# Viewing Direction Classification: Application to View Planning

Sebastian Riedel, Zoltan-Csaba Marton, Simon Kriegel
Institute of Robotics and Mechatronics, German Aerospace Center (DLR), 82234 Oberpfaffenhofen, Germany
Email: {firstname.lastname}@dlr.de

## I. Introduction

An autonomous household robot passively observes the environment while navigating, possibly while performing other tasks, and has spare processing power to identify the different objects it encounters. For example, while setting the table for breakfast, the robot might spot a cup somewhere. Later, when the robot is asked to fetch a cup, it does not need to actively start searching for it, but instead retrieves the cup's location from memory.

We investigated such a passive world-state logging during long-term operation in household environments in the context of object classification [1], [2], [3]. However, manipulating these objects also requires the estimation of object poses. In recent work, we described a method to estimate an object's orientation using Bingham Mixture Models [5] based on standard local Fast Point Feature Histogram (FPFH) shape features [6] which can deal with high uncertainties and ambiguities.

In this work, we focus on analysing which object views and object parts are ambiguous, respective informative, with respect to estimating the object's orientation in the world. Being able to tell which object views and, more specifically, which object parts are informative is valuable information for active perception (view planning especially in the face of occlusions) as well as passive perception (saving computation by purposively selecting the images to do computation on). Encouraged by the performance improvement and simplicity of the approach to feature selection for categorization by [7], a similar approach, but targeting feature selection for orientation estimation, will be presented and applied to orientation estimation. In our previous work [5], an approach to orientation estimation based on viewing direction classification was presented. This viewing direction classification is motivated by the idea, that the classification allows an insight into which viewing directions and also which object parts are significant for the orientation estimation of the object. This work now investigates this idea for the simulated datasets and shows how a local model of informativeness is obtained using the previously described classification framework.

## II. Related Work

In the work of [7], a system for category recognition of objects is presented which improves over several other methods by incorporating a measure of feature informativeness. Standard features (they also used FPFH) describing the local geometry around a point are computed densely over training views of several objects in different categories. The implemented classification pipeline in this work is conceptually similar, except that we train probabilistic logistic regression (LR) classifiers and are interested in informative object parts for orientation estimation rather than classification.

A different approach, specifically targeted towards feature selection for pose estimation is presented in [8]. Their pose estimation is an extension of [9]. This is a voting approach based on a hash table which maps point pair features to object poses. Despite the approaches for feature selection described in [7] and [8], a probabilistic model of the feature distribution over the object's surface would also lend itself towards an analysis of surface point informativeness. In [10], a dense probabilistic model over 3d features is built by clustering the feature descriptors into words and estimating a distribution over feature orientation and position on the object for features belonging to one word. Such a formulation was presented in [5], and will be used in this work.

## III. View Informativeness

As described in detail in [5], the classification pipeline is trained to predict the training view direction a feature is observed from. As the used LR classifier outputs a probability over training view directions $p(D = m|f)$ given a single feature $f$, the classification pipeline implicitly encodes a model of view-related surface ambiguity. The broader and more uniform a feature's view distribution is, the less it tells us about how the object is oriented with respect to the camera.

We can exploit the classifier's model of surface ambiguity by evaluating which training view directions have features which identify the view correctly. The training view which results in the most unambiguous classification in this sense is the most informative for estimating the object's orientation. For every training direction $m'$ separately, the training views are analyzed by first extracting features with the same settings as used by the online applied classifier. The features are ranked according to the entropy of their classification distribution and the top $N_{feat}$ features are selected to estimate the view's informativeness. This basic feature ranking and selection procedure is the same as performed before orientation estimation and described in [5]. The selected feature's view distributions are summed and a measure of correctness is obtained by calculating the discrete Kullback-Leibler (KL) divergence between the correct distribution $p^*$ and the extracted summed distribution $p_{sum}$. The KL divergence measures the difference of the extracted distribution from the theoretically correct distribution and is defined as

$$d_{KL}(p^*||p_{sum}) = \sum_m p^*(m) \ln \frac{p^*(m)}{p_{sum}(m)} \quad (1)$$

$$= \ln \frac{1.0}{p_{sum}(m')} \quad (2)$$

where the correct distribution is defined as 1.0 for the training view direction of concern $m'$ and 0.0 everywhere else. The simplification in the second equality is thus possible due to the form of $p^*$. We can see that the KL divergence is zero for a perfect summed feature distribution with $p_{sum}(m') = 1.0$ and goes to positive infinity as $p_{sum}(m')$ approaches zero. For an expected informativeness ranking of a viewing direction, the KL divergences of all training point clouds for that direction are averaged and ranked in ascending order. The resulting ranking is illustrated in figure 1 using simulated data for the cartoon and mug model. For interpretation purposes, some of the viewing directions are illustrated by means of a rendering of the object observed from that viewing direction.

For the mug model one can observe that the best viewing directions lie on the plane defining the reflective symmetry of the mug whereas the least informative views show large parts of the body of the mug. This is intuitively correct as features on the body of the mug can be observed from many directions. Furthermore, as the features are rotational invariant and the classification is based on individual features and therefore local information, most views, even if they show the handle, are ambiguous as the reflective symmetry cannot be resolved. The most unambiguous views are therefore correctly identified as the ones on the reflection plane which additionally show large parts of the handle or the inside of the mug.

For the cartoon model there is no clear intuitive ranking of views from a human perspective, but we will shed light on which parts of an object are informative and thus the reason for this ordering within the next section.

## IV. Model Surface Informativeness

The view ranking in the previous section was based on accumulating information of several features of a view into a summed distribution and assessing the correctness of this distribution. This way we obtained information about the informativeness of a viewing direction. In this section, we accumulate information of features within a small neighborhood of a point on the object's surface and thus assess how informative a surface point is.

For the method presented here, we assume the availability of a set of points $S = \{s_0, \ldots, s_N\}$, $s_i \in \mathbb{R}^3$ representing the complete surface of the object. For the cartoon and mug object in this evaluation, we have 3d models and thus the set $S$ was generated by sampling the surface of the 3d model with uniform density using the stratified sampling approach described in [11] and [12]. Our objective is now to score every surface points $s_i$ by means of how informative the point is or more precisely, how informative features originating from that point are. As a surface point may be visible from more than one viewing direction, at first a scoring matrix $K \in \mathbb{R}^{N \times M}$ is computed which ranks the $N$ surface points separately with respect to the $M$ training view directions. For a given surface point $s_i$ and viewing direction $m$, the nearest neighbor points $\mathrm{NN}_m(s_i)$ within a radius of $0.5\mathrm{cm}$ are obtained in the training point clouds for that viewing direction. The corresponding features $\mathrm{NNF}_m(s_i)$ in the training point clouds are extracted

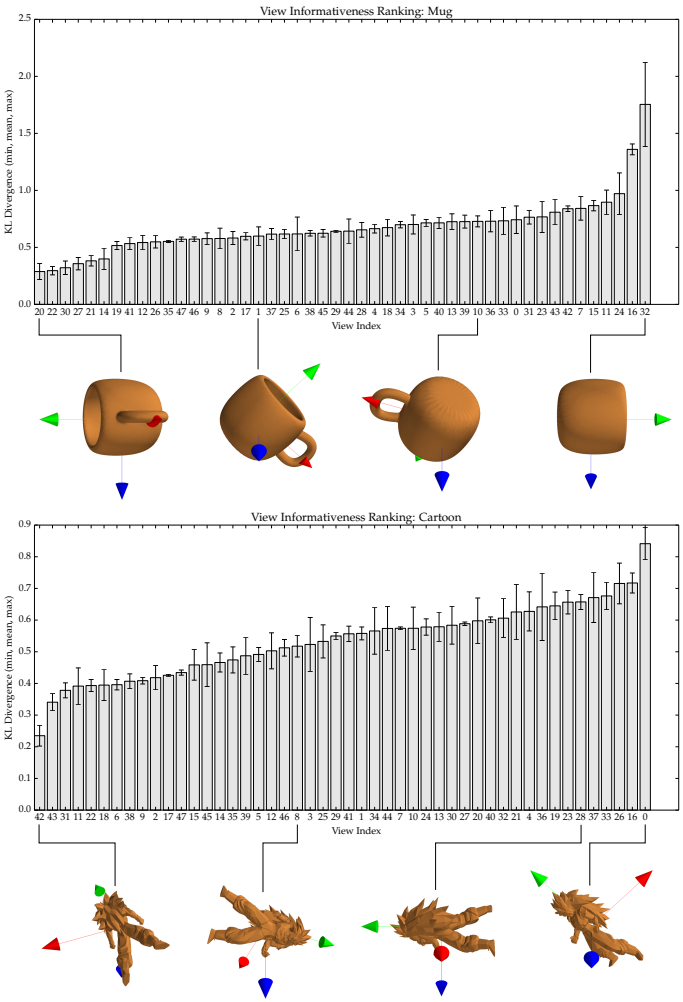


Fig. 1: Training directions ranked by average view KL divergence for mug and cartoon model. Below the bar charts, selected views are illustrated by renderings of the object from those views.

and the average KL divergence between the correct and the predicted view distributions is calculated and stored

$$K[i,m] = \frac{1}{|\mathrm{NNF}_m(s_i)|} \sum_{f \in \mathrm{NNF}_m(s_i)} d_{KL}(p^* || p(D|f)) \quad (3)$$

For further reference, the average KL divergence stored at $K[i,m]$ will be termed view-conditional score of surface point $i$ to viewing direction $m$. If a surface point is not observable from a direction and hence no nearest neighbors could be found, the score is set to -1 to indicate this. The scoring matrix $K$ thus encodes the visibility and view-conditional informativeness of every surface point by assessing the average prediction correctness of features computed at these locations. A global, not view-conditional measure of a points informativeness is obtained by averaging the point's scores for all viewing directions it was observed from.

In figure 2 the view-conditional scores for a set of views - the same views as in figure 1 - are illustrated through a heatmap visualization. In other words, every pair of rendered

and heatmap images visualizes a specific column of the score matrix $K$ for that object. The colormap was chosen so that the color white corresponds to a KL divergence of zero and black corresponds to the median KL divergence of the complete scoring matrix (ignoring the -1 for non-visibility). This way, the same colormap is used for all shown views of an object and the heatmaps can be compared to each other. White color indicates, that features at this surface point are reliably recognized as originating from the viewing perspective shown. The views presented are ordered left to right by the overall view ranking extracted in the previous section and thus we clearly see which object parts make the most informative view (most left) better than the least informative view (most right). For the mug, our intuition that the handle is more informative than the body is now quantitatively proved. For the cartoon object, is seems that the concave regions within the character's hair as well as the rear part make the best view so informative.

Another interesting aspect is revealed when taking a closer look at the two left-most views of the mug. The upper handle part is colored white in both views which might seem contradictory at first as this means that features originating from the same physical region can be reliably classified to more than one view. This behavior can be explained by remembering that features are computed over geometry within a certain radius (here $3\text{cm}$) and thus encode the view-specific self-shadowing of the object, which turns out to be very descriptive.

In figure 3, the global KL score is visualized by means of averaging a surface point's view-conditional score over all viewing directions. The colormap is scaled on a per object basis to show white for the lowest observed KL score and black for the highest observed score (first column for each object) or the median of the observed scores (second column of each object). This measure and the illustration show where distinctive features on the object's surface can be expected, independent of the viewing direction. For the mug model, again, regions on and around the handle are generally distinctive. For the cartoon model, a general observation is the higher average classification correctness of features and surface points. For the cartoon character the average KL divergence over all model points is $1.69$ versus $2.15$ for the mug model. Via the equation (2) this results in an average probabilistic weight for the correct viewing direction of $18.4\%$ for the cartoon character versus $11.6\%$ for the mug model. Regions of high informativeness for the cartoon character appear to be within the character's hair, the hands and the rear. Also, pointed surface regions like the feet and hair tips show high ambiguity or wrong classification which is probably due to unstable normal estimation in those areas.

## V. Evaluation of Informativeness Values

In order to show the practical value of the extracted model surface informativeness, a proof-of-concept experiment using the simulated mug model under varying degrees of occlusion was conducted. The visible part of the mug was manually chosen to consist of regions of high informativeness according to the findings in the previous section and figure 3. 20 random view sequences with 20 views per sequence have been

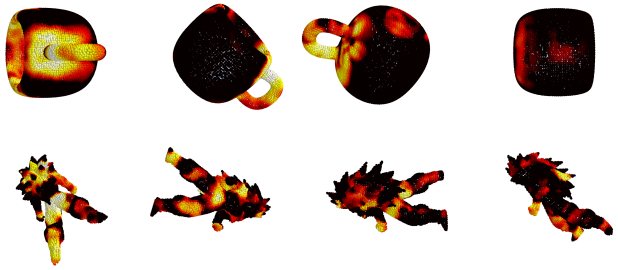

Fig. 2: View-conditional KL scores for chosen example viewing directions. For comparability, the same viewing directions as shown in figure 1 are shown. The view-conditional KL scores are shown via a heatmap-visualization, with white representing a KL value of zero (low classification ambiguity) and black representing the median view-conditional KL value of the score matrix K.

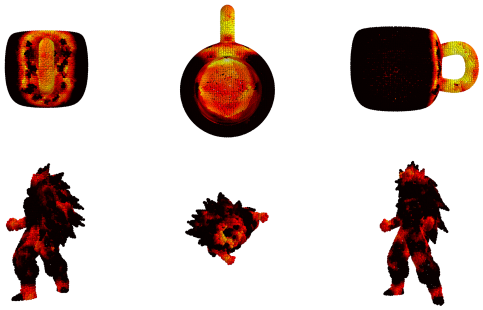

Fig. 3: Global KL score obtained by averaging view-conditional scores over all viewing directions for every surface point. For each object a view along x-axis, y-axis and z-axis are given. Bright colors signal low values. Values greater than the median are shown in black for clarity.

generated and the development of the maximum a posteriori (MAP) rotational error after each view is analyzed.

To simulate the occlusion of parts of the mug which have been found to be uninformative (mainly the body of the mug), two points on the mug handle where selected manually, one on the upper side of the handle and one on the lower side, which together with a visibility radius $r_{vis}$ around those points define the fixed observable region of the mug. The simulation pipeline proceeds by first generating a complete point cloud of the mug as seen from a given viewing direction and then selecting the sub-cloud within the distance $r_{vis}$ around the two selected points as final simulation output. Normal estimation and feature computation is then done on the extracted sub-cloud. The experiment was performed six times with visibility radii in $r_{vis} \in [3\text{cm}, 4\text{cm}, 5\text{cm}, 6\text{cm}, 7\text{cm}, 8\text{cm}]$. For radii of $6\text{cm}$ and larger, the visible part includes surface regions on the rim and the inside of the mug, which allow a unique orientation estimate in contrast to radii smaller than $6\text{cm}$.

A summarizing comparison between all visibility radii is given in figure 4 by displaying the median MAP error over all 20 sequences for the different visibility settings. As observable in the median plots, the sequential estimation converges slower as the visible surface region gets smaller. For visibility radii of

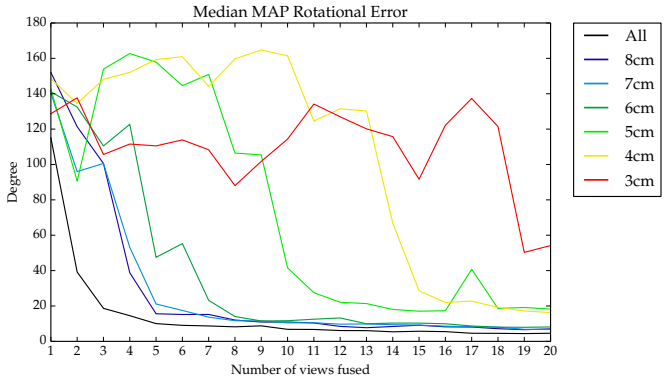


Fig. 4: Median MAP error over all sequences for all visibility settings. The visible surface area for all is 100% naturally, and 30.8%, 24.6%, 18.1%, 11.4%, 7.7%, 4.9%, respectively.

8cm, 7cm and 6cm the error after convergence is comparable to the baseline experiment with no occlusion (visibility radius 'All'). Starting with a visibility radius of 5cm and smaller, the error of convergence gets significantly larger. For the 5cm setting, this is largely due to an ambiguity between the upright (mug opening in positive z-direction) and the flipped orientation (opening in negative z-direction) which arises because the handle is symmetric and the visible surface area does not include the rim and inner surface of the mug. Due the orientation representation as Bingham mixture model, an interesting question here is whether the flip-ambiguity is present in the orientation estimate as two separate mixture components. An investigation for the 5cm case revealed, that this is, however, not the case and random sequences either converge to a unimodal distribution with the mode close to the flipped or the non-flipped orientation.

Overall, the results presented show the robustness of the orientation estimation to occlusions of up to 81.9% respective $r_{vis} \geq 6\text{cm}$. This is achieved due to the local and correspondence-less nature of the viewing direction classification. It also shows that model surface informativeness ranking extracts surface areas relevant for orientation estimation as the increase in orientation error when occluding presumably uninformative parts is small. Therefore, it seems valuable to actively plan views in such a way that regions of high estimated informativeness are visible.

## VI. Conclusions and Outlook

In summary, we presented an information-theoretic approach for estimating the informativeness of an object's local geometry with respect to orientation estimation. Identifying the regions and viewing directions which show the most unambiguous local features is of interest for long-term autonomy as it enables to actively plan to observe these features as well as to actively select image frames out of passively acquired image streams. In both cases, we expect an overall increased computational efficiency of the autonomous agent and we plan to integrate these results into our next-best-view and scene analysis system from [4].

A remaining challenge for long-term autonomy is the use of incomplete models that are built online on a mobile robot. A subject for further evaluation would therefore be using an online-trainable classifier, e.g. a Mondrian Forest classifier [13], as internal model of view-related ambiguity together with online surface models obtained with our approach in [4].

Another avenue for further research is how such local models of informativeness can be used to more efficiently guide sampling based pose estimation algorithms (e.g. [6]) which would potentially improve algorithm runtime as well as model storage requirements at the cost of a more expensive offline training stage.

**Acknowledgments:** This work has partly been supported by the European Commission under contract number H2020-ICT-645403-ROBDREAM.

## References

[1] N. Blodow, D. Jain, Z.-C. Marton, and M. Beetz, "Perception and Probabilistic Anchoring for Dynamic World State Logging," in *10th IEEE-RAS International Conference on Humanoid Robots*, Nashville, TN, USA, December 6-8 2010, pp. 160–166.

[2] Z. C. Marton, D. Pangercic, N. Blodow, and M. Beetz, "Combined 2D-3D Categorization and Classification for Multimodal Perception Systems," *The International Journal of Robotics Research*, vol. 30, no. 11, pp. 1378–1402, September 2011.

[3] M. Beetz, F. Balint-Benczedi, N. Blodow, D. Nyga, T. Wiedemeyer, and Z.-C. Marton, "RoboSherlock: Unstructured Information Processing for Robot Perception," in *IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, Washington, USA, 2015, Best Service Robotics Paper Award.

[4] S. Kriegel, M. Brucker, Z.-C. Marton, T. Bodenmuller, and M. Suppa, "Combining object modeling and recognition for active scene exploration," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 2384–2391.

[5] S. Riedel, Z.-C. Marton, and S. Kriegel, "Multi-view Orientation Estimation using Bingham Mixture Models," in *Proceedings of the International Conference on Automation, Quality and Testing, Robotics*. IEEE Computer Society – Test Technology Technical Council, 2016.

[6] A. Aldoma, Z.-C. Marton, F. Tombari, W. Wohlkinger, C. Potthast, B. Zeisl, R. B. Rusu, S. Gedikli, and M. Vincze, "Tutorial: Point cloud library: Three-dimensional object recognition and 6 dof pose estimation," *Robotics & Automation Magazine, IEEE*, vol. 19, no. 3, pp. 80–91, 2012.

[7] M. Madry, H. M. Afkham, C. H. Ek, S. Carlsson, and D. Kragic, "Extracting essential local object characteristics for 3D object categorization," *Intelligent Robots and Systems, 2013 IEEE/RSJ International Conference on*, pp. 2240–2247, Nov. 2013.

[8] O. Tuzel, M.-y. Liu, Y. Taguchi, and A. Raghunathan, "Learning to Rank 3D Features," in *Computer Vision (ECCV), 2014 European Conference on*. Springer, 2014, pp. 520–535.

[9] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3D object recognition," in *Computer Vision and Pattern Recognition, 2010 IEEE Computer Society Conference on*. Ieee, Jun. 2010, pp. 998–1005.

[10] J. Glover, G. Bradski, and R. B. Rusu, "Monte Carlo pose estimation with quaternion kernels and the bingham distribution," *Robotics: Science and Systems*, 2012.

[11] D. Nehab and P. Shilane, "Stratified Point Sampling of 3D Models," in *Point-Based Graphics, IEEE/Eurographics Symposium on*, 2004.

[12] D. Doria, "Stratified Mesh Sampling for VTK," *The VTK Journal, March*, vol. March, 2010.

[13] B. Lakshminarayanan, D. M. Roy, and Y. W. Teh, "Mondrian forests: Efficient online random forests," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3140–3148.