A Saliency-Based Framework for 2D-3D Registration

Mark Brown, Jean-Yves Guillemaut and David Windridge
Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, United Kingdom
{*m.r.brown, j.guillemaut, d.windridge*}*@surrey.ac.uk*

Keywords: pose estimation, registration, saliency

Abstract: Here we propose a saliency-based filtering approach to the problem of registering an untextured 3D object to a single monocular image. The principle of saliency can be applied to a range of modalities and domains to find intrinsically descriptive entities from amongst detected entities, making it a rigorous approach to multi-modal registration. We build on the Kadir-Brady saliency framework due to its principled information-theoretic approach which enables us to naturally extend it to the 3D domain. The salient points from each domain are initially aligned using the SoftPosit algorithm. This is subsequently refined by aligning the silhouette with contours extracted from the image. Whereas other point based registration algorithms focus on corners or straight lines, our saliency-based approach is more general as it is more widely applicable e.g. to curved surfaces where a corner detector would fail. We compare our salient point detector to the Harris corner and SIFT keypoint detectors and show it generally achieves superior registration accuracy.

1 INTRODUCTION

The increase in available 3D data over the last decade has naturally led to the problem of its registration with 2D images. It has applications in object recognition, robotics and medical imaging, where in particular real-time solutions are required [van de Kraats et al., 2004, Gendrin et al., 2011].

While there is a significant amount of 3D data that contains texture information, for example acquired from Structure from Motion or multi-view reconstruction techniques, here we consider the problem where the data is untextured; often obtained from LiDAR or other forms of laser scanner. This constraint makes it significantly more challenging, if the 3D data were textured, point correspondences could be initialised by, for example, using the SIFT descriptor from Viewpoint Invariant Patches [Wu et al., 2008]. However, for untextured 3D data these descriptors are inapplicable since there is no longer a common modality to be described.

The 2D / 3D registration problem can be seen as an instantiation of the more general multi-modal registration problem, an area that has received significant attention in computer vision [van de Kraats et al., 2004, Gendrin et al., 2011, Mastin et al., 2009, Egnal, 2000]. While there exist a number of methods that can perform this with a good initial alignment, e.g. 2D / 3D registration in medical imaging using fiducial markers [van de Kraats et al., 2004] or cross-spectral stereo matching [Egnal, 2000], the less constrained case without these priors remains largely unsolved. This is because the methods relying on a good initial alignment often use computationally expensive measures such as Mutual Information [Egnal, 2000] and so are infeasible when considering all possible alignments.

The main contribution of this paper is its advocation of saliency-based filtering for the 2D / 3D registration problem. Saliency is a broad term that refers to the idea that certain parts of an image are more informative or distinctive than other areas and these parts represent the intrinsic and underlying aspects of the object. As such, it is not domain or modality specific, making it a very general approach to multi-modal registration. Further, it is seen as an approximation to the Human Visual System (HVS) which is capable of solving a variety of vision and registration tasks very efficiently. The structure of saliency detectors can be varied since the definition is broadly specified; the aim is often to extract features that have a high repeatability with respect to keypoints people have labeled as salient [Judd et al., 2012]. As such, the two detectors the saliency detector is compared to - the Harris Corner detector and SIFT - could arguably both

be considered saliency detectors in certain contexts.

While there are a range of saliency detectors to choose from [Itti, 2000, Kadir and Brady, 2001, Judd et al., 2012, Lee et al., 2005], we use the Kadir-Brady detector due to its principled information-theoretic approach. We generalise it over an arbitrary number of dimensions and implement a curvature based extension to the 3D domain. It extracts salient points by maximising the Shannon entropy in scale-space; a measure that has been used in 2D / 3D and cross-spectral registration before in the form of Mutual Information (MI) [Mastin et al., 2009, Egnal, 2000] however MI is too costly in our case as we do not assume a good initial alignment.

After extracting salient points from both domains, we seek to align them. However, as descriptors are typically inapplicable in multi-modal data no correspondences can be initialised; hence the 'Simultaneous Pose and Correspondence' (SPC) problem has to be solved. For this we use the SoftPosit algorithm [David et al., 2002] to initially align the data and the registration is refined by aligning the edges of the silhouette with edges from the image. Whereas other methods e.g. [Mastin et al., 2009] put a strong prior on the initial alignment, the SoftPosit point based method is able to consider a large variety of alignments (all rotations and limited translations) within a reasonable amount of time (5 minutes on a standard single core CPU).

The structure of this paper is as follows: In Section 2, we give an outline of related work in 2D / 3D registration. In Section 3 we present our methodology; in Section 4 our experiments and conclusions are presented in Section 5.

2 Related Work

2D / 3D registration has received a significant amount of attention, particularly in the medical domain, where multi-modal (e.g. CT, MRI, 3D Rotational X-ray [van de Kraats et al., 2004]) registration is of great importance. [van de Kraats et al., 2004] evaluate a number of algorithms for this and classify them as *intensity-based*, *gradient-based*, *feature-based* or *hybrid-based*. Algorithms outside of the medical imaging domain are usually *feature-*, *intensity-,* or *learning-based*. Registration in medical imaging typically involves using fiducial markers for an initial alignment and so the problem becomes one of refinement.

Intensity based methods typically project the 3D data and use correlation measures such as mutual information or cross-correlation to refine the registration [Kotsas and Dodd, 2011]. These techniques have been used in cross-spectral stereo matching [Egnal, 2000] where the offset that maximises the mutual information or correlation is determined to be the correct offset. It has also been used by [Mastin et al., 2009] where the authors match a LiDAR scan to an image using Mutual Information. However, the authors acknowledge the sensitivity to initialisation, saying that they do not initialise the yaw or pitch to more than 0.5 degrees from the ground truth in their experiments. It would be computationally infeasible when sampling without these priors.

Feature based methods extract features such as points or lines, and match these up, often without assuming any knowledge about their correspondences (hence solving for the pose *and* the correspondences). A number of papers have attempted to solve the Simultaneous Pose and Correspondence (SPC) problem from these points, a problem also addressed in 3D-3D registration. With a good enough initial alignment, matching can be achieved through Iterative Closest Point (ICP). This can be made more robust through Expectation-Maximisation ICP [Granger and Pennec, 2002] or Levenberg-Marquardt ICP [Fitzgibbon, 2001] who further describes the use of different kernels to deal with outliers and shows their proposed method increasing the basin of convergence over other variants of ICP. A well known solution to SPC, and the one used here, is the SoftPosit algorithm [David et al., 2002]. Unlike ICP, this allows weighted correspondences between points rather than the binary assignment in ICP, allowing it to better avoid local maxima during convergence. It will be discussed further in Section 3.3.

More recently [Moreno-Noguer et al., 2008] solved the SPC problem by modeling initialisations as a Gaussian Mixture Model and using each component to initialise a Kalman filter. They also introduced priors on the camera pose; for example the camera is always above the ground and pointing towards the object. It performs just as well as SoftPosit in a similar amount of time except in large amounts of clutter, where SoftPosit is outperformed by it. Also, [Enqvist et al., 2009] solved the SPC problem by determining which pairs of correspondences are infeasible and sought to maximise the number of correspondences that are feasible. This is done using an heuristic branch and bound technique and the authors achieve results comparable to state of the art.

Machine learning algorithms are evident in the literature for the 2D / 3D registration problem, specifically in face recognition. In general, the main drawback with this approach is the time spent in labeling and learning the model. [Yang et al., 2008]

use Canonical Correlation Analysis (CCA) for face recognition to learn a correlation between image features and 3D features of face data. In this application, an image and a 3D mesh of a face are compared: if there is sufficient correlation then it is deemed a match.

2.1 Saliency-Based Methods

We use Kadir-Brady saliency [Kadir and Brady, 2001, Shao et al., 2007] due to its applicability in different modalities resulting from a principled information-theoretic approach. Other forms of saliency are not so easy to extend to 3D e.g. Itti-Koch saliency [Itti, 2000] relies on colour and centre-surround operations that aim to replicate how the human visual system works. [Lee et al., 2005] determine the saliency of a 3D mesh through weighting curvatures in scale space to return a saliency value for each point; this is not as general as Kadir-Brady which only requires a histogram.

Saliency has already been applied in the 2D / 3D registration problem in medical imaging by [Chung et al., 2004]. Here, the authors learn a saliency map through eye-tracking of volunteers. By focusing on registering only on the region of the image that is salient, the authors increase the accuracy of the results whilst reducing the computation time. This is different from what is proposed here because [Chung et al., 2004] uses a ground truth saliency map constructed through using eye tracking software; no salient algorithm was used. Instead, we propose a framework for 2D / 3D registration that can be applied to the general multi-modal registration problem.

3 Methodology

We wish to register a 3D model to an image. To do so, N_I salient image points $\{\mathbf{x}_i\}_{i=1}^{N_I}$ and N_W salient mesh points $\{\mathbf{X}_j\}_{j=1}^{N_W}$ are extracted in the **first stage** of the methodology. This is achieved using the generalised Kadir-Brady saliency detector in each modality (see Sections 3.1 and 3.2).

The **second stage** of the methodology aims to determine the transformation that matches these points. This requires the SPC problem to be solved since no putative matches may be found, for which SoftPosit [David et al., 2002] is used.

However, this is only possible if the two sets of salient points have a sufficiently high repeatability. This may not be the case: in particular the image points may exhibit *projective saliency*, whereby they

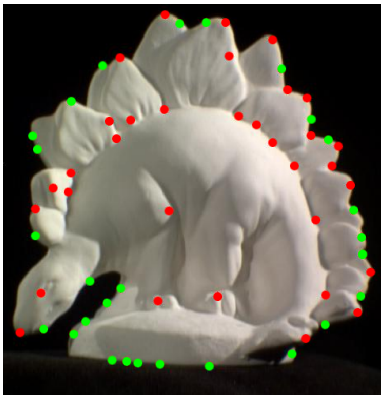

Figure 1: An illustration of different types of saliency. The red points are salient due to the intrinsic properties of the model whereas the green points exhibit *projective saliency*.

are salient because they lie on the edge of the projected model, due to the viewing angle (i.e. transformation between modalities) rather than the model's intrinsic properties. An example of this is given in Figure 1. Projective saliency is accounted for by extracting points on the edge of the silhouette (S) of the mesh and measuring their alignment with $\{\mathbf{x}_i\}_{i=1}^{N_I}$ - a detailed explanation is given in Section 3.3.

The **final stage** of the methodology takes the best scoring 1% of transformations from the previous stage and refines them by further matching edges from 2D (extracted using the Canny edge detector) with the boundary of the 3D silhouette - a detailed explanation of this is given in Section 3.4. A diagram showing the pipeline of the methodology is given in Figure 2.

3.1 The Generalised Kadir-Brady Saliency Detector

The Kadir-Brady detector is here abstracted to $\mathbb{R}^n$. This is constructed using a set of points $\{\mathbf{x}_i \in \mathbb{R}^n\}$, a set of scales $\{s_1, \ldots, s_K\}$ and a histogram with values $\{v_{1,\mathbf{x},s_k}, \ldots, v_{R,\mathbf{x},s_k}\}$ associated with a point $\mathbf{x}$ over a given scale s_k. The Kadir-Brady detector uses this histogram about each point for each scale and determines where the histogram's entropy is peaked in scale space. The higher this peak is above its neighbour's scales, the higher the saliency of that point is defined to be.

More formally, for a point $\mathbf{x}$, a scale s_k, and a histogram with values $\{v_{1,\mathbf{x},s_k}, \ldots, v_{R,\mathbf{x},s_k}\}$, the entropy of $\mathbf{x}$ is defined as:

$$H_{\mathbf{x},s_k} = -\sum_{i=1}^{R} P(v_{i,\mathbf{x},s_k}) \log_2(P(v_{i,\mathbf{x},s_k})) \qquad (1)$$

where $P(v_{i,\mathbf{x},s_k})$ is the probability of the histogram

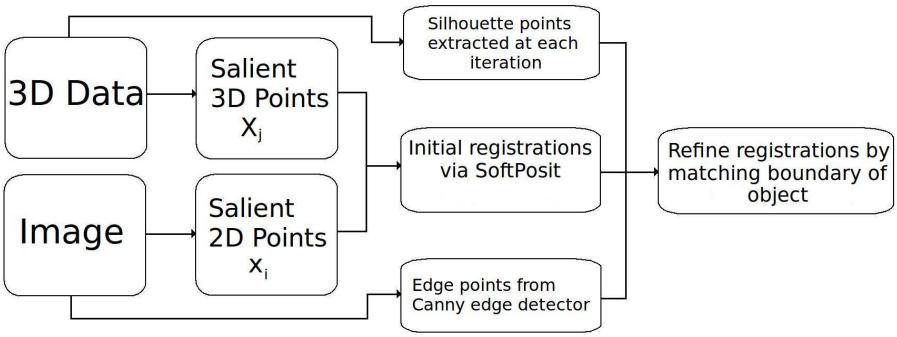

Figure 2: Pipeline of methodology for 2D / 3D registration.

taking the value v_i at scale s_k from $\mathbf{x}$. This is taken from a frequentist approach, i.e. $P(v_{i,\mathbf{x},s_k}) = v_{i,\mathbf{x},s_k} / \sum_{j=1}^{R} v_{j,\mathbf{x},s_k}$. [Shao et al., 2007] make this step more robust by weighting points by twice as much if they are within s_{k-1} of $\mathbf{x}$. This is used here as the authors demonstrate an increase in the repeatability by 10% compared to no weighting.

After computing the entropy of each point at each scale (as above) only the scale-space points for which the entropy is peaked above its neighbouring two scales are retained. This is then weighted by how dissimilar the PDF is from these two scales:

$$W_{\mathbf{x},s_k} = \frac{1}{2}\Big(C_1 \sum_{i=1}^{R} \left|P(v_{i,\mathbf{x},s_{k+1}}) - P(v_{i,\mathbf{x},s_k})\right| + C_2 \sum_{i=1}^{R} \left|P(v_{i,\mathbf{x},s_k}) - P(v_{i,\mathbf{x},s_{k-1}})\right|\Big) \quad (2)$$

where C_1 and C_2 are constants that make the comparison scale invariant:

$$C_1 = \frac{N_{s_{k+1}}}{N_{s_{k+1}} - N_{s_k}} \qquad C_2 = \frac{N_{s_k}}{N_{s_k} - N_{s_{k-1}}} \quad (3)$$

and where N_{s_k} is the number of points within s_k of $\mathbf{x}$, i.e. $N_{s_k} = \sum_{i=1}^{R} v_{i,\mathbf{x},s_k}$. The final saliency measure $Y_{\mathbf{x},s_k}$ is then the product of the two measures $H_{\mathbf{x},s_k}$ and $W_{\mathbf{x},s_k}$.

These salient points are subsequently clustered into circular salient regions using an heuristic clustering algorithm according to [Kadir and Brady, 2001]. Its purpose is to make it more robust to noise as this would otherwise act as a randomiser and increase the entropy. The centre of each circular region is defined to be the salient point returned by the detector, and has an associated scale and saliency value. An example of points extracted from the generalised Kadir-Brady saliency detector is given in Figure 3.

The original Kadir-Brady detector for images ($\mathbb{R}^2$) constructs a 256-bin histogram of pixel intensities about each pixel taken from a circular neighbourhood of radius s_k. The radius ranges from three pixels to 21 in intervals of three.

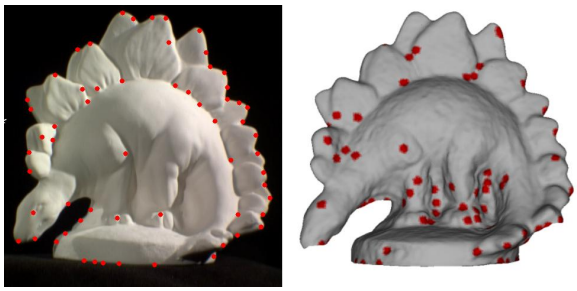

Figure 3: Detected points of the *Middlebury dinosaur* using the standard 2D Kadir-Brady detector (left) and its extension to 3D mesh-based data (right).

3.2 The 3D Kadir-Brady Saliency Detector

While the scale-space concept is the same in 3D, the descriptor histogram to use is less obvious. After some experiments, we adopt a curvature-based detector, the same as in [Lee et al., 2005]. For this the mean curvature (the average of the two principal curvatures) is used, where the principal curvatures at $\mathbf{x}$ are defined to be the eigenvalues of the shape operator at $\mathbf{x}$, and are calculated using the algorithm presented by [Taubin, 1995]. Since this algorithm can produce abnormally large curvatures for some points due to floating point errors, in our algorithm those with the top 0.5% of all curvature values have their curvature value capped. Denoting the maximum curvature by M, a 30-bin histogram around point $\mathbf{x}$ with scale s_k is constructed as follows: For each point $\mathbf{x}_j$ that is within s_k of $\mathbf{x}$ denote its curvature as $\mathbf{x}_j^c$. Then insert this into bin number $\lfloor \frac{30\mathbf{x}_j^c}{M} \rfloor$, i.e. construct bins of uniform width ranging from curvatures 0 to M. Thus the value of each bin in the histogram can be expressed

as:

$$v_{i,\mathbf{x},s_k} = \sum_{\mathbf{x}_j:|\mathbf{x}-\mathbf{x}_j|<s_k} \delta(\lfloor \frac{30\mathbf{x}_j^c}{M} \rfloor, i) + \sum_{\mathbf{x}_j:|\mathbf{x}-\mathbf{x}_j|<s_{k-1}} \delta(\lfloor \frac{30\mathbf{x}_j^c}{M} \rfloor, i) \quad (4)$$

where $\lfloor \mathbf{x} \rfloor$ denotes the greatest integer smaller than $\mathbf{x}$ and δ the Kronecker delta function.

Seven scales are used as in the 2D case. Here the lowest scale is 0.3% of the length of the diagonal of the bounding box of the model, and this increases in 0.3% intervals up to 2.1%.

3.3 The SoftPosit Algorithm

The SoftPosit algorithm [David et al., 2002] attempts to solve the SPC Problem: for image points $\{\mathbf{x}_i\}_{i=1}^{N_I}$ and model points $\{\mathbf{X}_j\}_{j=1}^{N_W}$ it aims to find:

$$\underset{R,T,P}{\operatorname{argmin}} \sum_{i=1}^{N_I} \sum_{j=1}^{N_W} P_{i,j} ||K(R\mathbf{X}_j + T) - \mathbf{x}_i||^2 \quad (5)$$

where K is the known camera intrinsics, $R \in SO(3)$, T is a 3x3 translation vector and P is a permutation matrix whose entries are $P_{i,j} = 1$ if $\mathbf{x}_i$ matches $\mathbf{X}_j$ and 0 otherwise. A brief outline of the SoftPosit algorithm is given here, with more detail in [David et al., 2002].

The SoftPosit algorithm iteratively switches between assigning correspondences (P) (involving *multiple*, weighted correspondences from each point) and solving the pose (R and T) from these weighted correspondences. Furthermore, the parameter that controls this weighting (β) can be seen as a simulated annealing parameter, meaning the correspondences tend towards the binary 'one correspondence per point' as required in Equation 5. For comparison with Section 3.4, the update step is as follows:

Inputs: N_W world points, N_I image points, hypothesised pose.
Outputs: Updated pose.

1. For each world / image coordinate pair ($\mathbf{X}_j$, $\mathbf{x}_i$), project $\mathbf{x}_i$ to the same depth as $\mathbf{X}_j$ using a perspective projection, then project this back onto the image plane under a Scaled Orthographic Projection (SOP) - call this $\mathbf{p}_i$. Define the projection of $\mathbf{X}_j$ under a SOP as $\mathbf{q}_j$. Let $d_{i,j} = |\mathbf{p}_i - \mathbf{q}_j|$.
2. Compute a matrix of weights $m_{i,j} = \gamma \exp(-\beta(d_{i,j} - \alpha))$ for $\beta = 0.004$, $\alpha = 1$, $\gamma = 1/(\max\{N_W, N_I\} + 1)$. Also $\forall\ i, j$ set $m_{i,N_W+1} = m_{N_I+1,j} = \gamma$: this represents the probability of a point having no correspondence.
3. Use Sinkhorn's method to normalise each row and column (apart from the last ones) of this matrix. This is achieved by alternately normalising each row and column.
4. Use $m_{i,j}$ as the weights for weighted correspondences to update the pose by minimising a weighted sum (specifically Equation 5 with $P_{i,j} = m_{i,j}$). This is achieved simply by setting the derivative of the objective function to zero.

The algorithm is iterated from the first stage with β increasing by a factor of 1.05 each iteration - this has the effect of weighting close points significantly higher.

What is defined here as *projective saliency* is now taken into account by extracting points on the edge of the silhouette (S) of the model and aligning these with the image points. Therefore, (5) is altered and the objective is to find:

$$\underset{R,T}{\operatorname{argmin}} \sum_{j=1}^{N_W} \min_{i=1}^{N_I} ||K(R\mathbf{X}_j + T) - \mathbf{x}_i||^2 + \sum_{i=1}^{N_I} \min_{j \in W} ||K(R\mathbf{X}_j + T) - \mathbf{x}_i||^2 \quad (6)$$

where W represents the union of the original salient 3D points and the 3D points on the boundary of the silhouette S (and is thus dependent on R and T). Note that (6) no longer depends on P; this is because the problem is no longer symmetric (image points may lie on the boundary but not vice versa). Obtaining the silhouette at every iteration is an expensive operation and is only relevant when the hypothesised pose is within the vicinity of the solution. Therefore, the points are initially aligned according to (5) by SoftPosit and scored according to (6), which is subsequently minimised in Section 3.4.

In the current implementation, 50 iterations are carried out from 500 random initial poses, for which random rotation matrices were generated using the method described in [Arvo, 1992]. Further, OpenGL is used to render the 3D model and points that are not visible are not included in the update step. As this is computationally expensive, it is only done when the pose is sufficiently different from the previous time the visibility condition was checked. This difference is measured as the sum over all elements in R and T of the absolute difference between the two poses. From these 500 initialisations, the lowest scoring 1% are selected according to (6).

3.4 Boundary matching

After an initial alignment is obtained, the top 1% are subsequently refined by matching the boundary of the

silhouette with the outer edges of the image as well as matching the original 2D and 3D salient points. A Canny edge detector is first used on the image and points are extracted on the exterior contour from this. With these E exterior points, there are now N_W world coordinates and $(N_I + E)$ image points and (6) is minimised with this new set of image points. The update step is adjusted as follows:

Inputs: N_W world points, $(N_I + E)$ image points, hypothesised pose, 3D mesh.
Outputs: Updated pose.

1. Compute $d_{i,j}$ between the N_W world points and all $(N_I + E)$ image points.
2. For each of the $(N_I + E)$ image points, compute the distance between it and the *nearest* point that is on the edge of the silhouette. (Computing the distance between it and all points on the edge of the silhouette is too computationally expensive.)
3. Construct the matrix of weights $m_{i,j}$ as appropriate, taking into account an image point may be matched with its nearest point on the silhouette. This leads to a $(N_W + 2) \times (N_I + E + 1)$ matrix. Apply Sinkhorn's method for the same purpose as in Section 3.3.
4. Use these weighted correspondences to update the pose, similar to Section 3.3.

Again, β is increased by a factor of 1.05 at each iteration. The score is calculated in the same way as before, except now it uses all $(N_I + E)$ image points instead of the N_I originally extracted. The lowest scoring match is deemed the correct alignment.

4 Experiments

4.1 Experimental Setup

We applied our algorithm to six datasets; two real and four synthetic. The synthetic datasets are the Stanford bunny and Stanford dragon, and the horse and angel from the Large Geometric Models Archive from Georgia Institute of Technology. For the synthetic datasets, textureless images were generated using POV-Ray from a random angle [Arvo, 1992] and fixed translation, using a point light source at the same location as the camera. The two real datasets are the dinosaur and temple from Middlebury's multi-view reconstruction dataset [Seitz et al., 2006]. Since this does not include the ground truth, we use the 3D reconstruction provided by [Guillemaut and Hilton, 2011]. These images contain texture information of the model while the 3D data does not, adding another layer of difficulty. An example image from each dataset is shown on the right of Figure 4. In all cases, we attempted to align 30 images independently to the model.

For each attempted registration, we extracted 160 points from the 3D structure and 80 from the image. We extract these using the generalised Kadir-Brady saliency method described here, and using the Harris corner detector and the SIFT keypoint detector, both of which have a respective implementation in 3D (As there is no texture in the models, SIFT uses the curvature value instead). All three of these keypoint detectors return a response value which is used to rank them so as to obtain the optimal keypoints. We then follow the methodology described in Section 3, initialising SoftPosit with 500 random poses.

To generate the poses, we assume no prior information on the rotation and so we generate random rotation matrices according to [Arvo, 1992]. However, we initialise the translation such that the mean of the 3D data is in the centre of the image (which is the case in the synthetic data, and is sufficiently close for the real data) and initialise a random depth within 5% from the ground truth. Whilst we could have sampled over a larger range of translations, this is unrealistic as often an image of a model is focused at a suitable location to capture all of the structure.

4.2 Results and Discussion

We define three distance measures to compare the hypothesised registration with its ground truth. The first measure is a comparison of rotations. There are a number of ways to compare rotation matrices [Huynh, 2009]. Here, we define the angle between the rotation R and its ground truth rotation R_{gt} as $E_{rot}(\theta) = \arccos(R\mathbf{u} \cdot R_{gt}\mathbf{u})$ with $\mathbf{u} = \frac{1}{\sqrt{3}}[1,1,1]^T$. Secondly, we measure the distance between the hypothesised camera centre and the ground truth camera centre in normalised 3D coordinates (E_{dist}). Finally we introduce the projection error: it is calculated by projecting the model using the hypothesised projection and comparing it with the ground truth projection by comparing the two generated areas A and B. The error is defined as $E_{proj}(\%) = \frac{(A \cup B) \setminus (A \cap B)}{A \cup B}$.

The cumulative frequency curves for these measures are shown in Figure 4 for individual datasets and their average across all datasets.

Overall, the saliency detector produced the best results, particularly on the real data and the horse. It performed similarly to the other detectors for the Stanford bunny and the angel. This is shown by the different areas under the cumulative frequency curve

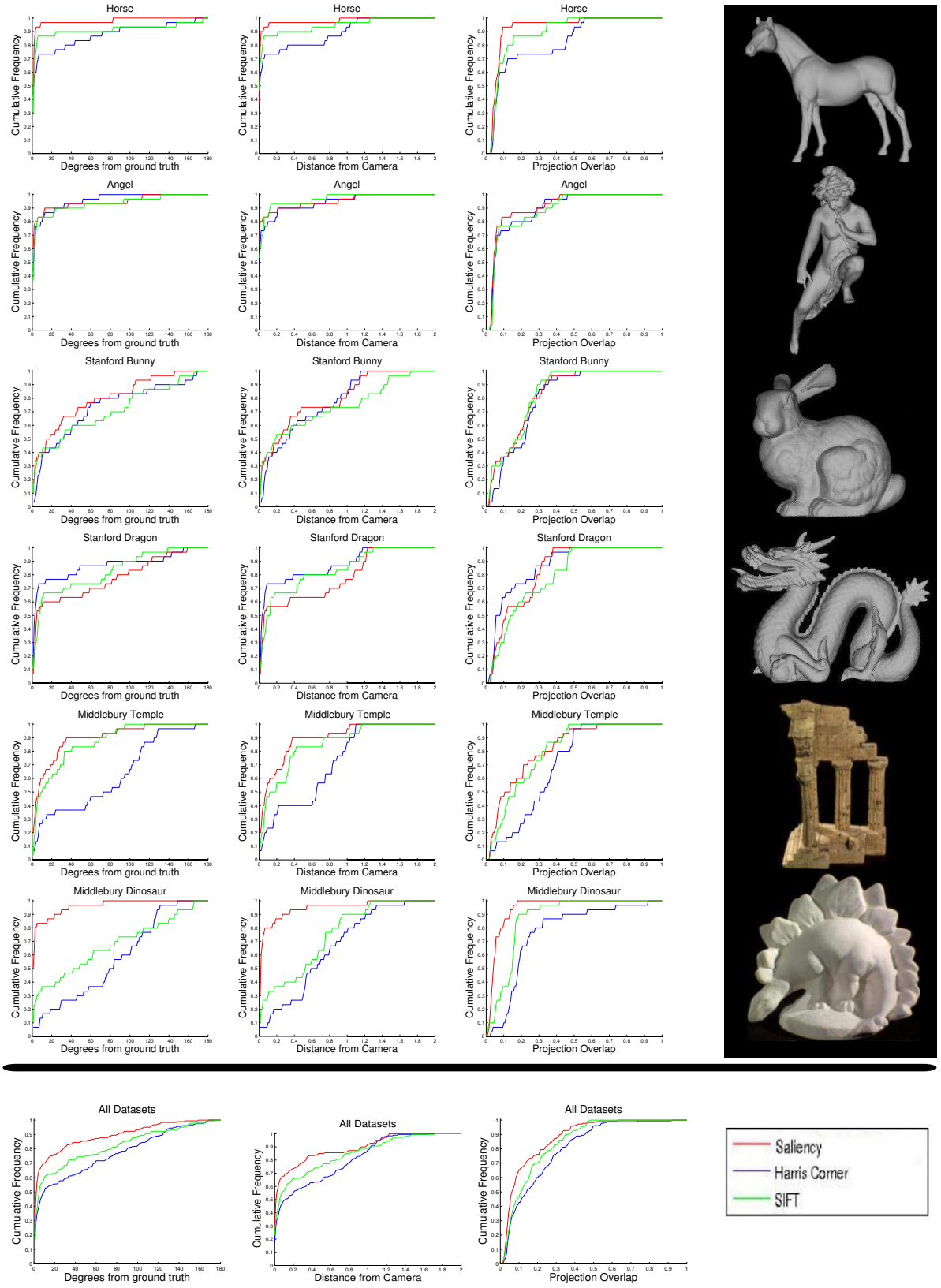

Figure 4: Images of the datasets and cumulative frequency curves for each error measure. From left to right: Rotation error; Distance between camera centres; Projective error; image of dataset. From top to bottom: horse; angel; Stanford bunny; Stanford dragon; Middlebury temple; Middlebury Dinosaur; Average across all datasets.

for each point detector when averaged across all datasets:

Dataset / Error Type	$E_{rot}(\theta)$	E_{dist}	$E_{proj}(\%)$
Harris	140	1.63	0.81
SIFT	149	1.70	0.85
Saliency	**169**	**1.72**	**0.88**

Whilst our method worked well on the Stanford dragon, it was outperformed by the Harris corner detector and SIFT; this may be due to the large amount of small corners on it, allowing the other methods to be better suited. In particular, Kadir-Brady saliency may not work as well for small features since their entropy may not be peaked in scale space (the entropy is high for the smallest scale and then decreases). Many of the results could potentially have been localised to within 1° if a better method had been used in the refinement stage. In particular, numerous edges were extracted from images of the temple, producing a lot of noise for the refinement. This could be solved by, for example, using Mutual Information for refinement as in [Mastin et al., 2009].

5 Conclusion

In this paper we have presented a generalisation of Kadir-Brady saliency to an arbitrary number of dimensions and provided a novel curvature based extension to 3D. Further, we have used this as a filter for the 2D / 3D registration problem and shown saliency to be a superior filter for point-based registration, demonstrating its consistency across different modalities. This is due to its principled information-theoretic approach, however it is only a proxy for true saliency and improvements can be made here. Future work may include line or curve-based saliency, estimating the focal length of the camera as well as its extrinsics and the extension of the principle of saliency-based filtering to other multi-modal problems.

Acknowledgements

This research was executed with the financial support of the EPSRC and EU ICT FP7 project IMPART (grant agreement No. 316564).

REFERENCES

Arvo, J. (1992). Fast random rotation matrices. pages 117–120.

Chung, A. J., Deligianni, F., Hu, X.-P., and Yang, G.-Z. (2004). Visual feature extraction via eye tracking for saliency driven 2D/3D registration. In *Proc. of the 2004 symposium on Eye tracking research & applications*, ETRA '04, pages 49–54.

David, P., DeMenthon, D., Duraiswami, R., and Samet, H. (2002). Softposit: Simultaneous pose and correspondence determination. In *Proc. of the 7th European Conference on Computer Vision-Part III*, ECCV '02, pages 698–714.

Egnal, G. (2000). Mutual information as a stereo correspondence measure. Technical report, Univ. of Pennsylvania.

Enqvist, O., Josephson, K., and Kahl, F. (2009). Optimal correspondences from pairwise constraints. In *International Conference on Computer Vision*.

Fitzgibbon, A. W. (2001). Robust registration of 2D and 3D point sets. In *British Machine Vision Conference*, pages 662–670.

Gendrin, C., Furtado, H., Weber, C., Bloch, C., Figl, M., Pawiro, S. A., Bergmann, H., Stock, M., Fichtinger, G., Georg, D., and Birkfellner, W. (2011). Monitoring tumor motion by real time 2d/3d registration during radiotherapy. *Radiother Oncol*.

Granger, S. and Pennec, X. (2002). Multi-scale em-icp: A fast and robust approach for surface registration. In *European Conference on Computer Vision (ECCV 2002), volume 2353 of LNCS*, pages 418–432.

Guillemaut, J.-Y. and Hilton, A. (2011). Joint multi-layer segmentation and reconstruction for free-viewpoint video applications. *Int. J. Comput. Vision*, 93(1):73–100.

Huynh, D. Q. (2009). Metrics for 3d rotations: Comparison and analysis. *J. Math. Imaging Vis.*, 35(2):155–164.

Itti, L. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506.

Judd, T., Durand, F., and Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. Technical report, MIT.

Kadir, T. and Brady, M. (2001). Saliency, scale and image description. *Int. J. Comput. Vision*, 45(2):83–105.

Kotsas, P. and Dodd, T. (2011). A review of methods for 2d/3d registration. World Academy of Science, Engineering and Technology.

Lee, C. H., Varshney, A., and Jacobs, D. W. (2005). Mesh saliency. *ACM Trans. Graph.*, 24(3):659–666.

Mastin, A., Kepner, J., and Fisher III, J. W. (2009). Automatic registration of lidar and optical images of urban scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*.

Moreno-Noguer, F., Lepetit, V., and Fua, P. (2008). Pose priors for simultaneously solving alignment and correspondence. In *Proc. of the 10th European Conference on Computer Vision: Part II*, ECCV '08, pages 405–418.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. (2006). A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. of the 2006 IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition - Volume 1, CVPR '06, pages 519–528.

Shao, L., Kadir, T., and Brady, M. (2007). Geometric and photometric invariant distinctive regions detection. *Inf. Sci.*, 177(4):1088–1122.

Taubin, G. (1995). Estimating the tensor of curvature of a surface from a polyhedral approximation. In *Proceedings of the Fifth International Conference on Computer Vision*, ICCV '95.

van de Kraats, E. B., Penney, G. P., Tomazevic, D., van Walsum, T., and Niessen, W. J. (2004). Standardized evaluation of 2d-3d registration. In *MICCAI (1)*, pages 574–581.

Wu, C., Clipp, B., Li, X., Frahm, J.-M., and Pollefeys, M. (2008). 3d model matching with viewpoint-invariant patches (VIP). In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008)*.

Yang, W., Yi, D., Lei, Z., Sang, J., and Li, S. Z. (2008). 2D-3D Face Matching using CCA. In *8th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2008)*, pages 1–6.