Tina Memo No. 2010-003
Submitted to BMVC 2010 (Rejected)

Presented at the Experimental Psychology Society meeting, Bangor July, 2013, see also memo 2009-003.

The Merits of 2D View vs. 3D Based Model Matching.

N. A. Thacker and S.Coupe.

Last updated
8 / 05 / 2010

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

The Merits of 2D View vs. 3D Based Model Matching

Abstract

A comparison of view-based and 3D model-based methods for localisation of man-made objects is made in the context of a working system. A projection validation approach is taken in order to confirm location hypotheses, which is based upon quantitative statistical models of feature detection and orientation. Results are provided which suggest that 3D data from stereo vision systems might be better employed for the prediction of novel views of objects, rather than as a generator of spatial representation suitable for geometric reasoning.

Introduction

Early work in object recognition by Marr [10] suggested that it would be infeasible to recognise 3D objects without using view-based invariant features. He argued that extraction of 3D surface data as the ideal approach to construction of an invariant shape representation. This approach is intended to eliminate the effects of view point and orientation which provide unwanted variation during recognition. However, the motivation for attempting a view- based approach was bolstered by Rosenfeld [1] and a number of studies which indicated that humans only store specific views of objects for the purposes of recognition [3, 5, 20, 19].

It has long been established that human vision is heavily dependent on the analysis of projected intensity edge contours. Our ability to effortlessly recognise objects from fragmentary line drawings supports the hypothesis that object recognition may be mediated by 2D projected contour-based pattern association, rather than surfaces. Furthermore, edge features have been shown to be very compact and powerful shape descriptors, offering a high degree of invariance to illumination and background clutter.

One approach to recognising 3D objects based upon extended edge contours is that adopted by Chen and Stockman (1997) [7]. This work involved the higher-level detection and recognition of freeform objects' defining curves, as extracted from a Canny [6] edge map. Carmichael introduced another histogram-type point operator for the recognition of image edges in 2004 [12]. The idea here was to use a circular reference window with around 40 uniformly sampled edge 'probes'. Again for a given oriented point of reference, each probe serves to sample the edge density in its Gaussian weighted local neighbourhood. A number of publications by Selinger and Nelson came to light by the end of the 20th century [13][14], detailing another curved edge reference feature-based scheme for the view-based learning and recognition of freeform 3D objects. Large-scale tests of their recognition system across a set of 24 complex objects through scale, clutter and occlusion indicated that recognition was robust across large complex scenes[13]. The authors suggested that these results were the best in the literature to their knowledge in 1997 [14]. However, the system is only able to operate if its sampled features are intact, although the simpler nature of their base descriptor makes the representation potentially more viable than that proposed by Chen and Stockman.

The standard approach for the recognition of visual data is the construction of a representation followed by statistical pattern recognition. Within this methodology, invariant and view-based approaches need not be entirely contradictory. It has been found possible to construct local view-based representations with a number of invariance characteristics without necessarily extracting complete curves or a 3D representation. 'Shape contexts' (2002) [16] have been proposed as a representation suitable for recognising objects by virtue of their extended edge distributions. They are are radially symmetric feature histograms that encode the frequency of edges passing through each bin surrounding a reference point. Although oriented relative to the main reference line upon which they are sampled, each histogram bin only details how many edge feature pixels pass though it without reference to their orientation. The shape context can be regarded as a simplified SIFT type detector [4]. One problem with these approaches is that having constructed such representations it is unclear to what extent shape information is actually encoded. One way to understand this relates to the concept of representational completeness, i.e. Can the original shape information be reconstructed from the feature representations used in recognition? The approach adopted here is the Pairwise Geometric Histogram [11], which aside from completeness, has a number of characteristics which make it robust to the effects generally found in real images, including curve fragmentation, occlusion and clutter.

In this work, taken from the lead authors PhD thesis, we take two approaches to predicting the locations of objects in images. Predictions are then used as the starting point for a validation process which quantitatively assesses the amount of edge data present in the scene corresponding to the best object model projection.

Our projected view validation strategy is based upon the ideas presented in [17], which allows us to construct independent quantitative hypothesis tests for the detection of an edge at a specific location and orientation in an image. This method varies from previously published methods in that the statistical distributions are based upon the characteristics of noisy image formation, rather than positional errors of the projected model [9]. The edge hypothesis is based upon a bootstrap approach for computing the probability that random configurations of data in the region of the hypothesis would have resulted in defining the location under consideration as an edge. The orientation hypothesis is a more conventional approach based upon propagating errors from the observed grey levels through to those on estimated orientation. A validation score for extended curves (or the entire object) is then based upon the fraction of edge data which satisfies a given hypothesis threshold. The associated theory also provides a robust Likelihood formulation for object location, which is used to refine the estimate provided by the view-based and stereo-based match results. The algorithm incorporates a lateral feature position adjustment to accommodate systematic effects of image formation [17]. Indeed, as the hypothesis test is defined for distributions generated at the correct location, we cannot compute a useful validation score without this step. In essence, the validation step provides a well behaved estimate of the proportion of projected model data which accords with the presence of a suitable edge.

The 3D localisation system is based upon an optimised version [2] of work presented in [8]. This framework comprises a feature-based stereo matching algorithm, curve fitting and view hypothesis generation using a 3D model matcher. As the intention is to identify and match extended features which are consistently detected and located in 3D, this excludes the use of occluding boundaries (which vary in position as a function of viewpoint and do not provide meaningful stereo measurements). Over recent years, the system has been upgraded to increase the reliability of stereo matching and 3D curve fitting, to the point we believe that both now provide data which is as good as can be extracted, matched and fitted for the scene for 99% of true step edge features. The 3D matcher uses a heuristic (non-statistical) geometrical clique matching strategy to generate location hypotheses which are then used as the starting point for iterative alignment and validation (as described above).

The view-based localisation system is based upon use of histograms of geometric co-occurrence (Pairwise Geometric Histograms). A shape-based matching system is constructed for the view sphere using a dual linear model of deformation of these histograms. The statistical basis for this approach and the way in which the resulting piecewise linearisation accords with our expectations of the topology of data from 3D projection are described in another paper submitted to this conference.

Methods

3D Model Matching

The 3D-based model matching system operates by forming an edge-based depth map of a scene from a pair of calibrated stereo images. An implementation of the Canny edge detector [6] is used to extract images' significant edge features before a 'stretch correlation' algorithm is applied for stereo reconstruction [15, 18]. Extended edge structures are recursively fitted to geometrical primitives such as lines and ellipses in 2D and then projected into 3D using the correspondences identified from stereo matching.

Object model to scene matching is performed by searching for sets of three-dimensionally distributed edge features with the expected geometric configurations. Tables are created for both model and scene geometries, detailing the relative 3D distances and orientations of each feature within each set. Because representative object models may potentially incorporate hundreds of features, matching is initially based on reduced sets of features (e.g. 10 at a time), so as to avoid combinatorial search problems. A further reduced set of the most important features is specified as a focus feature set, and used to seed the matching process.

Model matching proceeds by finding any potential matches for each specified focus feature with the scene geometry, thus forming a potential match seed list. For each entry on the seed list, an exhaustive search is made between the remaining features in the (reduced-) model and (full-) scene geometry sets to find

any other coinciding features. This is performed by checking whether each potential feature pairing is of a compatible type and if so analysing whether the pairwise 3D geometrical relationships accord with those expected (using quite loose, non-statistical, tests). The initial match lists may contain many matches for each focus feature including combinatorial combinations. A subsequent stage of processing then extracts any unique combinations of features. For each such feature set with a prescribed minimum number of features (e.g. 3), a further search is conducted over the remaining features in the full wireframe model, with each match list being extended accordingly. Each matching pair of features is weighted by length, with any cumulative match lists being ranked by the sum of such weights. Candidate match hypotheses are ordered in terms of this weighting and used to compute approximate location parameters for the model, using simple least-squares. These estimates are then suitable for initialisation of an iterative localisation using more realistic statistical assumptions.

View Based Matching

For a given wireframe object, a specified number of the longest linear edge features (typically 12) are selected for a fixed number of (equally spaced) vertices from the view sphere. Geometric histograms are constructed with dimensions of 20x64 bins, corresponding to ranges of $-50 \rightarrow +50$ pixels, and $0 \rightarrow 4\pi$ radians in perpendicular distance and relative orientation. Care is taken to make entries so that they will change smoothly as a function of view direction, with uncertainty encoded as blurring at the level of 1 bin on each axis. A piecewise triangulated linear model is then learned for those features jointly visible across the view patch (to remove representational discontinuities).

The weighting process of histogram formation leads to their statistical interpretation as a Poisson counting process. This justifies the use of the Bhattacharyya measure for matching, provided that correlated behaviours in histogram formation are modelled. This is done via approximation, based upon a minimum reconstruction match score (eg: 0.9), using a strategy of iterated refinement. So that, if the linear model does not approximate the match score at the centre of the triangular patch or along the bounding edges, it is subdivided generating 4 new triangles (Figure 1). The results of training are pre-computed and stored to data files for use during matching.

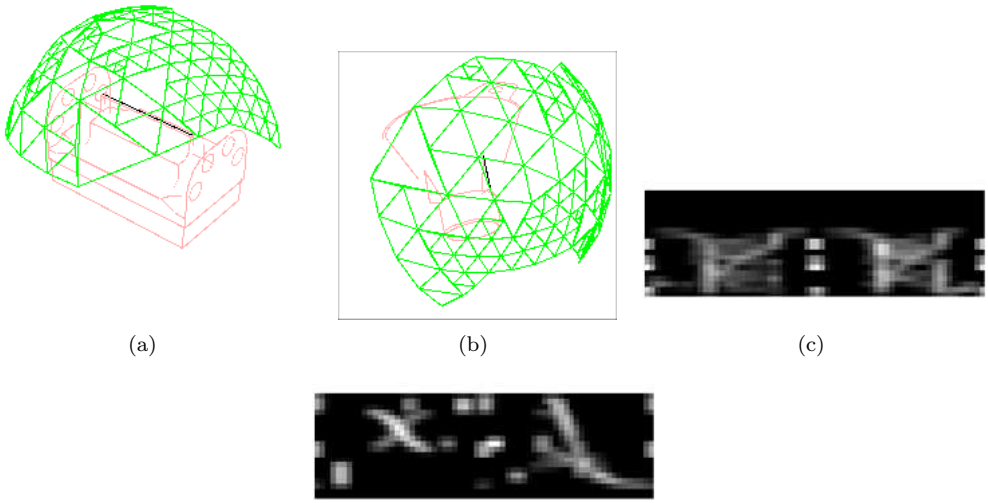
Object location hypotheses are generated by matching a linear model of histograms from scaled lines ($\pm$ 10 %) in the scene to the view-based models and accumulating evidence for most likely generators of the observed data as a function of view point and scale. By basing the matching structure upon geometric histograms (which encode all expected relationships between the features of an object in one matchable structure), and in contrast to the 3D approach described above, combinatorial search is avoided at the expense of an increase in base level (linear) computation per feature. As all data is represented as linearised approximations, curve categorisation errors are also avoided. The matching of triangles is done taking into account the linear boundaries of triangular patches. Matches from the full model that are found to lie outside the patch are recomputed, constrained along the line between the vertices. Matching the full set of view point triangles is avoided by treating the matching process as an optimisation. Starting at the central view direction, any triangle connected to the best match result so far is searched recursively until no improvement in match score is obtained.

Once a minimum amount of evidence (e.g. 3 features per view and scale) has been accumulated, the features consistent with this hypothesis are used to initialise iterative alignment and validation (as described above).

Evaluation

The view and 3D-based methods of location hypothesis generation can therefore be interchanged within the overall system. Comparison of the performance for these approaches is based upon the number of valid alignments generated. The aim is to illustrate key aspects of data behaviour which limit the success of these approaches.

To evaluate the performance of the proposed view-based model matching system, a selection of 15 3D objects covering a diverse range of shapes and materials has been obtained (figure 2). A 3D wireframe model has been constructed for each object. View-based feature visibility is approximated with reference to manually composed visibility files, which are currently uniformly sampled with 42 views around the view-sphere. Using these constructs as surrogates for the corresponding tangible objects, each object's



(a) (b) (c)

(d)

Figure 1: Recursive triangulation (green) of geometric histogram variation as a function of view direction for the wiper (a) and funnel (b), shown together with typical geometric histograms (perpendicular distance vs relative orientation) for the mean view of the selected line (black).

range of PGH appearance has been learned in accordance with the method outlined above.

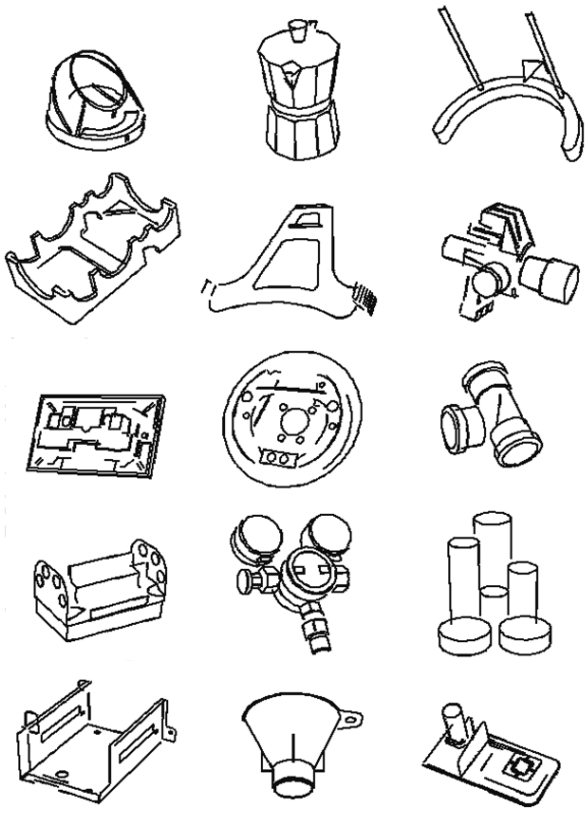


Figure 2: Wireframe models of; grill, pot, stand, guide, aframe, pump, plug, brake, pipe, wiper, valve, tidy, tray, funnel and widget (top left to bottom right).

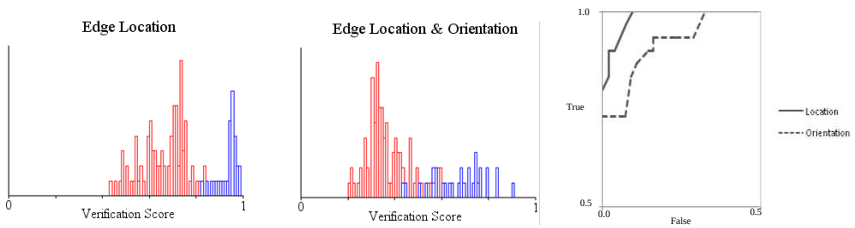
[17] outlined a quantitative statistical framework for the verification of projected edge feature points. Now that an associated model matching system is available, it is possible to evaluate the utility of the proposed approach to verification in terms of discrimination between valid and invalid instances of model matches. For these experiments, 2 random digital images (1152/864 down-sampled 50%) of each of the 15 objects in the test dataset have been accurately model matched with valid model views and the proportion of sampled edge feature points passing a specified hypothesis test have been sampled. For each of the 30 test images, 3 other random objects from the set of 15 have been optimally matched to the image edge data with the same criteria. The experiments have been repeated with and without the additional constraint of edge orientation to allow any advantages to be observed in terms of valid/ invalid class separability.

Each of the 15 test objects has also been imaged in stereo (verge angle approximately 20 degrees) against a plain cloth background from 14 view-points approximating the 8 corners and 6 faces of an encompassing virtual cube (210 stereo pairs in all). The cameras were then calibrated using a 2D tile by minimising the error on projected vertices from a grid of squares. Resulting residuals were accurate to a fraction of a pixel. The view-based model matching system has then been applied to the right image. The longest 12 linear edge features (relative to a central viewpoint) from each of 42 views are matched to the longest 24 linear image edge features. A larger set of image edge features is sampled to account for potential interference from background artefacts (e.g. shadows) or extended line fragmentation. These numbers represent a performance compromise and were selected in order to generate execution times of no more than a few minutes for each object.

For both systems, performance has been quantified in terms of general pose with an indication of the quality of model alignment. Alignment quality has been manually graded as very good (VG), good (G) or fair (F) with any poorly aligned model matches being classified as match errors. As the 3D-based matching strategy also resulted in quite variable execution times we have also documented them here.

Finally, we have run the view based localisation technique on images of multiple occluded objects from our data base, with the minor modification to the baseline algorithm of processing additional scene lines until a view hypothesis is generated.

Results



(a) (b) (c)

Figure 3: Total validation score for valid matches (blue) and invalid matches (red), defined as the percentage of projected edge data above the 1% hypothesis limit. Results are given for edge only (a) and edge plus orientation (b), together with associated ROC curves.

The validation process was tested for a range of hypothesis limits (i.e. 100 - % of data consistent with the model). Results for the use of the best view validation score corresponding to a 1% hypothesis are shown in figure 3. Tables 1 and 2 show the performance figures for the view-based and stereo localisation algorithms. The data demonstrates that the view-based system works very well on most objects (even for cluttered scenes when matching to sufficient numbers of scene lines, see figure 4), but the 3D vision system has systematic difficulties with some classes of object, particularly those for which the surface properties generate difficulties for curve extraction and stereo matching.

Object (14 views)	Correct Solution	Symmetric Solution	Alignment Quality			Match Error
			VG	G	F	
Aframe	13	0	8	2	3	1
Brake	5	3	4	3	1	6
Funnel	11	3	14	0	0	0
Grill	14	0	14	0	0	0
Guide	14	0	11	1	2	0
Pipe	9	4	7	3	3	1
Plug	9	2	6	3	2	3
Pot	9	5	13	1	0	0
Pump	14	0	13	1	0	0
Stand	14	0	14	0	0	0
Tidy	12	0	10	2	0	2
Tray	13	1	12	2	0	0
Valve	6	1	2	3	2	7
Widget	14	0	13	1	0	0
Wiper	14	0	14	0	0	0
Mean	81.5%	9.0%	73.8%	10.5%	6.2%	9.5%

Table 1: View-based localisation performance, evaluated on 14 views of 15 objects.

Object (14 views)	Correct Solution	Symmetric Solution	Alignment Quality			Match Error	Extended Params. Required	>5mins
			VG	G	F			
Aframe	10	2	12	0	0	2	1	
Brake	1	1	1	1	0	12	1	3
Funnel	8	3	9	2	0	3	2	
Grill	4	1	2	1	1	9	1	2
Guide	6	0	6	0	0	8	1	
Pipe	1	0	0	0	1	13	0	
Plug	9	1	9	0	0	4	1	8
Pot	0	3	2	1	0	11	1	4
Pump	2	0	2	0	0	12	2	
Stand	7	0	6	1	0	7	0	
Tidy	4	0	2	2	0	10	0	
Tray	10	0	9	0	1	4	2	4
Valve	1	0	0	1	0	13	0	5
Widget	7	1	7	1	0	6	1	
Wiper	8	1	9	0	0	5	1	1
Mean	37.1%	6.2%	36.2%	4.8%	1.4%	56.7%	6.7%	12.9%

Table 2: Localisation performance for the stereo-based 3D system from 210 calibrated stereo pairs. The format is similar to Table 1, except extra table entries are included to indicate occasions where the match parameters (focus and cliche feature group sizes) required modification, and the effects of combinatorial search on execution times.

Conclusions

The results indicate that using the orientation hypothesis does not improve the reliability of location validation beyond that already obtained from edges alone. We believe that this is because the bootstrap approach to construction of an edge detection hypothesis is more robust than use of the orientation distribution computed via error propagation. We believe this to be due to the factors associated with image formation (such as interactions between illumination and local surface properties). Moreover, as the modelled features are in any case extended, a separate measurement of orientation may be redundant.

View-based alignment was found to be quite reliable, completing execution in a time commensurate with the linear complexity of the object model (up to a few minutes). All failures of the localisation could be attributed to poor views in which insufficient structure was detected due to object orientation or

surface illumination. Tables of localisation performance demonstrate that 3D model-based alignment is far less reliable and with execution times being polynomially dependant upon object complexity and often exceeding 5 minutes.

A 3D approach can be made to work in restricted cases (for example a limited set of possible views). Though both approaches were affected by the non-detection of features and object symmetry, inspection of the failing data sets from these experiments suggest that the stereo-based 3D matching was compromised in three fundamental ways. First; the extraction/labelling of curves, though working as well as we can expect given the local image data, was often inefficient or fragmented, leading to lost or mis-identified features. Second, the stereo process required any feature to be jointly visible and metrically useful in both images, any structure present in only one image (or oriented parallel to an epi-polar) will be lost. Thirdly, the inability to incorporate occluding boundaries, or correctly extract them using stereo, resulted in less real data and more additional clutter. We believe that in our software these processes operate at a level which is now far more significant than the errors introduced by curve fitting or stereo matching.

Three dimensional model matching may sound like a good way to solve the scene interpretation problem, but the advantage gained by using an invariant description is quickly lost by the inability to represent a large fraction of visible features and the inability to address appropriately the statistical errors on reconstructed data (which inevitably vary as a function of viewpoint). The use of a statistically based matching strategy (such as the PGH matched with a Bhattacharyya score) gives better reliability during the identification of candidate matches than the essentially heuristic non-statistical approaches based upon 3D models in the absence of an adequate description of measurement error.

As things stand, we can say that the original purpose for the stereo algorithms has been undermined by these results. However, stereo now offers a reasonable starting point for the automatic generation of the view-based wireframe models upon which the view-based approach can be trained. Stereo data may thus have more utility for the prediction of changes in appearance over view point, than as a Marr-like 3D sketch.

References

[1] A. Rosenfeld. Recognizing Unexpected Objects: a Proposed Approach. *International Journal of Pattern Recognition and Artificial Intelligence*, 1(1):71–84, 1987.

[2] A.Lacey, N.A.Thacker, P.Courtney and S.Pollard. TINA 2001: The Closed Loop 3D Model Matcher. *Proc. BMVC*, pages 203–212, 2001.

[3] H. Bulthoff and S. Edelman. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Nat. Aca of Sci.*, 89:60–64, 1992.

[4] D. Lowe. Object Recognition from Local Scale-Invariant Features. *Proc. IEEE International Conference on Computer Vision, Greece*, 2:1150–1157, 1999.

[5] S. Edelman and H. Bulthoff. Orientation dependence in the recognition of familiarand novel views of three-dimensional objects. *Proc. Nat. Ac. of Sci USA*, 32:2385–2400, 1992.

[6] J. Canny. A Computational Approach to Edge Detection. *IEEE Trans. PAMI*, 8:679–714, 1986.

[7] J. Chen and G. Stockman. 3D Free-form Object Recognition using Indexing by Contour Features. *Computer Vision and Image Understanding*, 71(3):334–355, 1998.

[8] J.Porrill, S.B.Pollard, T.Pridmore, J.Bowen, J.E.W.Mayhew and J.P.Frisby. Tina: A 3D vision system for pick and place. *Proc. Third Alvey Vis. Conf.*, pages 65–72, 1987.

[9] M. Boshra, H. Zhang. Accomodating Uncertainty in Pixel-Based Verification of 3-D Object Hypotheses. *Pat. Rec. Let*, 20(7):689–698, 1999.

[10] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Henry Holt and Company, 1982.

[11] N. A., Thacker, P. A. Riocreux and R. B. Yates. Assessing the Completeness Properties of Pairwise Geometric Histograms. *Image and Vision Comp.*, 13(5):423–429, 1995.

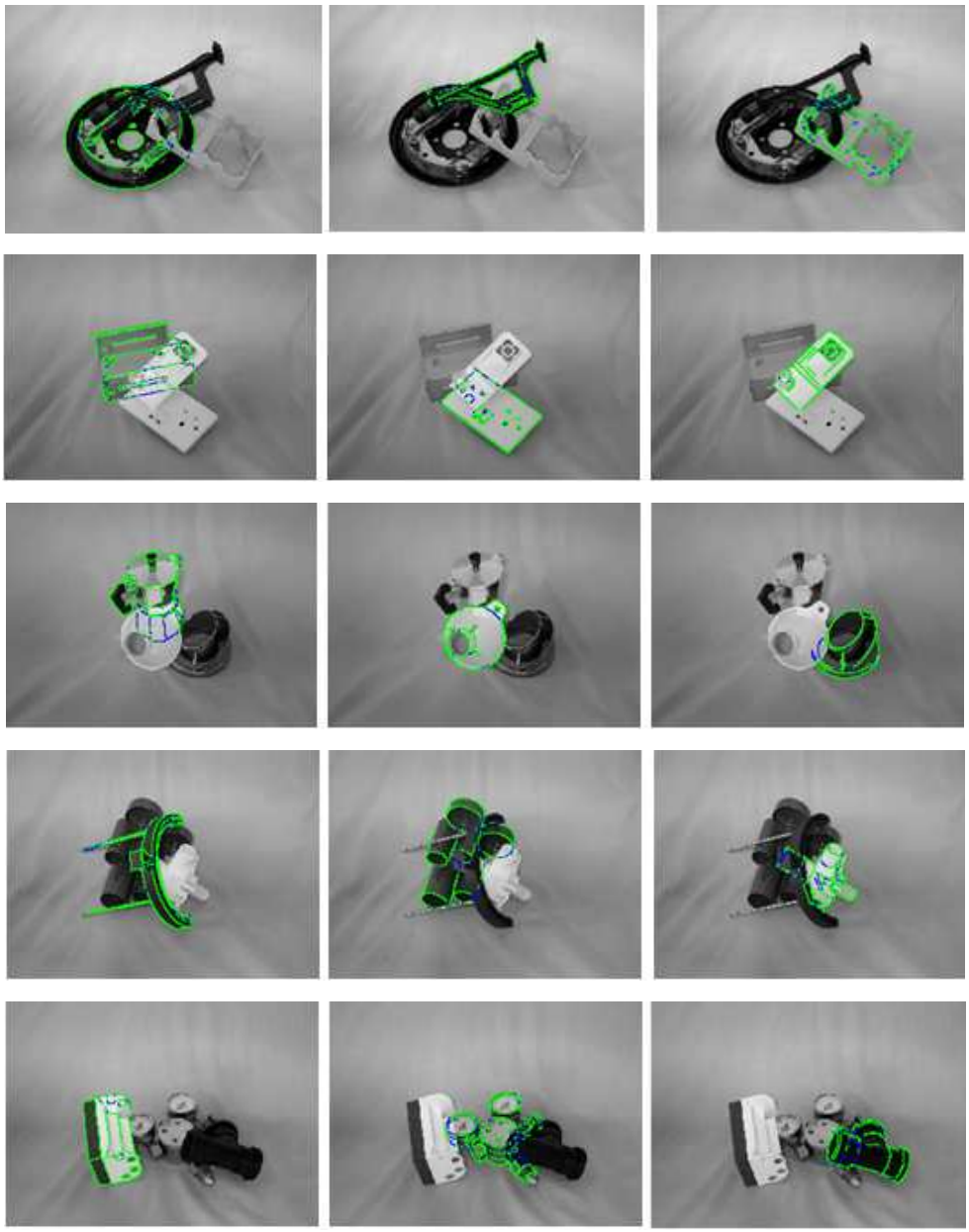


Figure 4: Localisation performance for the view-based system in occluded scenes showing verified (green) and non-verified (blue) feature points.

[12] O. Carmichael and M. Hebert. Object Recognition by a Cascade of Edge Probes. *Proc. BMVC*, pages 102–122, 2002.

[13] R. Nelson and A. Selinger. Large-Scale Tests of a Keyed, Appearance-Based 3-D Object Recognition System. *Vision Research: Special Issue on Computational Vision*, 38:15, 1998.

[14] R. Nelson and A. Selinger. A Perceptual Grouping Hierarchy for Appearance-Based 3D Object Recognition. *Computer Vision and Image Understanding*, 76(1):15, 1999.

[15] R.Lane, N.A.Thacker and N.L.Seed. Stretch Correlation as a Real-Time Alternative to Feature Based Stereo Matching Algorithms. *Image and Vision Comp.*, 12(4):203–212, 1994.

[16] S. Belongie, J. Malik, and J. Puzicha. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Trans. PAMI*, 24(4):509–522, 2006.

[17] S.Coupe and N.A.Thacker. Quantitative verification of projected views using a power law model of feature detection. *Proc. CRV, Canada.*, 2008.

[18] S.Crossley, A.J.Lacey, N.A.Thacker and N.L.Seed. Robust Stereo via Temporal Consistency. *Proc. BMVC*, pages 659–669, 1997.

[19] M. Tarr. Rotating objects to recognize them: a case study of the role of viewpoint dependency in the recognition of three-dimensional objects. *Psy. Bull. and Rev.*, 1(2):52–82, 1995.

[20] M. Tarr and S. Pinker. Mental rotation and orientation dependence in shape recognition. *Cog. Psy.*, 28(21):233–282, 1989.