adaptive visual surveillance system for tracking low-resolution colour targets in dynamically changing scenes. *Image and Vision Computing 21*(10), 913 – 929.

KaewTraKulPong, P. & R. Bowden (2004). Probabilistic learning of salient patterns across spatially separated, uncalibrated views. In *Intelligent Distributed Surveilliance Systems, IEE*, pp. 36–40.

Koller, O., H. Ney, & R. Bowden (2013). May the force be with you: Force-aligned signwriting for automatic subunit annotation of corpora. In *IEEE Int. Conf. on Face and Gesture Recognition*. Shanghai, PRC.

Krejov, P. & R. Bowden (2013). Multitouchless: Real-time fingertip detection and tracking using geodesic maxima. In *10th IEEE Int. Conf on Face and Gesture Recognition FG2013*.

Lebeda, K., J. Matas, & R. Bowden (2013). Tracking the untrackable: How to track when your object is featureless. In J.-I. Park and J. Kim (Eds.), *Comp Vis - ACCV 2012 Workshops*, Volume 7729 of *LNCS*, pp. 347–359.

Lucas, B. D. & T. Kanade (1981). An iterative image registration technique with an application to stereo vision. In *In Proc. Intl Conf. on Artificial Intelligence*, pp. 674–679.

Matas, J., K. Zimmermann, T. Svoboda, & A. Hilton (2006). Learning efficient linear predictors for motion estimation. In P. Kalra and S. Peleg (Eds.), *Computer Vision, Graphics and Image Processing*, Volume 4338 of *Lecture Notes in Computer Science*, pp. 445–456. Springer Berlin Heidelberg.

Micilotta, A. S. & O. R. Bowden (2004). View-based location and tracking of body parts for visual interaction. In *Proc. of BMVC.*, Volume 1, pp. 849–858.

Micilotta, A. S., E. Jon, & O. R. Bowden (2005). Detection and tracking of humans by probabilistic body part assembly. In *Proc. of BMVC.*, Volume 1, pp. 429–438.

Moore, S. & R. Bowden (2009). The effects of pose on facial expression recognition. In *BMVC'09*, pp. 1–11.

Moore, S. & R. Bowden (2011). Local binary patterns for multi-view facial expression recognition. *Comp Vis and Image Understanding 115*(4), 541 – 558.

Moore, S., E. Jon Ong, & R. Bowden (2010). Facial expression recognition using spatiotemporal boosted discriminatory classifiers. In *Image Analysis and Recognition*, Volume 6111 of *LNCS*, pp. 405–414.

Okwechime, D., E.-J. Ong, A. Gilbert, & R. Bowden (2011a). Social interactive human video synthesis. In *Proc. of the 10th Asian Conf. on Comp Vis - Volume Part I*, Volume 6492 of *LNCS*, pp. 256–270.

Okwechime, D., E.-J. Ong, A. Gilbert, & R. R. Bowden (2011b). Visualisation and prediction of conversation interest through mined social signals. In *IEEE Int. Conf. on Face and Gesture Recognition and Workshops (FG 2011)*, pp. 951–956.

Ong, E.-J. & R. Bowden (2004). A boosted classifier tree for hand shape detection. In *Face and Gesture Recognition, 2004. Proc.. Sixth IEEE Int. Conf. on*, pp. 889–894.

Ong, E.-J. & R. Bowden (2008). Robust lip-tracking using rigid flocks of selected linear predictors. In *IEEE Int. Conf. on Face and Gesture Recognition*.

Ong, E.-J. & R. Bowden (2011a). Learning sequential patterns for lipreading. In *Proc. of the BMVC.*, pp. 55.1–55.10.

Ong, E.-J. & R. Bowden (2011b). Learning temporal signatures for lip reading. In *Comp Vis Workshops (ICCV Workshops), 2011 IEEE Int. Conf. on*, pp. 958–965.

Ong, E.-J. & R. Bowden (2011c). Robust facial feature tracking using shape-constrained multiresolution-selected linear predictors. *PAMI, IEEE Trans. on 33*(9), 1844–1859.

Ong, E.-J., H. Cooper, N. Pugeault, & R. Bowden (2012, june). Sign language recognition using sequential pattern trees. In *Comp Vis and Pat Rec (CVPR), 2012 IEEE Conf. on*, pp. 2200 –2207.

Ong, E.-J., Y. Lan, B. Theobald, R. Harvey, & R. Bowden (2009). Robust facial feature tracking using selected multi-resolution linear predictors. In *Comp Vis, 2009 IEEE 12th Int. Conf. on*, pp. 1483–1490.

Pugeault, N. & R. Bowden (2011). Spelling it out: Real-time asl fingerspelling recognition. In *Comp Vis Workshops (ICCV Workshops), 2011 IEEE Int. Conf. on*, pp. 1114–1119.

Sheerman-Chase, T., E.-J. Ong, & R. Bowden (2009). Online learning of robust facial feature trackers. In *Comp Vis Workshops (ICCV Workshops), 2009 IEEE 12th Int. Conf. on*, pp. 1386–1392.

Sheerman-Chase, T., E.-J. Ong, & R. Bowden (2011). Cultural factors in the regression of non-verbal communication perception. In *Comp Vis Workshops (ICCV Workshops), 2011 IEEE Int. Conf. on*, pp. 1242–1249.

Sheerman-Chase, T., E.-J. Ong, & R. Bowden (2013). Non-linear predictors for facial feature tracking across pose and expression. In *10th IEEE Int. Conf on Face and Gesture Recognition FG2013*.

Shotton, J., A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, & A. Blake (2011). Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, Washington, DC, USA, pp. 1297–1304. IEEE Computer Society.

Viola, P. & M. J. Jones (2004, May). Robust real-time face detection. *Int. J. Comput. Vision 57*(2), 137–154.