Dimensionality Reduction in images for Appearance-based camera Localization

Silvia Luengo, Alberto Jaenal, Francisco A. Moreno, and Javier Gonzalez-Jimenez
Machine Perception and Intelligent Robotics Group (MAPIR-UMA)
Malaga Institute for Mechatronics Engineering and Cyber-Physical Systems (IMECH.UMA)
University of Malaga, Spain

Abstract

Appearance-based Localization (AL) focuses on estimating the pose of a camera from the information encoded in an image, treated holistically. However, the high-dimensionality of images makes this estimation intractable and some technique of Dimensionality Reduction (DR) must be applied. The resulting reduced image representation, though, must keep underlying information about the structure of the scene to be able to infer the camera pose. This work explores the problem of DR in the context of AL, and evaluates four popular methods in two simple cases on a synthetic environment: two linear (PCA and MDS) and two non-linear, also known as Manifold Learning methods (LLE and Isomap). The evaluation is carried out in terms of their capability to generate lower-dimensional embeddings that maintain underlying information that is isometric to the camera poses.

Keywords: Appearance-based Localization, Dimensionality Reduction, Manifold learning.

1 INTRODUCTION

Images are two-dimensional projections of a 3D scene, where the intensity of each pixel is originated by the light captured by the pixel area. This representation, despite being 2D, is suitable to provide information about the 3D structure of the environment (e.g.: walls, objects, etc.). Furthermore, we can also infer some underlying information from them, such as the camera pose where the image was taken at. This is the well-known Visual Localization [14] problem.

During the last decades, most localization approaches on Computer Vision and Robotics have relied on image local features [12, 6] to infer 3D information, operating by extracting, matching and re-projecting the most salient elements in the image to produce 3D representations of the landmarks in the scene.

As an alternative, Appearance-based Localization (AL) models images as high-dimensional vectors, produced by stacking all their pixels into a single 1D array. This way, when the camera moves on a plane, for example, it is assumed that this image vector changes following a three dimensional surface embedded on the image space, that is, forming a manifold whose latent variables are those of the camera pose $p = (x, y, \theta)$.

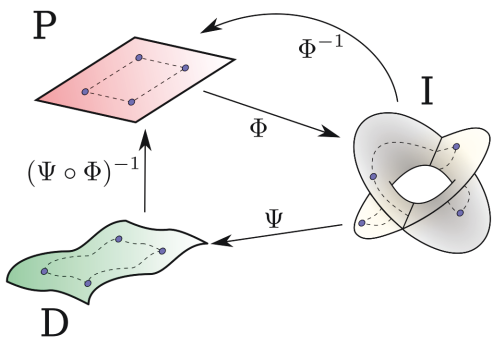


Figure 1: Relationships between the involved spaces: the pose space $\mathbf{P} \in SE(2)$, the image space $\mathbf{I} \in \mathbb{R}^N | N \gg 100$, and the low-dimensional descriptor space $\mathbf{D} \in \mathbb{R}^3$. The imaging function Φ relates poses and images, while the embedding function Ψ is in the core of DR and maps images to descriptors. From these, appearance-based localization can be performed.

Thus, in Appearance-based Localization, we assume the existence of a certain *imaging function* Φ that defines a continuous map $I = \Phi(p)$ between images I and camera poses p. Then, the problem of localization consists of reverting such function $p_q = \Phi^{-1}(I_q)$ to regress the pose p_q of a certain query image I_q, given a map of the environment, that is, a set of image-pose pairs.

However, the extremely high dimensionality of the image space (e.g.: $\sim \mathbb{R}^{3*10^5}$ for 480×640 images) renders impossible to find this relationship between the pose and the image [7]. Consequently, it is necessary to apply some Dimensionality Reduction (DR) technique to the image vector in order to remove non-relevant information from it, thus making the problem more tractable. DR methods can be classified depending on whether the intrinsic space is considered linear or non-linear

https://doi.org/10.17979/spudc.9788497498418.0721

(also known as Manifold Learning (ML)). Optionally, we can apply some intermediate transformation before DR, generating an image holistic feature vector (typically through Deep Learning-based methods) which lies into a lower dimensional and less intricate space than that of images.

This work describes and analyzes the problem of DR in images in the context of AL, evaluating four different well-known DR techniques (linear and manifold-based) in terms of the isometry between their lower-dimensional image embeddings and the underlying camera pose space, also exploring the use of a holistic feature vector before applying DR.

The paper is structured as follows: in Section 2 the problem of Appearance-based Localization is stated. The fundamentals of DR in general, and ML in particular, are discussed in detail in Section 3. Sections 4 and 5 examine the effect that DR methodologies have in AL supported by a set of experiments and related works. Finally, a conclusion is given in Section 6.

2 APPEARANCE-BASED LOCALIZATION

Given an environment, let us define the *appearance map* $\mathcal{M} = \{\mathbf{P}, \mathbf{I}\}$ as the set of images $\mathbf{I}$ and the poses $\mathbf{P}$ they were captured at.

The *imaging function* Φ is the function that maps poses to images, which we assumed to be bijective, that is, two images taken at different poses must be also different. This way,

$$\Phi : \mathbf{P} \rightarrow \mathbf{I} \;\middle|\; I_i = \Phi(p_i) \text{ and } p_i = \Phi^{-1}(I_i). \quad (1)$$

As stated before, estimating the *localization function* Φ^{-1} from $\mathcal{M}$ is not possible mainly for the high dimensionality of the images, and some kind of DR must be performed.

For that, we explore the alternative of applying DR to holistic representations of the images by means of an *embedding function* $\Psi : \mathbf{I} \rightarrow \mathbf{D}$, obtaining descriptors D_i that lie in a lower dimensional space but where the essential information of the image still remains.

We claim that, if Ψ is an appropriate embedding function, the information needed to obtain the pose of the image is still present in the descriptor, that is, the function $(\Psi \circ \Phi)^{-1}$ can be estimated, and, not only that, the problem of AL becomes more tractable because of the reduction of dimensionality.

Consequently, the resulting framework (see Fig. 1) depends on two fundamental elements: i) the descriptors, or equivalently the embedding Ψ, and ii) the estimation of $(\Psi \circ \Phi)^{-1}$. In this work, we focus on the first element, studying, for different cases, how distinct embeddings maintain the underlying pose structure.

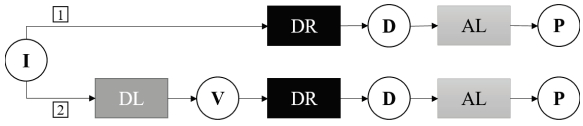


Figure 2: General scheme of performing DR and AL from 1) images ($\mathbf{I}$) and 2) DL-based image holistic feature vector ($\mathbf{V}$).

Finally, we also study the effect of the application of DR to Deep Learning (DL)-based description [1] of the images, which became prominent in recent years due to their visual invariance and abstraction capabilities. This way, our study explores two scenarios: applying DR to whole reduced images and applying DR to DL-based characterizations, depicted as $\mathbf{V}$ in Fig. 2.

3 DIMENSIONALITY REDUCTION

The situation where the dimension of the sample data points is comparable or even larger than the amount of existing samples is called *the curse of dimensionality*, and represents one of the main challenges of working with images for AL.

The general approach to handling this is to reduce the image dimensionality while trying to keep as much information as possible.

This leads to two sets of strategies denoted as *feature selection* and *feature extraction* (refer to Fig. 3). The first approach selects a representative subset of the data, such as keypoints, lines or planes [6] in the images, effectively reducing the images dimensionality by means of working with a selection of these local characteristics. On the other hand, *feature extraction* builds new features from the whole image by combining the original data through either linear or non-linear transformations. This work explores some of the most popular methods within this second category.

3.1 LINEAR DIMENSIONALITY REDUCTION

Principal Component Analysis (PCA) [13] or Multidimensional Scaling (MDS) [9] are well-known DR algorithms that successfully retrieve the underlying lower-dimensional representation of data provided it has a linear structure. In short, PCA looks for the linear combinations of the initial fea-

https://doi.org/10.17979/spudc.9788497498418.0721

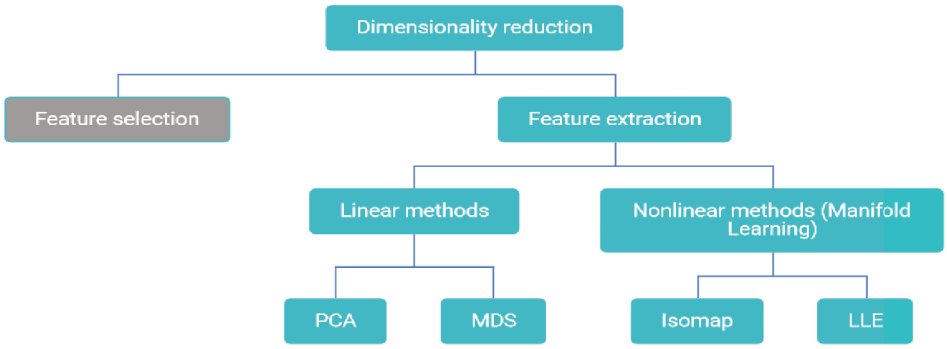


Figure 3: Classification of Dimensionality Reduction techniques

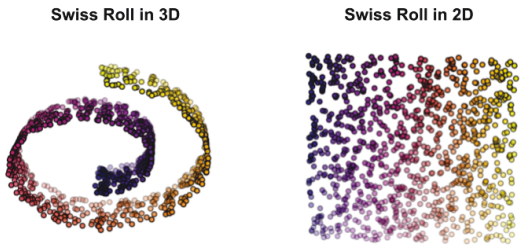


Figure 4: Example where the intrinsic dimensionality is recovered. Isomap proves to be able to unroll the Swiss Roll and to find its intrinsic geometry.

tures that explain the largest variance of the data. The objective of MDS, in turn, is to reduce the dimensionality of the input while preserving distances (typically Euclidean) between data points.

3.2 MANIFOLD LEARNING

In a nutshell, *Manifold Learning* (ML) [15] represents the set of techniques for non-linear DR grounded on the *manifold hypothesis* [17], which claims that high dimensional data tend to lie on low dimensional manifolds embedded in such high-dimensional spaces. Driven by this theory, ML methods try to recover the underlying geometry of the manifold that contains the data, which receives the name of *intrinsic space*, and which can thoroughly characterize the data by itself. The idea behind ML can be better understood visually with 3 dimensional manifolds (surfaces) which are embedded in the plane as it happens with the Swiss-Roll dataset illustrated in Fig. 4. Even though this figure lives in the three dimensional space, only two dimensions are needed to describe its structure.

In general, the dimensionality of the intrinsic space is unknown and, to this day, its retrieval has only been suitable for few particular datasets [19]. But, even though this is not yet feasible with most complex data, ML algorithms are still capable of finding more compact and tractable representations of the data while preserving important properties of the structure.

In this work, we will consider Isomap and Locally Linear Embedding (LLE) as representative methods within ML.

3.2.1 Isomap

Isomap [19] upgrades MDS by replacing Euclidean distances with *geodesics*, i.e. distances in the manifold. Geodesics are estimated from the known points of the manifold, first by constructing a weighted distance graph with K-Nearest-Neighbours (KNNs), and then by approximating the geodesic as the shortest path between the vertices in the graph. The geodesic distance between two nodes is the sum of the edge weights between them. Finally, MDS is applied to the estimated geodesics to get the lower-dimensional embedding.

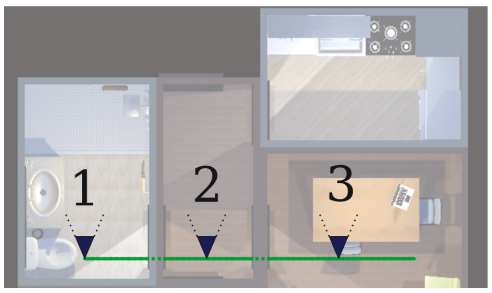
3.2.2 Locally Linear Embedding (LLE)

The idea behind the LLE method [16] is to exploit the fact that manifolds locally resemble Euclidean spaces. Based on this, LLE linearly reconstructs each point from a combination of its closest KNNs, and later the resulting weights are used to find the lower dimensional geometry which also minimizes the linear reconstruction error in the new space.

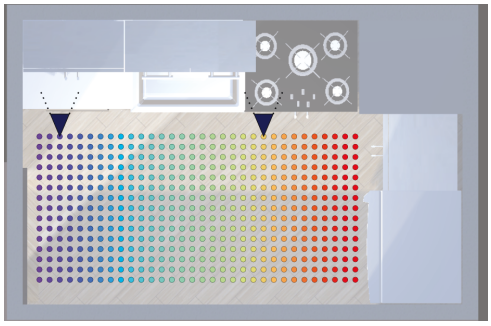
4 EXPERIMENTS

This section experimentally analyzes the performance of different linear and non-linear DR techniques for AL, measuring the isometry between the resulting lower dimensional spaces and the intrinsic pose space for each scenario.

https://doi.org/10.17979/spudc.9788497498418.0721



(a) Unidimensional case, where the camera trajectory is shown in green and each separated room is tagged with a number.



(b) Two-dimensional scenario, a position grid captured in the kitchen.

Figure 5: Set-ups where the experiments take place, both with the camera orientation depicted in blue.

4.1 Experimental setup

We have employed the synthetic dataset *Robot@VirtualHome* [5] to obtain the images for the experiments, concretely moving the camera in the *House 11*. Two different cases have been considered:

- **Set-up 1:** The camera visits three different rooms, moving transversely following a 3 m trajectory while gathering ~ 250 images (see Fig. 5a). This dataset captures the diverse appearances of the rooms and the occlusions of the passages between them.
- **Set-up 2:** The camera is placed within a unique room, capturing 480 images at a position grid of 1.5 m $\times$ 1 m with fixed orientation (see Fig. 5b). In this case, the whole image set has similar appearance, being more devoted to delve into local changes.

For both scenarios, two appearance representations of the images have been compared: (i) the grayscale images, taken as 1D vectors, and, (ii) using NetVLAD [1], a 4096-sized Visual Place Recognition learned descriptor. We have then embedded each of these characterizations in simpler spaces employing two linear DR methods (PCA and MDS) and two ML methods (Isomap and LLE). Each resulting embedding has been evaluated through the correlation coefficient between its retrieved geometry and the camera position.

4.2 SETUP 1: One-dimensional motion

As the camera moves along a single direction without rotation, we should evaluate a one-dimensional intrinsic space. We have compared how each DR descriptor recovers the original geometry both in *local* and *global* terms. In this context, the local perspective consists of recovering the intrinsic space of the camera within each independent room without overlap, thus maintaining a similar appearance. On the other hand, the global perspective considers the whole sequence, navigating along the full trajectory.

Table 1: Absolute value of correlation coefficients for the whole images.

	PCA	MDS	LLE	ISOMAP
Room 1	0.87	0.6	0.92	**0.96**
Room 2	0.3	0.27	0.93	**0.99**
Room 3	0.97	0.86	0.99	**0.998**
Globally	0.33	0.06	0.98	**0.996**

Table 2: Absolute value of correlation coefficients for the NetVLAD descriptor.

	PCA	MDS	LLE	ISOMAP
Room 1	0.94	0.27	0.98	**0.997**
Room 2	0.14	0.22	0.92	**0.998**
Room 3	0.96	0.23	0.68	**0.98**
Globally	0.35	0.29	0.37	**0.92**

The isometry between the retrieved low-dimensional descriptors and the camera positions using whole images as appearance representations is depicted in the Table 1, where values closer to 1 represent large correlations between the involved variables. As can be seen, ML algorithms prove to recover better the intrinsic geometry in all cases, while linear methods have more difficulty to find the relationship between images and poses, especially on the second room (which has more occlusions) and in the global case. This suggests that the relation between position and image becomes highly non-linear in complex environments. Table 2 compares the results after pre-processing the image with NetVLAD (whose last layer consists on a PCA transform). Such processing has no significant impact on PCA and Isomap, while it proves to be inconsistent with

https://doi.org/10.17979/spudc.9788497498418.0721

Table 3: Computation times measured in seconds for images (I) and appearance vectors (V).

	PCA		MDS		LLE		ISOMAP	
	I	V	I	V	I	V	I	V
Room 1	12.47	0.012	2.9	0.006	2.84	0.026	3.11	0.015
Room 2	17.846	0.02	5.22	0.02	4.72	0.06	4.91	0.02
Room 3	13.02	0.023	3.051	0.014	2.77	0.03	2.93	0.011
Globally	32.69	0.06	14.38	0.13	13.74	0.19	15.14	0.07

Table 4: Absolute value of correlation coefficients for the whole images.

	PCA		MDS		LLE		ISOMAP	
	Var 1	Var 2	Var 1	Var 2	Var 1	Var 2	Var 1	Var 2
Coord x	0.24	0.67	0	0.67	0.67	0.65	**0.97**	**0.12**
Coord y	0.04	0.34	0.08	0.24	0.14	0.4	**0.12**	**0.91**

the underlying idea of MDS and LLE.

On the other hand, Table 3 shows the computational cost of the Dimensionality Reduction expressed in seconds, comparing the performance between using whole images (I) and appearance vectors (V) for each case. As expected, extracting the underlying geometry from the image space takes longer than using appearance vectors, with the linear methods spending longer times for all cases, suggesting that non-linear methods are competitive in such scenarios.

In conclusion, ML techniques (and specifically, Isomap) demonstrate to outperform the remaining methods in identifying the relationship between images and poses for this setup.

4.3 SETUP 2: Bidimensional motion

In this scenario, the camera movement has two degrees of freedom (refer to Fig. 5b), therefore each method is evaluated through a correlation matrix. Now, the ideal situation occurs if the absolute values of the elements in one of the diagonals are equal to 1 and the remaining ones are zeros.

Table 4 shows that, when using whole images as the appearance representation, Isomap retrieves the underlying geometry with higher accuracy, while the remaining methods struggle with the non-linearities of the scene. This can also be seen in the first row of Fig. 6 where the structure of the Isomap result looks more similar to the original regular grid.

By contrast, the appearance space in which NetVLAD features lie demonstrates a more linear behaviour, since the performance of PCA and MDS improves substantially, while ML methods yield worse results as shown in Table 5. This is illustrated in the second row of Fig. 6, with MDS standing out among the rest. The computational cost in this case is higher due to the increased complexity (see Table 6), especially for linear methods with images. Concretely, in the case of NetVLAD, the weighting layer of this network proves to ease the performance of PCA.

5 RELATED WORK

Different DR methodologies [3, 7, 18, 10] have been previously suggested to address the task of Appearance-based Localization.

The study carried out in [3] is restricted to local environments where the appearance does not vary substantially. In this situation, PCA correctly identifies the latent geometry of the data, which allows the pose to be estimated through linear interpolation. This is coherent with the results obtained in our first setup, when the underlying geometry is retrieved for each individual room. However, linearity is proved to disappear in larger domains, as in the global case.

For this reason, some works propose variants of ML algorithms, as Locally Linear Projection (LLP) [7], an adaptation of LLE to Visual Localization which has been shown to outperform PCA in multiple-room environments, as demonstrated in our results. The authors of [18] propose an iterative model based on Isomap (the technique with best performance in our experiments), that seeks to minimize a stress function, accurately retrieving the geometry of a dense indoor environment with omnidimensional images. Finally, [10] faces Visual Localization in unidimensional image sequences through the Grassman-Stiefel Embedding (GSE), which grounds on the similarity between tangent spaces.

https://doi.org/10.17979/spudc.9788497498418.0721

Table 5: Absolute value of correlation coefficients for the NetVLAD descriptor.

	PCA		MDS		LLE		ISOMAP	
	Var 1	Var 2	Var 1	Var 2	Var 1	Var 2	Var 1	Var 2
Coord x	0.06	0.85	**0.91**	**0.25**	0	0.46	0.74	0.62
Coord y	0.95	0	**0.24**	**0.96**	0.47	0.56	0.56	0.75

Table 6: Computation times measured in seconds for images (I) and appearance vectors (V).

PCA		MDS		LLE		ISOMAP	
I	V	I	V	I	V	I	V
67.6	0.22	39.92	8.71	35.79	0.51	34.63	0.36

6 CONCLUSIONS

This work has addressed the application of Dimensionality Reduction (DR) methodologies for Appearance-based Localization (AL). We have stated the mathematical framework for the localization problem, as well as described some well-known linear and non-linear DR methods.

Two experiments have been presented, testing the performance in AL achieved by such methods in a synthetic environment. This performance is measured by applying DR to the images and measuring the isometry between the resulting geometry and the camera positions.

Linear DR techniques obtain good performance in local environments but are unable to generalize to multiple appearances at wider environments. However, employing a global appearance representation (e.g. NetVLAD) before applying DR improves the performance of linear methods. Non-linear methods, in turn, outperform linear ones in almost all scenarios, showing promising results for AL that could be exploited in future works. This demonstrates that the relationship between images and the poses where they were captured at is strongly non-linear.

Acknowledgement

This work is supported by the research projects *HOUNDBOT* (P20_01302), funding by Andalusian Regional Government, and *ARPEGGIO* (PID2020-117057GB-I00), funded by Spain National Research Agency.

References

[1] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J., (2016) NetVLAD: CNN architecture for weakly supervised place recognition, *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5297-5307

[2] Bernstein, A. (2017, February). Manifold learning in machine vision and robotics. *In 2016 International Conference on Robotics and Machine Vision* (Vol. 10253, pp. 81-86). SPIE.

[3] Crowley, J.L., Pourraz, F. (2001) Continuity properties of the appearance manifold for mobile robot position estimation, *Image and Vision Computing*, pp 741-752.

[4] Cummins, M., Newman, P. (2008) FAB-MAP: Probabilistic localization and mapping in the space of appearance, *The International Journal of Robotics Research*, pp 647-665.

[5] Fernandez-Chaves, D., Ruiz-Sarmiento, J., Petkov, N., Gonzalez-Jimenez, J. Robot virtualhome, an ecosystem of virtual environment tools for realistic indoor robotic simulation (2022). *To appear.*

[6] Gomez-Ojeda, R., Moreno, F. A., Zuniga-Noel, D., Scaramuzza, D., Gonzalez-Jimenez, J. (2019). PL-SLAM: A stereo SLAM system through the combination of points and line segments. *IEEE Transactions on Robotics*, 35(3), 734-746.

[7] Ham, J., Lin, Y., Lee, D.D., (2005) Learning nonlinear appearance manifolds for robot localization, *In Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp 2971-2976.

[8] Jaenal, A., Moreno, F.A., Gonzalez-Jimenez, J., (2021) Experimental Analysis of Appearance Maps as Descriptor Manifolds Approximations, *Computer Analysis of Images and Patterns*, pp 109-119.

[9] Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2), 115-129.

 https://doi.org/10.17979/spudc.9788497498418.0721

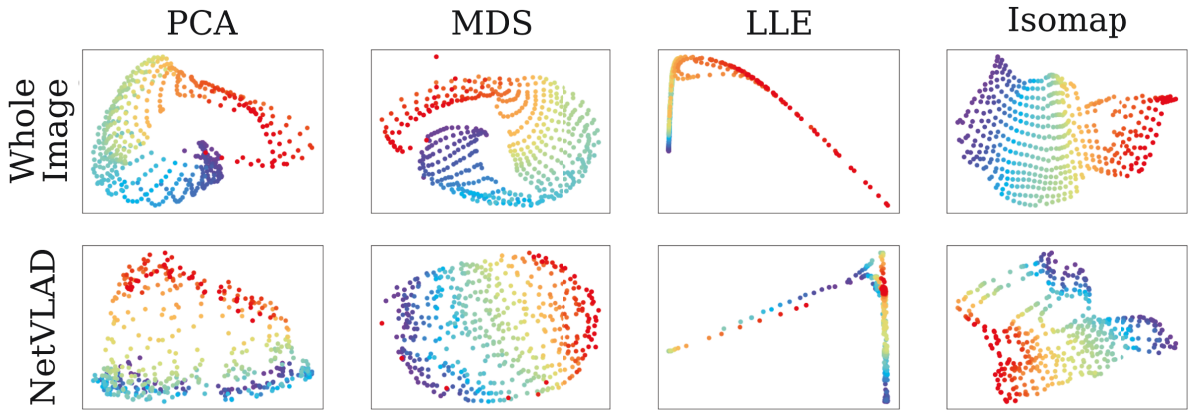


Figure 6: Embeddings for the bidimensional motion. The first row corresponds to the application of DR techniques (from left to right: PCA, MDS, LLE, Isomap) to the space of corrected images. The second row corresponds to the application of the same techniques to the space of NetVLAD descriptors. The colors represent the camera poses illustrated in Fig. 5b.

[10] Kuleshov, A., Bernstein, A., Burnaev, E., Yanovich, Y. (2017, December). Machine learning in appearance-based robot self-localization. *In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE.* pp. 106-112

[11] Masone, C., Caputo, B. (2021) A Survey on Deep Visual Place Recognition, *IEEE Access*, vol. 9, pp. 19516-19547.

[12] Mur-Artal, R., Montiel, J. M. M., Tardos, J. D. (2015). ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE transactions on robotics*, 31(5), pp 1147-1163.

[13] Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11), 559-572.

[14] Piasco, N., Sidibe, D., Demonceaux, C., Gouet-Brunet, V. (2018). A survey on visual-based localization: On the benefit of heterogeneous data. *Pattern Recognition*, 74, 90-109.

[15] Pless, R., Souvenir,R. (2009) A survey of Manifold Learning for images, *IPSJ Transactions on Computer Vision and Applications*, pp 83-94.

[16] Roweis, S.T., Saul, L.K. (2000) Nonlinear Dimensionality Reduction by locally linear embedding, *Science*, pp 2323-2326.

[17] Seung, H.S., Lee, D.D. (2000) The Manifold Ways of Perception, *Science*, pp 2268-2269.

[18] Schwartz, A., Talmon, R. (2018) Intrinsic Isometric Manifold Learning with Application to Localization. *SIAM Journal on Imaging Sciences*,12(3), pp 1347-1391

[19] Tenenbaum, J.B., De Silva, V., Langford, J.C. (2000) A global geometric framework for nonlinear Dimensionality Reduction, *Science*, pp 2319-2323.

© 2022 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution CC-BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by-nc-sa/4.0/deed.es).