Proceedings of the INMM & ESARDA Joint Virtual Annual Meeting
August 23-26 & August 30-September 1, 2021

DEEP LEARNING FOR NUCLEAR SAFEGUARDS

Erik Wolfart
European Commission - Joint Research Centre, Ispra, Italy
erik.wolfart@ec.europa.eu

Carlos Sanchez-Belenguer
European Commission - Joint Research Centre, Ispra, Italy
carlos.sanchez-belenguer@ec.europa.eu

Vitor Sequeira
European Commission - Joint Research Centre, Ispra, Italy
vitor.sequeira@ec.europa.eu

ABSTRACT

Nuclear inspectors benefit from location-based services in several ways: i) to independently verify the current position and navigate within nuclear facilities; ii) to access to information that is contextual to the current position, for example to measurements, notes and observations acquired during previous inspections; iii) to carry out specific tasks according to the current location. Furthermore, the inspector can tag measurements and observations taken during the inspection with the current position to facilitate later analysis of the data and increase the efficiency of follow-up inspections. Current mobile devices typically use GPS to acquire position information, i.e. they cannot be used for inspection activities inside nuclear facilities. This paper presents an indoor localization system developed for nuclear safeguards applications, which recognizes the current location using visual information. In an off-line mapping phase, we use a 3D laser sensor mounted on a backpack with a calibrated spherical camera i) to generate the data for training a deep neural network and ii) to build a database of georeferenced images for an environment. Thanks to the 3D laser measurements and the spherical panoramas, we can efficiently survey large indoor areas in a very short time. The underlying 3D data allows us to identify images observing the same place and effectively train a deep neural network that maps an image to a signature, which is representative for the given location. During the online localization phase, the inspector acquires an image and queries the trained network to efficiently retrieve the location of the most similar signature in the database of georeferenced images. The paper presents the architecture of the underlying neural network and shows how the concept can be applied to other applications in nuclear safeguards, e.g., for verifying spent nuclear fuel using gamma emission tomography. In this case, the network can be trained to map the sinogram generated by the tomograph to a unique signature, which is representative for the given fuel assembly layout and which can then be used to verify the operator declaration.

INTRODUCTION

Recognizing places from visual information is a well-known problem that has been present in the literature for a long time. Recently, visual place recognition has gained increasing attention due to the amount of geo-localized image datasets [1-4], the increase of portable acquisition devices (i.e. mobile phones) and the limitations of GPS localization systems in indoors and urban environments.

Visual place recognition is a challenging problem due to three major facts: (1) the appearance of a place can change drastically over time, (2) places are not always observed from the same viewing point and (3) cameras are strongly affected by light conditions. It can be used for a wide range of applications like loop detection in SLAM approaches [5] or autonomous navigation [6].

Traditional approaches facing this topic relied either on global methods [7], which process the image as a whole using global descriptors, like GIST [8], or on local methods that extract selectively parts of the image using local-invariant feature extractors (such as SURF [9], or SIFT [10]) and match them using Bag-Of-Words [11] or voting schemes. Even though both methods differ in the way they approach the problem, they share a common feature: descriptors are always hand-crafted.

In the last years Deep Learning has revolutionized many disciplines. Particularly, in the Computer Vision field, Convolutional Neural Networks (CNNs) have achieved state-of-the-art performance on several recognition and classification tasks. The key idea behind CNNs is their ability to automatically learn high-level global features, in contrast with hand-crafted ones. Similarly to the rest of machine learning approaches, CNN-based place recognizers require large sets of training data to perform properly. For this reason, researchers working in this field dedicate lots of resources to the data collection stage [12, 13].

Works using CNNs for solving place recognition problems have mostly focused their efforts on urban outdoor environments [3, 14-16]. The reason for that is related to the availability of training data: acquiring outdoor georeferenced training images can be easily achieved with a consumer GPS-enabled camera. However, the number of available public datasets for visual indoor localization is significantly smaller. Existing datasets can be classified based on the way data is acquired: (1) manual approaches like [12, 13], (2) semi-manual approaches like [17, 18], (3) datasets acquired with dedicated infrastructure like [19, 20] and (4) automatic approaches relying on RGB-D sensors like [21]. The first two categories suffer from low performance and are more error-prone. The other two categories are limited in terms of covered area and usually work at room scale. Only few datasets are available at building scale covering several rooms or floors [17, 18, 22, 23]. They are generated from 3D scanners, simultaneously acquiring geometry and image data.

METHOD

Our localization technique uses a 3D map of the environment and associated images with their known position that are acquired in an a-priori mapping phase with a 3D scanning backpack. During the localization phase, a simple stereo camera is used to identify the most similar images in the a-priori map and then compute the current pose. In this section we start by introducing our data acquisition platform that allows us to automatically generate accurate 3D maps of indoor spaces and to collect thousands of georeferenced images. Then, we introduce our CNN-based place recognizer that is trained using these data and, finally, we present our pose refiner that transforms the coarse poses reported by the place recognizer into accurate estimations of the user's position inside the building.

Acquisition of training data

Our data collection procedure is based on the MLSP (Mobile Laser Scanning Platform) backpack: a portable system equipped with two LiDAR sensors and an inertial unit that can work in two modes: mapping and tracking. When working in mapping mode the system produces a high resolution point cloud [24] with no drift [25]. When working in tracking mode, the system estimates the pose of the laser head within the reference point cloud with centimeter accuracy [26]. The MLSP is an in-house development of the EC-JRC and was awarded with the first place in the 2015 Microsoft Indoor Localization Competition and was used for refereeing in the 2016, 2017 and 2018 editions. It has many commercial applications for surveying and mapping; in the nuclear safeguards domain it is used for DIV inspections, for example in the geological repository for final disposal currently under construction in Finland [27].

We use the backpack in mapping mode to create the 3D point clouds of the buildings. Each point cloud defines the reference frame, the *map*, in which all data will be localized. Then, we revisit the building in tracking mode while acquiring continuous sequences of image data under different conditions (e.g. different lighting, appearance changes or different paths). The trajectory estimated by the backpack is used to locate the images. In this sense, the modified MLSP can be seen as an accurate indoor GPS with calibrated sensors around [28].

As Figure 1 shows, the sensors mounted during our data collection are:

- *Spherical camera*: a Garmin Virb 360 camera directly connected to the backpack on-board computer, acquiring 1080p images (1920x1080 pixels) at 15 Hz in an equirectangular format (stitched on-board with a predefined LUT) and compressed as a sequence of JPEG images.

- *Stereo camera*: a wide-angle field of view ZED 2 camera from Stereolabs, connected to an external nVidia Jetson TX2 embedded computer, acquiring 720p stereo pairs (2x1280x720 pixels) at around 15 Hz (varying framerate depending on the compression time) stored in a lossless format. Additionally, the camera is equipped with accelerometers, gyros, a barometer, a magnetometer and a temperature sensor whose readings are also recorded.
- *Smartphone*: an Asus ZenFone AR with Android 7.0 from which we acquire the available WiFi access points and their signal intensity, the phone cell towers in range, the magnetic field intensity and direction, the data from the gyros and accelerometers, the noise level, ambient temperature, air pressure and the ambient light conditions. The smartphone can provide auxillary information to further improve the localization; it is however not used in the current work.

The spherical camera provides an excellent 360 degree coverage for mapping and fingerprinting the environment. On the other side, the final localization typically works with regular cameras. We chose the ZED 2 because it also provides a stereo image and out-of-the-box visual odometry.

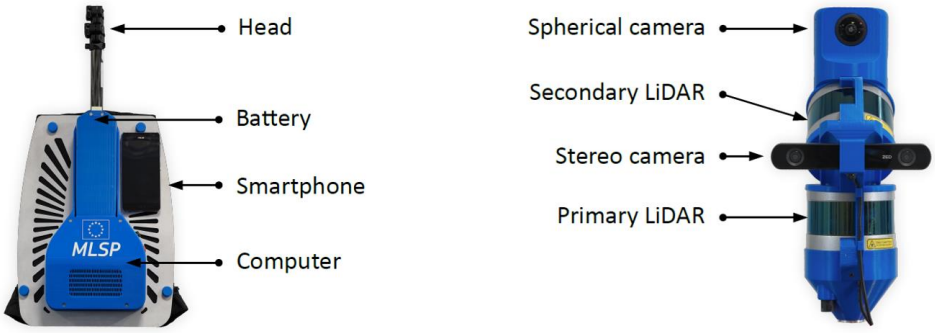

Figure 1. MLSP backpack used for data acquisition. (left) main body of the backpack with the smartphone. (right) Sensors head mounted on the top of the backpack.

CNN for place recognition

We propose a deep CNN that maps images into embeddings [29] (i.e. vectors of numbers that provide a compact representation of the original image). The goal of the training is to compute the weights of the neural network that satisfy one single condition: when two images are observing the same place their associated embeddings have to be very similar (i.e. small Euclidean distance). When two images are observing different places this distance has to be large.

In this sense, our CNN learns a similarity metric that enables deciding whether two *sensor readings* in the same domain are observing the same *scene* or not. As the last section will show, this architecture can be extended to many other verification domains, where the sensors do not need to be cameras, do not need to be identical (as long as the output data is in the same domain), the scene can be anything that the sensor can perceive and the similarity metric can be arbitrarily defined based on the problem to be solved. The task of the CNN is to learn which parts of the input data allow deciding how similar the two input samples are.

Given two images, I_1 and I_2, and their associated embeddings, e_1 and e_2, respectively, we define the dissimilitude between the two images $d(I_1, I_2)$ as:

$$d(I_1, I_2) = \|e_1 - e_2\| \#(1)$$

During the training stage we aim to learn the weights that minimize the following error function:

$$E = \left[d(I_1, I_2) - \left(1 - \text{overlap}(I_1, I_2)\right)\right]^2 \#(2)$$

where $\text{overlap}(I_1, I_2)$ represents the number of pixels observed in one image that are present in the other and vice versa. Notice that this ratio can be computed analytically exploiting the data provided by our acquisition pipeline: for each training image we have the camera projection parameters, its pose inside the environment

and an accurate 3D model of the building. Combining all this information we can estimate automatically the real geometric overlap between each pair of images used for training the neural network. This enables us to evaluate equation (2) and, thus, to back-propagate the prediction errors during the training.

To train the CNN we proceed as illustrated in Figure 2: we define two instances of the same CNN in a siamese configuration (forcing weights to be equal in both instances). Then, we feed the system with two different training images and compute the estimated dissimilarity between their embeddings using (1) (left side of Figure 2). Simultaneously, we compute the ground truth overlap between both images using the training 3D data (right side of Figure 2). Finally, we evaluate the error function using (2) and back-propagate to optimize the weights of both instances of the CNN simultaneously.

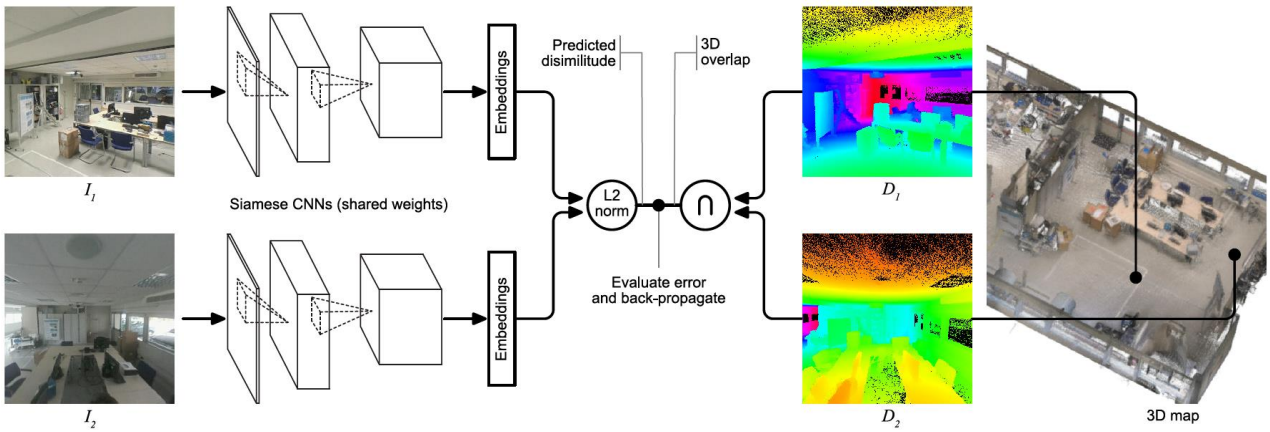

Figure 2. CNN training. Given two images (I_1 and I_2 on the left), their 3D poses and projection parameters and a 3D map (right), we compute the 3D position of each pixel by re-projecting the map into both image planes (D_1 and D_2, respectively). Then, we compute the ratio of overlapping points (*3D overlap*). The L2 norm of the two embeddings (*predicted dissimilitude*) has to be equal to (*1 - overlap*).

Once the CNN is trained, to use it for place recognition purposes, we populate a database with the set of georeferenced images. To do so, instead of storing all pixel intensities, we feed the images into the trained CNN and compute their associated embeddings. Each entry of the database consists on a pair of vectors $\langle e_i, \Gamma_i \rangle$, where e_i is the embedding and Γ_i is the associated pose.

The localization phase exploits the pose information stored in the database and requires only a simple video camera during execution: as the camera moves inside the environment, new images are acquired with no pose information. For each of them the corresponding embedding is calculated using the pre-trained CNN. To retrieve similar images from the database (e.g. observing the same place) and their associated poses in the map a nearest neighbor radius search is performed in the high-dimensional embedding space. This is efficiently resolved by using a kd-tree pre-allocated with all the embeddings present in the database.

Since the computed pose is based on the nearest images in the database, which might have been acquired at some distance from the current position, it is only an approximate result. To achieve the accuracy required for Augmented Reality applications, a pose refinement is carried out as described in the next section.

Pose refinement

The pose refinement component of our system takes as input a coarse estimation of the user's pose inside the environment and returns a more accurate estimation. This operation is especially important when building AR applications: without an accurate positioning system the blending between the real world and synthetic data does not work.

Pose refinement is necessary in our system in two scenarios: (1) after executing the place recognizer, since the returned pose is never the exact position of the camera and (2) as the user moves inside the environment. For this second case it is important to remark that, once the place recognizer has identified the pose of the

user inside the map, it is no longer necessary to further execute it: the pose is tracked in real-time using visual odometry and the output is periodically refined to keep it in the map's reference frame.

The pose refinement operation is performed using classic Computer Vision techniques and exploiting the 3D data of the database: images acquired with the MLSP backpack can benefit of depth information by re-projecting the map into the image plane (as shown in Figure 2-right). Given one image from the live camera and a coarse estimation of the user's pose, we retrieve from the database the closest image. For both images, we detect visual features and match them based on similarity. Exploiting these matches and the 3D information from the database, the camera pose is calculated by simple triangulation (solving the well-known PnP problem). In our case, we rely on the BRISK keypoint detector and descriptor extractor, combined with a brute-force feature matcher and a RANSAC inlier selection strategy. This process is illustrated in Figure 3.

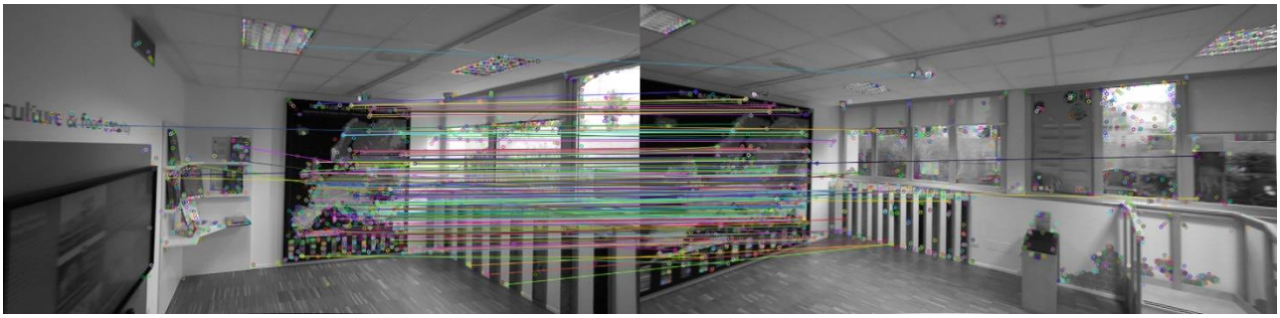

Figure 3. Pose Refinement. Given two images, one from the live camera (left) and another one from the database (right), we detect keypoints in both (colored points), match them based on their descriptor similarity and select only valid matches (colored lines) using a RANSAC approach. The 3D data available in the image from the database is exploited to triangulate the location of the live camera.

RESULTS

To train and evaluate the performance of our system, we generated a new public dataset for indoor localization: RISEdb [28]. It covers a large variety of indoor spaces, including office buildings, a workshop, a restaurant, an exhibition area and a conference building. For each building we provide a high-resolution and accurate 3D point cloud. Together with the reference map, we provide spherical and stereo sequences with a reliable ground-truth information and multiple sensor readings from a smartphone. The three main benefits from our dataset are: (1) instead of providing still images we offer full sequences that allow exploiting the temporal nature of the data, (2) the area we can cover with our acquisition platform is unconstrained, so the buildings we mapped are extremely large (more than 100m side and 4 floors in the case of the office building) and (3) thanks to the spherical camera we can provide a full coverage of the area mapped, which has proven to be very effective for machine learning approaches. Figure 4 shows some 3d models and images from our dataset.

Figure 4. The RISEdb dataset: (top) two of the 3D point clouds (workshop and exhibition area). (bottom) geo-referenced training images capturing the same environment under different conditions.

The dataset contains 30 sequences, covering the 5 described buildings for an acquisition time of more than 6 hours and a walked distance of 20.7km. This yields to more than 1 million geo-referenced images considering both panoramas and stereo pairs. Table I shows the details of the dataset, where the first row is the number of sequences acquired per building (i.e. number of times that the building has been re-visited under different conditions), the second row is the total acquisition time, the third is the total length walked and the fourth and fifth rows are the number of images acquired (panoramas and stereo pairs, respectively).

	office	conference	workshop	exhibition	restaurant
Sequences (#)	6	4	5	8	7
Time	2:16:11	45:27	55:21	1:20:16	1:00:14
Length (m)	8,727	2,233	2,690	3,896	3,142
Spherical (#)	117,144	40,860	49,775	72,999	54,084
Stereo pairs (#)	145,475	52,719	59,097	68,048	49,983

Table I. Dataset details.

Using this data, we trained and evaluated the performance of our place recognizer under different circumstances. We also compared it against RelocNet [30], a state-of-the-art technique that improved previous approaches like PoseNet [31] and its subsequent geometric extensions [32].

As results show (Figure 5-left), our place recognizer proved to be robust against light changes, appearance changes, outliers and different points of view. Execution times using the GPU enable real-time applications, with an average response time of 5 milliseconds, and the degree of accuracy achieved allows to effectively estimate the pose of the user within the building with a few observations.

To perform the comparison against RelocNet, we trained both systems for the "exhibition" building with the images acquired in 3 different sequences and used a fourth one for evaluation. The training of each system was executed on an nVidia GeForce GTX 1070 GPU during 48 hours using the almost 4 million training samples that the 3 sequences generated. As Figure 5-right shows, when querying RelocNet with a novel image, the first result returned is on average 89 cm away from the real pose of the camera and, as we retrieve more results, the positional error increases much faster than the same system trained using our technique. Our approach had an average error of 64 cm for the first result.

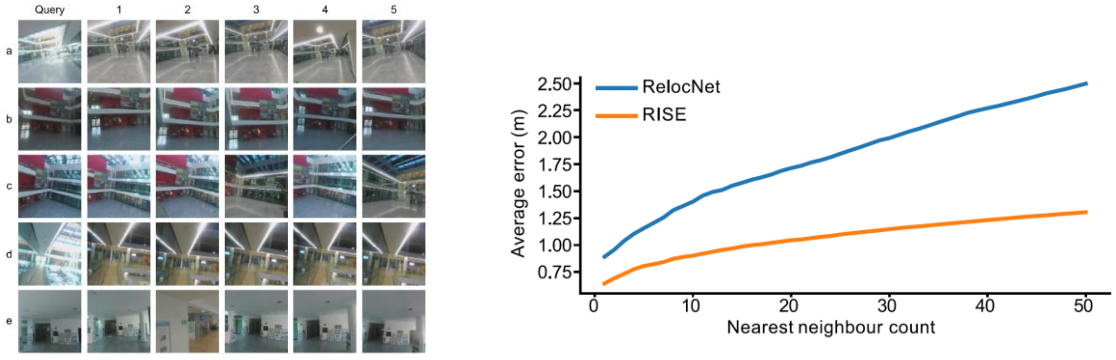

Figure 5. Place recognizer results. (left) Given a query image (first column), best 5 results returned from the database (columns 2 to 6). (right) Comparison between our approach (RISE) and RelocNet. Average positional error in the pose estimation w.r.t. the number of images retrieved from the database.

Figure 6 shows the final results achieved by our system after enabling the pose refinement. The yellow trajectory represents the ground truth as reported by the MLSP backpack and the green one is the output of our system. Notice how, during the initialization, the place recognizer quickly identified the place where the user was. After that, the visual odometer and the pose refiner were able to successfully track in real time the pose of the user as he moved inside the building, only deviating in the right side of the image, where the

training data was deliberately removed to show the magnitude of the odometer's drift. Once the user entered again into the mapped area the pose refiner took less than a second in compensating the accumulated drift.

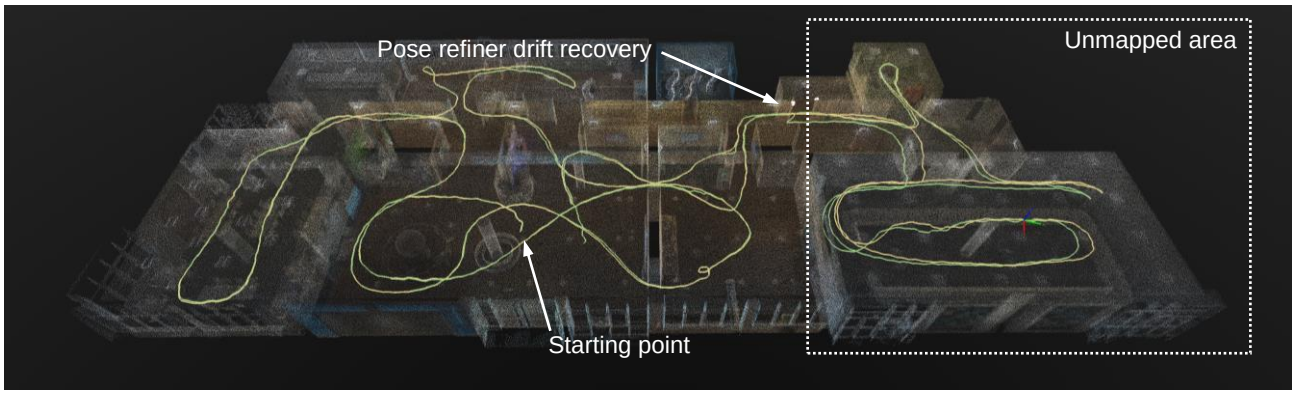

Figure 6. Pose estimation results. Comparison between the ground truth trajectory reported by the MLSP backpack (yellow line) and the trajectory estimated by our technique using only the camera (green line). Notice that both trajectories are almost identical except on the right side of the image, where the training data was deliberately removed to illustrate the drift of the visual odometry without pose refinement.

OTHER SAFEGUARDS APPLICATIONS

The CNN proposed for solving the place recognition problem can have other applications in the nuclear safeguards field: during the training stage it *only* learns how to compare two inputs in the same domain according to a user-defined similarity metric. This makes it especially attractive for verification problems, where nuclear inspectors need to decide whether if a measurement taken from their instruments matches the operator's declaration or not. To illustrate this idea, we implemented a proof-of-concept for spent fuel verification using the Passive Gamma Emission Tomograph (PGET) tool.

In this scenario, given a Spent Fuel Assembly (SFA) and the operator's declaration about its contents (in terms of internal layout and fuel pin composition), nuclear inspectors need to verify the correctness of the declaration. However, due to operational and safety constraints, this verification needs to be performed over complete assemblies and using non-destructive techniques. To address this problem, several stakeholders of the safeguards community jointly developed the PGET tool, which the IAEA approved for safeguards use in December 2017. It uses two opposing arrays of gamma detectors that measure the gamma radiation emitted by the fuel pins. The arrays rotate around the Spent Fuel Assembly (SFA) step-by-step, generating a 2D sinogram. The gamma radiation is measured in several broad energy windows in order to identify different fission products in the SFA, each window generating its own sinogram.

Current approaches for solving this problem rely on classic tomographic filtered back-projection algorithms to perform the reconstruction of each individual sinogram and a recommendation system that helps nuclear inspectors to manually compare the reconstructed layout wrt the operator's declaration (see Figure 7-top for details). This traditional approach has some drawbacks since it is error-prone (manual comparison between reconstruction/declaration), the tomographic filtered back-projection algorithms do not consider inter-pin attenuations (which are very strong in the case of spent fuel), it is a lossy process (sinograms need to be de-noised and the reconstruction algorithm also degrades the data) and does not benefit from a joint analysis of the four energy windows available.

All these issues together put limitations on the application of the PGET tool: large assemblies with complex layouts (e.g. experimental fuel) cannot be reliably validated and, additionally, no quantitative analysis can be performed at pin-level.

We propose a new pipeline supported by our CNN architecture that can perform automatic comparisons between the raw sinograms and the operator's declaration. To achieve such a system, we must first address

the cross-domain issue in the input data: the raw sinograms and the operator's declaration are different representations of the same assembly. We face this problem by introducing a simulator that takes as input the declared layout and outputs the expected sinogram that the SFA would produce. By doing so, we not only solve the cross-domain issue, but we also simplify considerably the learning requirements for the CNN: the simulator copes with most of the heavy payload of the problem, addressing its physics nature and dealing with the gamma transport equations and sensor physical properties. This way, the neural network only needs to learn how to compare simulated data with real sensor data (i.e. how to cope with simulation imperfections and how to filter out the noise produced by the PGET tool). Figure 7-bottom illustrates our proposed pipeline.

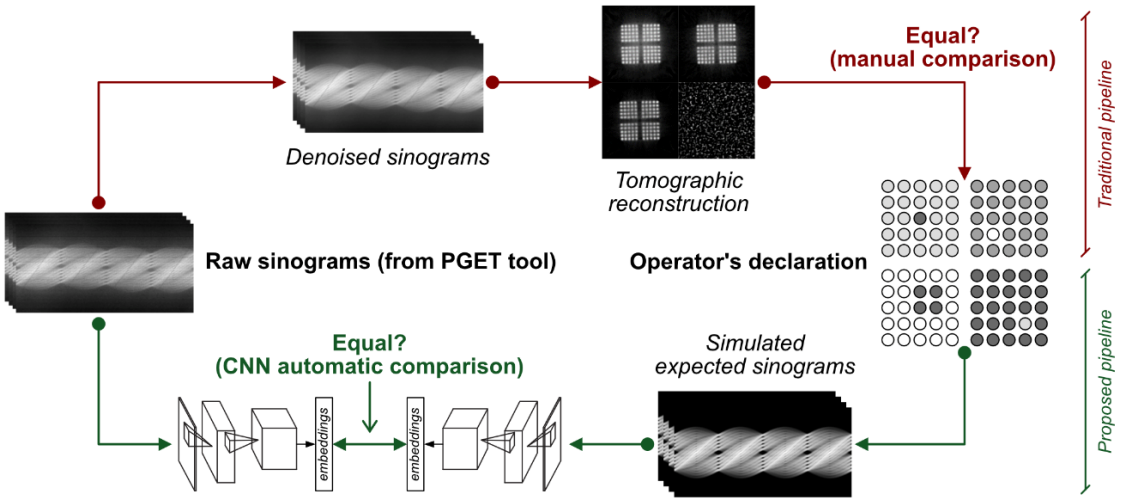

Figure 7. SFA PGET verification pipelines. Given a set of raw sinograms (left) and the operator's declaration (right), current approaches (top) follow the red path for verification. We propose an automatic approach based on our CNN (green-bottom) that exploits the similarity metric learnt by the neural network during the training.

As Figure 7 shows, the benefits from our proposed pipeline cope with all the previously enumerated limitations of traditional approaches: the comparison is no longer manual, inter-pin attenuations are considered by the simulator, raw sinograms are directly inputted to the CNN so there is no data loss and all four energy windows are jointly analyzed by the neural network in order to exploit potential correlations between different levels.

An additional benefit of this approach for nuclear applications is related to the training data itself: most of the training of the system can be performed using only synthetic samples generated by simulation. This way, in the first training stage, the CNN can learn about the structure of the data and where to *put attention* for finding similitudes/differences between two sinograms. Once the general structure of the problem has been learnt, only a few real samples should be sufficient for fine-tuning the internal weights of the CNN in order to be able to generalize what was learnt before to samples that suffer from sensor/background noise and that do not have simulation errors. In this sense, this architecture can be trained using a *few-shot learning* strategy, which is extremely convenient when real data is scarce and hard to access (like nuclear safeguards data).

CONCLUSIONS AND OUTLOOK

In this paper we have presented a full system for indoor localization that relies only on cameras. Results have shown that the degree of accuracy achieved and the execution time of our pipeline enables augmented reality applications, where responsiveness and precision are mandatory. These kind of applications can be directly applied to the nuclear safeguards field, where situational awareness tools can be developed for supporting nuclear inspectors as they move inside large facilities.

Our localization technique is based on two main building blocks: (1) a CNN-based place recognizer that can estimate the coarse pose of the user inside the environment with a single observation and (2) a pose refiner that allows achieving the degree of accuracy required for AR applications. Both systems require a reference map and some a-priori knowledge of the environment in the form of geo-localized images with depth information. To efficiently create this map and to collect the training images for the place recognizer we have introduced an automatic acquisition platform based on a 3D laser backpack. This system has allowed us to map very large environments in an almost-automatic way and to collect thousands of georeferenced images that have been published in a public dataset called RISEdb.

The CNN architecture introduced in this paper for place recognition exploits the 3D data and the georeferenced images to learn a distance function that allows deciding up to which degree two images are observing the same place. As results shows, it improves the accuracy of the state-of-the-art techniques and provides robust results under heavy changes in the environment due to different lighting conditions and appearance changes.

Finally, we have shown the potential of our CNN architecture in a complete different domain: spent fuel assembly verification. The natural ability of neural networks to work with raw high-dimensional data, together with the capability of our system of learning a distance function makes it ideal for verification tasks. Also, the fact that our CNN can be trained using a few-shot learning strategy makes it a perfect candidate for nuclear applications, where training data is scarce and hard to access.

Future developments include the deployment of the localization system in a commercial Augmented Reality headset and further research on the PGET data to demonstrate the potential of deep learning on tomography real data.

REFERENCES

[1] M. T. Islam, C. Greenwell, R. Souvenir, and N. Jacobs, "Large-Scale Geo-Facial Image Analysis," EURASIP Journal on Image and Video Processing (JIVP), vol. 2015, no. 1, p. 14, 2015.

[2] T. Weyand, I. Kostrikov, and J. Philbin, "Planet - photo geolocation with convolutional neural networks," in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 37–55.

[3] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[4] N. Sunderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons," Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA), p. 2013, 01 2013.

[5] X. Zhang, Y. Su, and X. Zhu, "Loop closure detection for visual slam systems using convolutional neural network," in 2017 23rd International Conference on Automation and Computing (ICAC), Sep. 2017, pp. 1–6.

[6] Y. N. Kim, D. W. Ko, and I. H. Suh, "Visual navigation using place recognition with visual line words," in 2014 11th Int. Conf. on Ubiquitous Robots and Ambient Intelligence (URAI), Nov 2014, pp. 676–676.

[7] P. Taddei, C. Sanchez, A. L. Rodrıguez, S. Ceriani, and V. Sequeira, "Detecting ambiguity in localization problems using depth sensors," in 3DV, 2014.

[8] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," Progress in brain research, vol. 155, pp. 23–36, 2006.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," Computer vision and image understanding, vol. 110, no. 3, pp. 346–359, 2008.

[10] D. G. Lowe, "Object recognition from local scale-invariant features," in Computer Vision, IEEE International Conference on, vol. 2. Los Alamitos, CA, USA: IEEE Computer Society, 1999, p. 1150.

[11] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in Proceedings Ninth IEEE International Conference on Computer Vision, 2003, pp. 1470-1477 vol.2.

[12] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database,” in Proc. of the 27th Int. Conf. on Neural Information Processing Systems - Volume 1, ser. NIPS’14. Cambridge, MA, USA: MIT Press, 2014, pp. 487–495.
[13] Z. Liu, L. Zhang, Q. Liu, Y. Yin, L. Cheng, and R. Zimmermann, “Fusion of magnetic and visual sensors for indoor localization: Infrastructure-free and more effective,” IEEE Transactions on Multimedia, vol. 19, no. 4, pp. 874–888, April 2017.
[14] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, “End-to-end learning of deep visual representations for image retrieval,” Int. J. of Computer Vision, vol. 124, no. 2, pp. 237–254, 2017.
[15] F. Radenovic, G. Tolias, and O. Chum, “Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples,” in European conf. on computer vision. Springer, 2016, pp. 3–20.
[16] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla, “Benchmarking 6dof outdoor visual localization in changing conditions,” in 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2018, pp. 8601–8610.
[17] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, “InLoc: Indoor visual localization with dense matching and view synthesis,” in CVPR, 2018.
[18] X. Sun, Y. Xie, P. Luo, and L. Wang, “A dataset for benchmarking image-based localization,” in 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5641–5649.
[19] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, “The euroc micro aerial vehicle datasets,” The Int. J. of Robotics Research, vol. 35, no. 10, pp. 1157–1163, 2016.
[20] C. Loffler, S. Riechel, J. Fischer, and C. Mutschler, “Evaluation criteria for inside-out indoor positioning systems based on machine learning,” 09/2018.
[21] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, “Scene coordinate regression forests for camera relocalization in rgb-d images,” in CVPR. IEEE, June 2013.
[22] F. Walch, C. Hazırbas¸, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, “Image-based localization using lstms for structured feature correlation,” 10 2017, pp. 627–637.
[23] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, “Matterport3d: Learning from rgb-d data in indoor environments,” in 2017 International Conference on 3D Vision (3DV), Oct 2017, pp. 667–676.
[24] S. Ceriani, C. Sanchez-Belenguer, P. Taddei, E. Wolfart, and V. Sequeira, “Pose interpolation slam for large maps using moving 3d sensors,” in Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on, Sept 2015, pp. 750–757.
[25] C. Sanchez-Belenguer, S. Ceriani, P. Taddei, E. Wolfart, and V. Sequeira, “Global matching of point clouds for scan registration and loop detection,” Robotics and Autonomous Systems, vol. 123, 10/2019.
[26] C. Sanchez-Belenguer, P. Taddei, S. Ceriani, E. Wolfart, and V. Sequeira, “Localization and tracking in known large environments using portable real-time 3d sensors,” Computer Vision and Image Understanding, vol. 149, no. C, pp. 197–208, Aug. 2016.
[27] Wolfart E; Ceriani S; Puig Alcoriza D; Sanchez Belenguer C; Taddei P; Sequeira V; Murtezi M; Turzak P; Zein A; Enkhjin L; Ingegneni M; Rocchi S; Yudin Y. Mobile 3D Laser Scanning for Nuclear Safeguards. Luxembourg (Luxembourg): Publications Office of the European Union; 2015. p. 71-/ 81. JRC104573
[28] C. Sanchez-Belenguer, E. Wolfart, A. Casado-Coscolla and V. Sequeira, "RISEdb: a Novel Indoor Localization Dataset," 25th Int. Conf. on Pattern Recognition (ICPR), 2021, pp. 9514-9521.
[29] C. Sanchez-Belenguer, E. Wolfart, and V. Sequeira, “Rise: A novel indoor visual place recogniser,” in 2020 IEEE Int. Conf. on Robotics and Automation (ICRA), 2020, pp. 265–271.
[30] V. Balntas, S. Li, and V. A. Prisacariu, “Relocnet: Continuous metric learning relocalisation using neural nets,” in The European Conference on Computer Vision (ECCV), September 2018.
[31] A. Kendall, M. Grimes, and R. Cipolla, “Posenet: A convolutional network for real-time 6-dof camera relocalization,” in Proceedings of the IEEE Int. Conf. on computer vision, 2015, pp. 2938–2946.
[32] A. Kendall and R. Cipolla, “Geometric loss functions for camera pose regression with deep learning,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6555–6564.