# Low-level global features for vision-based localization

Sven Eberhardt and Christoph Zetzsche

Cognitive Neuroinformatics,
Universität Bremen, Bibliothekstrae 1, 28359 Bremen, Germany
sven2@uni-bremen.de,zetzsche@informatik.uni-bremen.de

**Abstract.** Vision-based self-localization is the ability to derive one's own location from visual input only without knowledge of a previous position or idiothetic information. It is often assumed that the visual mechanisms and invariance properties used for object recognition will also be helpful for localization. Here we show that this is neither logically reasonable nor empirically supported. We argue that the desirable invariance and generalization properties differ substantially between the two tasks. Application of several biologically inspired algorithms to various test sets reveals that simple, globally pooled features outperform the complex vision models used for object recognition, if tested on localization. Such basic global image statistics should thus be considered as valuable priors for self-localization, both in vision research and robot applications.

**Keywords:** localization, visual features, dataset, spatial cognition

## 1 Introduction

The ability to make reliable assumptions about their own position in the world is of critical importance for biological as well as for man-made systems such as mobile robots. A number of sensors can be used and combined to achieve this feat (see e.g. [4]). Among these, vision is of particular importance. Although idiothetic information such as acceleration, velocity and orientation measurements can be used for dead reckoning, visual realignment can be essential to avoid the accumulation of errors in path integration. Furthermore, allothetic information in form visual input can be used for direct localization. For example, many place cells in the hippocampus can be driven by visual input alone [10]. But how exactly can vision support localization?

The default hypothesis would be that this is achieved by just the same estabished principles of visual processing used for other spatial tasks like pattern discrimination or object recognition. The corresponding standard view of the visual system assumes that the main task of the system is invariant object recognition, and that this is achieved by a feed-forward system of feature extraction in form of a hierarchy of neural layers with increasing levels of abstraction and of spatial granularity [5, 6, 16]. This standard model is supported by numerous

behavioral experiments and electroencephalography recordings, in particular by experiments showing that human discrimination between categories in object and scene classification is achieved as early as 150ms after stimulus onset (for an overview see [16]).

When looking at vision-based self-localization from static allothetic input alone, it can be formulated as a classification problem. A set of example images per location is trained with their location as the label and a new image can be attributed to one of the learned locations by testing the classifier. In this perspective, localization appears to be very similar to object classification problems such as those posed by popular object recognition datasets like Caltech-101 [3].

Generally, the features on which a classifier operates should be invariant to changes within a class but selective to changes between classes. Models designed for object recognition provide varying degrees of translation and scale invariance [13]. For example, the HMax features used for an animal detection task performed by [16] are designed to provide translation and scale invariance at local and global levels because animals may occur at different positions, sizes and 3D rotations in images.

However, whether object recognition and vision-based localization are really similar problems and can thus be solved with the same architecture has, to our knowledge, never been investigated systematically. In this paper, we ask whether visual features that are optimal for one task may be unsuited for the other and vice versa. To answer this question, we test how well feature outputs of a number of biologically inspired low-level vision models are able to discriminate among large numbers of locations and compare the results with benchmark performance on several object and scene recognition datasets.

## 2 Methods

**Streetview dataset** We use a novel dataset which has been sampled from Google Street View [1]. Street View has become popular as an outdoor dataset of natural scenes for self-localization , 3D map reconstruction , text recognition and image segmentation. Some unique key advantages to this dataset are its sheer amount of available data from many countries of the world, preprocessed in a standardized manner without bias to object centering [14]. Caveats include a bias to roads and populated areas, as well as relatively poor image quality with distorted edges and Google watermarks.

204 locations are selected by picking random points in the sampling region until a road for which street view data is available is found within 50m range. For each location, a full 360° yaw rotation in intervals of 10° for a total of 36 pictures per location is sampled. Field of view is 90° and pictures are stored as grayscale images with size 512x512 pixels. The Streetview dataset is sampled from random locations in France (SV-Country). To test localization on several distance scales, we generate two additional datasets from different sampling regions. For SV-City, we sample locations in Berlin city center only. SV-World consists of imagery from all countries where street view was available.

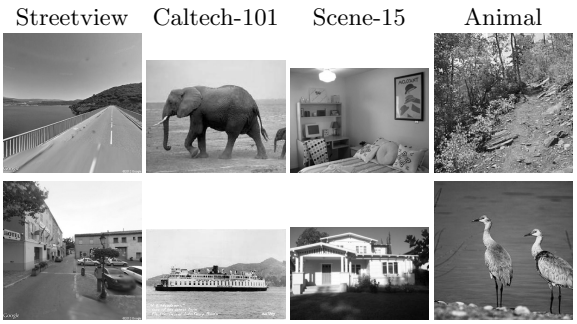


**Fig. 1.** Example images of the assayed datasets.

**Benchmark datasets** To compare localization with object recognition and scene classification tasks, we also use several established categorization databases. The first dataset is Caltech-101 [3], which is a very diverse collection of 101 object categories containing between 31 and 800 images each. Categories are diverse and include specific animals, musical instruments, food categories, vehicles and more. Image contents vary between isolated objects, comic depictions and scenes containing the object in use. The dataset has been used as a benchmark for object recognition by a number of algorithms in the past, including an implementation of HMax and Spatial Pyramids. Caltech-101 is sometimes criticized because low-level algorithms can perform relatively well on some categories due to their very similar sample images [14]. However, the large number of categories alleviates this.

For a scene classification test, we use Scene-15 [7], which is a dataset comprised of photos of 15 different indoor and outdoor scene categories such as kitchen, forest and highway. Each photo shows an open scene without any objects close to the camera. Scene-15 has been mostly used to benchmark holistic feature extraction models such as Gist and Spatial Pyramids.

Finally, we include the Animal detection dataset from Serre et al. [16], which is a two-class classification object recognition dataset showing mostly non-urban outdoor scenes both with and without animals.

**Models** We focus on low-level, biologically inspired models that produce a fixed-size feature vector for each input image. For all models, we use implementation code supplied by the authors if available.

**Textons** by Malik et al. [9] apply a set of Gabor filters to an image, resulting in a response vector for each pixel. The response vectors are clustered into 128 textons and each pixel is assigned the cluster with the least square distance to its response vector. The resulting output vector is a histogram of these texton assignments over the whole image. Textons have been used for image segmentation purposes [9] as well as scene classification [15].

**Gist** is also termed the *Spatial Envelope* of a scene by Oliva et al. It consists of the first few principal components of spectral components on a very coarse grid (8x8) as well as on the whole image. Gist has shown strong categorization performance on the Scene-15 dataset [12].

**Spatial Pyramids**, as described by Lazebnik et al. [7], calculate histograms over low-level features over image regions of different size and concatenates them to one large feature vector. The features used here are densely sampled SIFT [8] descriptors. For better comparability with the other models, we omit the custom histogram matching support vector machine (SVM) kernel used by Lazebnik in favor of a linear kernel and regression. We test the full pyramid up to level 2 (SPyr2) as well as outputs of the global histogram (SPyr0) only.

**HMax** is a biologically motivated multi-layer feed-forward model designed to mirror functionality found in the primate visual cortex ventral stream by Hubel and Wiesel [6]. It is based on the *Neocognitron* [5] and consists of alternating layers of simple and complex cells. Simple cell layers match a dictionary of visual patterns at all image locations and several scales, so units achieve selectivity to certain patterns. Complex cell layers combine the outputs of simple cells over a windows of locations and scales to achieve location and scale invariance. In this way, units of low layers have localized receptive fields and simple patterns, while units of higher layers respond to more complex patterns and are more translation and scale invariant. We use the CNS [11] implementation of HMax with parameter settings as chosen by Serre et al. [16]. The full feature vector of an image processed by HMax consists of randomly selected subsets of outputs of the C1, C2, C2b and C3 layers. In order to determine the effects of increasing invariance and matching to complex features, we also test performance when using only outputs of the C1, C2 and C3 layers respectively. To test if the task can be solved on trivial, low-level features, the classifier is also run on a luminance histogram and on a random subset of 2000 pixels from the images.

**Classification** is done on a normalized feature set which has been reduced to 128 features per image by principal component analysis. On these features, we perform regression with a linear kernel and leave-one-out cross validation to determine the regression parameter using the GURLS package [18] for MATLAB. Multi-class classification is performed by the one-versus-all rule. An equal number of training samples is taken at random from each class and all remaining elements are used for testing. Each run is repeated ten times with different test splits to yield the reported performance average and a standard deviation. Performance is defined as the averaged percent correct over all classes.

## 3 Results

All algorithms achieve between 28 and 76 percent correct performance on our dataset (figure 2a). Performance ranges are similar to those found in the benchmark sets, which shows that our dataset has a comparable difficulty. Despite the dataset similarity in difficulty, we find that classification on Texton features yield the highest rank on all tasks of the streetview dataset, while they rank

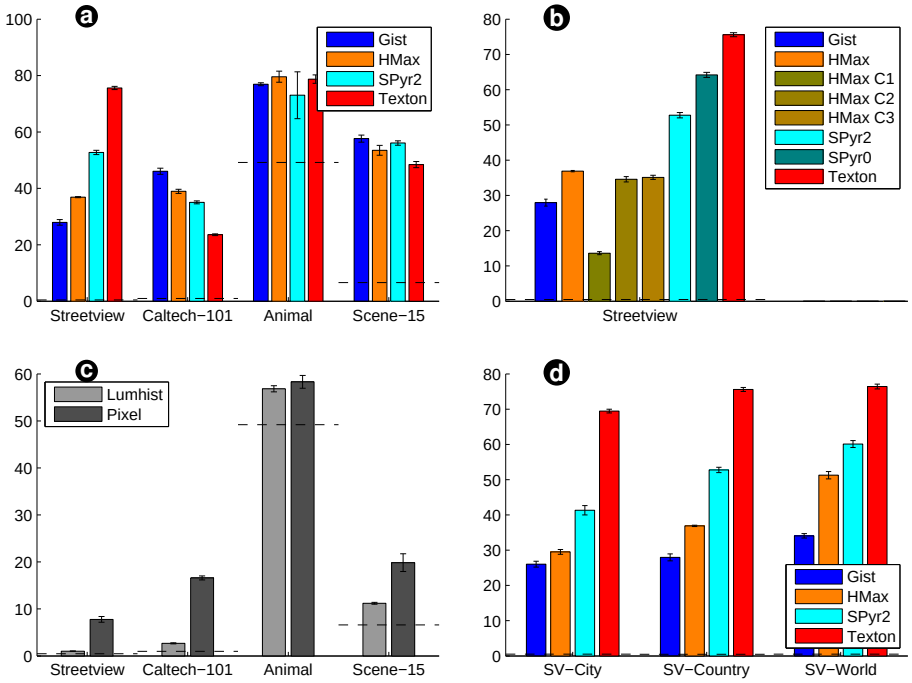


**Fig. 2.** Results performances of selected models and datasets in percent correct. Dashed lines mark chance level.

lowest on all other datasets. In particular, we do not observe this effect on the Scene-15 dataset, which hints that the requirements for scene classification are quite different from a true self-localization task. The strong performance of Textons is specially surprising, because they are the most basic and simple features in comparison with the outputs of HMax, Spatial Pyramids and Gist and they also output the least number of feature dimensions.

Spatial pyramids rank second on the performance scale. However, a test on the base level pyramid features (SPyr0 on figure 2b) reveals that the performance at level zero of the pyramid exceeds that of the full pyramid at level two. Since the base level is just a histogram over densely sampled SIFT descriptors, classification actually happens based on a global histogram similar to that of the Textons. This means that any information about spatial arrangement of features is actually detrimental to self-localization performance.

The results suggest that the task is too easy in the sense that low-level features are sufficient to achieve high performance. However, tests on global luminance histograms as well as random image pixels show low performance near chance level (figure 2c). In that sense, our self-localization dataset is harder than

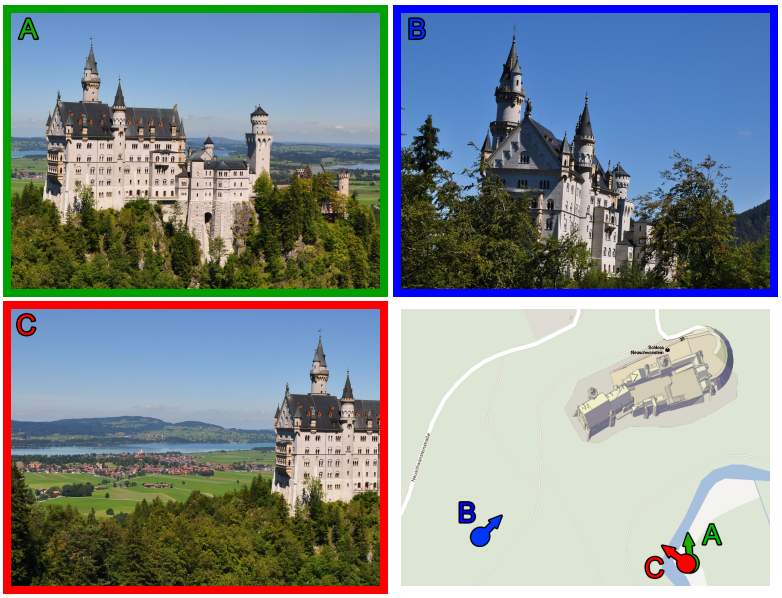


**Fig. 3.** Example for invariance requirements for object recognition versus localization: Views A and B show the same object from different locations, A and C show different views from the same location. An object classifier might pick up the similar castle features like towers and windows and put A an C into the same category. A self-localizer must not match such features and treat A and B as equal categories only.[1]

the benchmark datasets, for which 8-16% of all test samples could be classified based on raw pixel data alone.

Our findings generalize along different image sampling scales at SV-City, SV-Country and SV-World level (figure 2d). Performance is higher at larger sampling scales, because locations are more different on a world scale than on a city scale. However, the performance order among different models remains the same.

## 4 Discussion

The results show quite clearly that model performance is highly task-dependent and there are no universal features that are optimal for any vision-based task. The main reason for this finding is that there are key differences in the invariance properties required for self-localization compared to those inherent to object or scene classification [2, 20].

While object recognition needs to be tolerant to changes in scale and rotation, self-localization does not (see figure 3). Similarly, object recognition needs to be

[1] Photos: ©Stephen & Claire Farnsworth via flickr, license CC-BY-NC. Map: Google maps ©Google inc.

invariant to some feature rearrangements that occur when the object is seen from different angles. For self-localization, invariance to such rearrangements may be unwanted because if you see an object from a different angle, you are likely standing at a different position.

Concerning these invariances, HMax has both local translation and scale invariance built into the model. Thus it is not surprising that Streetview classification performance on these features is relatively poor. The differing invariance requirements also explain why neither Gist nor the pyramid structure of the spatial pyramid model could show strong performance on the dataset although both algorithms have been established for scene classification tasks [12]. Both models include features that are not completely location invariant, but contain the position in the image on a very coarse scale.

Classifying scenes in datasets like Scene-15 might actually be closer to a task like sorting photos, where photographers have a certain bias to how types of scenes are best portrayed and reflect that in the spatial arrangement of image features. Scene classification algorithms like Gist can catch on that common structure and use it for classification. However, when images from locations are recorded at random, unbiased angles, this method breaks down.

Although salient features are believed to be advantageous for localization [17, 19], we also find that the performance on complex SIFT descriptors is lower than on the more simple Textons. This is probably due to their high selectivity to particular objects, so they do not generalize well to matching on other, similar objects present in other views from the same location.

It appears surprising that Texton features, which have been designed for image segmentation [9], perform so well on a localization task. The reason seems to be that – among the models tested – they provide the best tradeoff between specificity to features present at individual locations and invariance to different views from the same location. The strong correlation of simple, global features with location suggests that very basic histogram features can be used as priors for self-localization algorithms for example in mobile robots instead of relying on geometric relations between complex features only. It also suggest that it might be worthwhile to check whether biological systems make use of such features to determine their own location.

**Acknowledgments.** This work was supported by DFG, SFB/TR8 Spatial Cognition, project A5-[ActionSpace].

## References

1. Google Street View, `http://google.com/streetview`
2. Eberhardt, S., Kluth, T., Zetzsche, C., Schill, K.: From Pattern Recognition to Place Identification. In: Spatial Cognition, International Workshop on Place-Related Knowledge Acquisition Research. pp. 39–44 (2012)
3. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object cate-

gories. In: IEEE. CVPR 2004, Workshop on Generative-Model Based Vision. vol. 106 (2004)
4. Filliat, D., Meyer, J.: Map-based navigation in mobile robots:: I. a review of localization strategies. Cognitive Systems Research 4(4), 243–282 (2003)
5. Fukushima, K.: Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biological Cybernetics 36, 193–202 (1980)
6. Hubel, D., Wiesel, T.: Receptive fields and functional architecture of monkey striate cortex. The Journal of physiology pp. 215–243 (1968)
7. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. vol. 2, pp. 2169–2178. Ieee (2006)
8. Lowe, D.: Object recognition from local scale-invariant features. In: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on. vol. 2, pp. 1150–1157 (1999)
9. Malik, J., Belongie, S., Leung, T., Shi, J.: Contour and Texture Analysis for Image Segmentation. International Journal of Computer Vision 43(1), 7–27 (2001)
10. Markus, E.J., Barnes, C.a., McNaughton, B.L., Gladden, V.L., Skaggs, W.E.: Spatial information content and reliability of hippocampal CA1 neurons: effects of visual input. Hippocampus 4(4), 410–421 (1994)
11. Mutch, J., Knoblich, U., Poggio, T.: {CNS}: a {GPU}-based framework for simulating cortically-organized networks. Tech. Rep. MIT-CSAIL-TR-2010-013 / CBCL-286, Massachusetts Institute of Technology, Cambridge, MA (2010)
12. Oliva, A., Hospital, W., Ave, L.: Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
13. Pinto, N., Barhomi, Y., Cox, D.D., Dicarlo, J.J.: Comparing State-of-the-Art Visual Features on Invariant Object Recognition Tasks. In: Applications of Computer Vision (WACV), 2011 IEEE Workshop on. pp. 463–470 (2011)
14. Ponce, J., Berg, T.L., Everingham, M., Forsyth, D.A., Hebert, M., Lazebnik, S., Marszalek, M., Schmid, C., Russell, B.C., Torralba, A., Williams, C.K.I., Zhang, J., Zisserman, A.: Dataset Issues in Object Recognition. Springer Berlin Heidelberg (2006)
15. Renninger, L.W., Malik, J.: When is scene identification just texture recognition? Vision research 44(19), 2301–2311 (2004)
16. Serre, T., Oliva, A., Poggio, T.: A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences 104(15), 6424–6429 (2007)
17. Sim, R., Elinas, P., Griffin, M., Little, J.: Vision-based SLAM using the Rao-Blackwellised particle filter. In: IJCAI Workshop on Reasoning with .... pp. 9–16 (2005)
18. Tacchetti, A., Mallapragada, P.K., Santoro, M., Rosasco, L.: GURLS : a Toolbox for Large Scale Multiclass Learning. In: presented at Workshop: "Big Learning: Algorithms, Systems, and Tools for Learning at Scale" at NIPS (2011), `http://cbcl.mit.edu/gurls/`
19. Warren, D.H., Rossano, M.J., Wear, T.D.: Perception of Map-Environment Correspondence : The Roles of Features and Alignment. Ecological Psychology 2(February 2013), 131–150 (1990)
20. Wolter, J., Reineking, T., Zetzsche, C., Schill, K.: From visual perception to place. Cognitive processing 10, 351–354 (2009)