ARCHITECTURAL HERITAGE RECOGNITION IN HISTORICAL FILM FOOTAGE USING NEURAL NETWORKS

F.Condorelli [1,*], F. Rinaudo [1], F. Salvadore [2], S. Tagliaventi [2]
[1] DAD, Department of Architecture and Design, Politecnico di Torino, Lab G4CH – Laboratory of Geomatics for Cultural Heritage
(francesca.condorelli, fulvio.rinaudo)@polito.it
[2] CINECA – HPC Department
(f.salvadore, s.tagliaventi)@cineca.it

Commission II, WG II/8

KEY WORDS: Deep Learning, Neural Networks, TensorFlow, Photogrammetric Workflow, Cultural Heritage, Historical Video Classification

ABSTRACT:

Researching historical archives for material suitable for photogrammetry is essential for the documentation and 3D reconstruction of Cultural Heritage, especially when this heritage has been lost or transformed over time. This research presents an innovative workflow which combines the photogrammetric procedure with Machine Learning for the processing of historical film footage. A Neural Network is trained to automatically detect frames in which architectural heritage appears. These frames are subsequently processed using photogrammetry and finally the resulting model is assessed for metric quality. This paper proposes best practises in training and validation on a Cultural Heritage asset. The algorithm was tested through a case study of the Tour Saint Jacques in Paris for which an entirely new dataset was created. The findings are encouraging both in terms of saving human effort and of improvement of the photogrammetric survey pipeline. This new tool can help researchers to better manage and organize historical information.

1. INTRODUCTION

Collecting data and information for the documentation of Cultural Heritage plays an important role in creating the basis of knowledge necessary to make the best decisions for protection and conservation. Documentation is a central aspect for the institutions and organizations that work in this field and, in recent years, they have shown an increased interest in new technologies that can support the documentation process. However, many new tools are not easy to manage or their utility is not appreciated. One of the major problems is, in fact, the difficult collaboration between those who collect the data and those who use them. Increasing awareness of the potentialities of technology to support the documentation process can enhance this collaboration.

This paper assesses the significance of a close relationship between information users and information providers and describes the implementation of ICT and geomatics technologies in a field of application to support the historical and restoration studies on Cultural Heritage. As the need to document heritage is important for the existing assets, the same need is fundamental in the case in which they no more exist. In these cases, the only remaining traces are represented by pictures and film footage in which parts of cities, urban environments and heritage monuments appear, which have been lost or transformed. These represent invaluable sources of cultural and historic information that are useful for architectural and restoration studies, especially when the heritage has not been surveyed before it was lost. However, one major issue of historical archives material such as old images and videos is their availability, often made difficult by an enormous quantity of unorganized data on historic heritage.

The motivation for this study stem from the need to perform historical research without the effort of the manual examination of videos in archives, and also the need to certify the metric quality of the 3D model of the monument obtained from the video process. To reach this objective, this research presents a possible approach to customised the standard photogrammetric pipeline for the processing of historical film footage and extends it with two new steps: the automatic detection with Deep Learning (DL) of frames in which architectural heritage appears and suitable to be process with photogrammetry; and the metric quality assessment of the model resulting from the reconstruction process.

This paper is divided into four parts. The first part examines open issues in collecting historical material and a state of the art of DL is presented. The second part describes the methodology used for experimenting the workflow and the metrics used to evaluate it. A case study is presented in the third part, focusing on the preparation of the datasets, finally the fourth part analyses and discusses the results.

2. EXPLOITING NEURAL NETWORK FOR CULTURAL HERITAGE DOCUMENTATION

The major critical factors of collecting historical material are the following:

(1) The difficulty of finding it often requires physical access to the archives, which in various cases allow on-site viewing but not data sharing. In this regard, the development of international projects (such as iMediaCities, Caraceni et al., 2017) that aim to limit the barriers to access data in video archives deserves special attention.

* Corresponding author

(2) The need to identify the object of interest within the amount of material that potentially contains it.
Metadata indexing for historical archive material is often incomplete or inaccurate and the associated search engines are consequently not very efficient. The human effort to find the data of interest therefore represents a significant percentage in the work of the geomatics specialist.
In computer science research, Machine Learning (ML) has acquired a fundamental importance not only in the research domain, but also in contemporary everyday life, as for example web search engines, fraud detection systems, spam filters, automatic text analysis systems and medical diagnosis systems, just to name a few. One of the reasons of this increasing importance is the successfully adoption of DL methods in fields as image classification (Krizhevsky et al., 2012; Szegedy et al., 2015; Simonyan and Zisserman, 2014) where Convolutional Neural Networks (CNNs) outperforms human level in object recognition and image search (Radenović et al., 2016; Tolias et al., 2016). However only few studies in Cultural Heritage has grown up in this area. To date, using this approach, researchers have been able to: classify elements of interest in images of buildings with an architectural heritage value (Llamas et al., 2017); detect different monuments based on the features of the monument's images (Saini et al., 2017); automatically annotate Cultural Heritage assets using their visual features as well as the metadata available at hand (Belhi et al., 2018); and develop a mobile app to recognize monuments (Palma, 2019). Besides the improvements of Machine Learning techniques, hardware development, in particular the use of Graphical Processing Units (GPUs), has given boost to the computational efficiency of such algorithms.

3. PROPOSED WORKFLOW

A photogrammetric workflow integrated with DL techniques for the input material preparation is proposed. As shown in Figure 1, the standard photogrammetric process based on Structure-from-Motion (SfM) pipeline is modified with the use of DL for primary data recovering.
To allow a better evaluation of the validity of the pipeline, the figure also represents an additional final validation phase in which the result is compared by using the benchmarks described in Condorelli and Rinaudo, 2019. The pipeline chosen as reference is the COLMAP (Schönberger et al., 2016) open-source Structure-from-Motion and Multi-View Stereo (MVS) algorithm implementation, developed by ETH of Zurich, (COLMAP, Johannes L. Schoenberger, 2019).

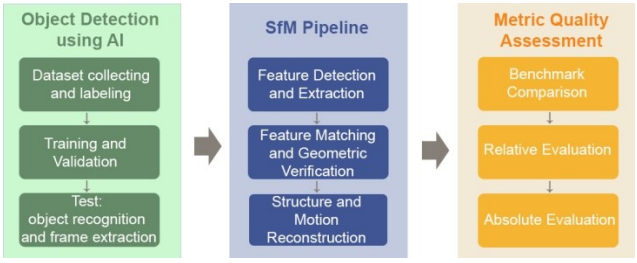

Figure 1. Flowchart of the proposed workflow.

3.1 Object detection using NN

The objective of the preliminary phase in which to use NN consists in extracting the video frames containing the object of interest. Object detection is a typical Deep Learning search and has been considered as effective in the present pipeline since it includes the image classification also in complex images and with bounding box extraction.
Considering the expected user for the pipeline, the easy use is a key aspect.
The Luminoth software (Rey et al., 2017) based on TensorFlow (Abadi et al., 2015) was selected. Luminoth implements object detection through state of the art networks. In particular:
• Faster RCNN (Ren et al, 2016) for more accurate solutions;
• SSD (Liu et al., 2016) oriented to reduced computational demand.
Luminoth provides these networks already pre-trained but allows to train them further by adding own elements to recognize. From the Cultural Heritage point of view, specific training is a necessary step. From an operational point of view, it is necessary to acquire valid starting data that will be used for training and validation. To annotate the bounding box of the architectural heritage, the VGG Image Annotator (VIA) tool (Dutta el al., 2016) was used.
In order to train a model using Luminoth, a simple configuration file has to be created specifying some required information, such as a run name, the dataset location and the model to use. Therefore training can be launched and the quality of the produced network can be tracked from the progress of the validation results. The authors have customized – using an *ad-hoc* Python script – the evaluation of the quality of the results against the most relevant metrics for their specific needs. In this regard, it is worth noting that a high quality for the bounding box is not considered crucial for photogrammetric purposes. Then, since the Luminoth output is a probability associated to the detected object, it was defined the probability threshold which was considered as the minimum value to accept the NN detection as positive. The selection of an effective threshold is not trivial and will be detailed in the discussion of results.

Considering a dataset of images, it is first defined P (N) as the number of images in which the object is present (not present), respectively. When performing the inference phase, these values are not known, and the network output is P' (N') that represents the images in which the network has found (not found) the object. In the test phase, where therefore P and N are known, it is possible to evaluate the images according to 4 statuses: True Positive (TP, image in both P and P '), True Negative (TN, image both in N and in N '), False Positive (FP, image in P' but not in P) and False Negative (FN, image in N 'but not in N). Obviously T=TP+FN and N=TN+FP.
In addition, typical parameter ratios are defined to highlight different properties of the network. Two typical indicators are:
• Sensitivity (SN), calculated as the number of correct positive predictions divided by the total number of positives:

$$SN = TP / (TP+FN).$$

• Specificity (SP), calculated as the number of correct negative predictions divided by the total number of negatives:

$$SP = TN / (TN+FP)$$

In particular, these indicators are the most representative in the case of validation sets containing all values of the same class (P or N) and in this case both SN and SP coincide with the accuracy:

$$AC = (TP + TN) / (TP + TN + FP + FN).$$

Unlike the validation set, which can be artificially created to investigate specific image types (see section 4), in the case of video analysis, it deals with realistic P and N distributions. It is reasonable to expect that in a generic historical video a certain searched object will appear in a large minority of frames, i.e. P << N. The specificity therefore reveals the ratio of videos that

need to be watched even if they do not actually contain the tower. Indeed:

$$SP = (N-FP) / (N-FP + FP) = 1-FP / N \sim 1-FP / (N\_FRAMES)$$

SP therefore corresponds to the ratio of the video sections that can be skipped, and, ultimately, to the saving factor of human time achieved thanks to the use of Machine Learning.

Finally, in the case of videos, it may be of interest to introduce the precision defined as:

$$PR = TP / (TP+FP).$$

The precision gives the percentage of results which are relevant.

In general, SN, SP, PR should all be maximized but which one is to prefer is a matter of usage context. In view of the type of use of the network within the photogrammetric pipeline, two extreme cases of use are considered:

(1) In the first situation in which the videos selected by the ML are then manually watched to decide which are the most suitable for photogrammetry, it is ideal to minimize the FN (and therefore the SN) to avoid losing useful information.

(2) In the second situation in which the pipeline is managed in a more automatic manner, it is instead preferable to minimize the FP (therefore the SP) to prevent incorrect images from entering the subsequent processing.

3.2 COLMAP SfM sequential processing pipeline

For the iterative reconstruction from historical film footage the steps are the following: (1) Feature detection and extraction: the intrinsic parameters are unknown a priori it is recommended to choose the camera model that is able to model distortion effects considering the following parameters: f, cx, cy, k1, k2, that is one focal length (f), two coordinates of the principal point (cx, cy) and two radial distortion parameters (k1, k2). (2) Feature matching and geometric verification: choosing a sequential matching mode, suitable for images acquired in sequential order by a video camera, allows to match only consecutive frames with visual overlap. (3) Structure and motion reconstruction.

For the precision analysis, the values of the Final Cost (residuals on re-projected points) from the bundle adjustment report of the SfM process were examined. Subsequently they were compared with a benchmark of the maximum metric quality that can be reached by implementing photogrammetry on videos, according to specific camera motion and a taking distance. Final Cost represents the average of the reprojection residuals over all image observations and it is expressed in pixel.

4. CASE STUDY AND DATASETS

4.1 Tour Saint Jacques

In order to test the algorithm, a case-study approach was adopted and the architectural heritage of the Tour Saint Jacques in Rue Rivoli in Paris's 4th arrondissement, inscribed in the UNESCO Heritage List since 1998, was chosen.

This bell tower, in flamboyant gothic architecture, represents the only evidence of the lost Saint-Jacques-de-la-Boucherie church and has been subjected to many transformations over time. The church began as a Carolingian chapel, then underwent more expansions along the centuries. The priory was torn down in 1797 after civil unrest, but the bell tower was protected because considered of high architectonic value (Meurgey, 1926). The tower was embedded in the urban pattern, but during the restoration in the 1850s it was moved from the original position and elevated on a decorative stone podium (O'Connell, 2001).

Figure 2. Transformation of the tower during the time.

Even now the tower exists and appears in many historical film footage (from the 1910s until 1960s) stored in several video archives in Paris (Lobster, CNC, Forum des Images). Moreover, a 3D model, obtained from a recent photogrammetric survey, date by 2015, of the existing tower made by Iconem is available. For this reason, the tower represents a good case study on which implemented the experimental pipeline.

4.2 Dataset

Despite the large quantity of data available on the tower, no dataset suitable for the training and validation of the Neural Network existed and a new dataset was created.

4.2.1 Training and validation datasets

It is well known that the quality of data used in training NN is crucial to achieve good results. In particular, a significant level of data entropy is needed to effectively learn the features of the object. To this aim, hundreds of contemporary and historical images of the tower with different lighting conditions, backgrounds and points of view were collected by the authors with three methods: (1) web crawling; (2) ad hoc photographic survey in the new location of the tower; (3) historical archives consultation in Paris.

The images resulting from this research were divided into four groups. First, both contemporary and historical images of the entire tower were included. Then, images representing only details of the tower were used because it is a typical situation when dealing with film footage. For the same reason, views with the skyline of Paris were considered, both with the tower and without. Moreover, images were selected which show monuments or architectures similar to the Saint Jacques tower in different contexts. These last images act as "negative matching" and can lower the incidence of false positive ratio in Machine Learning classification problems (Hu et al., 2014; Kalal et al., 2015).

The collected images from the four groups were used for the training and validation stages. In particular, random mix of images were used in different trainings to improve the quality of the network according to the requirements of the investigation, as explained in the next section. Moreover, four validation datasets with 80 images from each group, respectively as valid1, valid2, valid3 and valid4, were used to assess the quality of results from different perspectives.

4.2.2 Test dataset

To test the performance of the algorithm in a realistic case, a dataset collecting historical videos from archives in Paris has been considered. Despite the criticalities in retrieving these materials (see section 2), a significant set of footage were collected, and their characteristics are described in Table 2.

Dataset	Description	From web	From survey	From historical photographs	Number in training RUN A	Number in training RUN B	Number in training RUN C	Number in validation
1	Tour Saint Jacques	x	x	x	400	400	400	80
2	Landscape	x						80
3	Negative matching	x		x		200	200	80
4	Tour Saint Jacques Parts	x	x		80	80	80	80

Table 1. Training and validation datasets description.

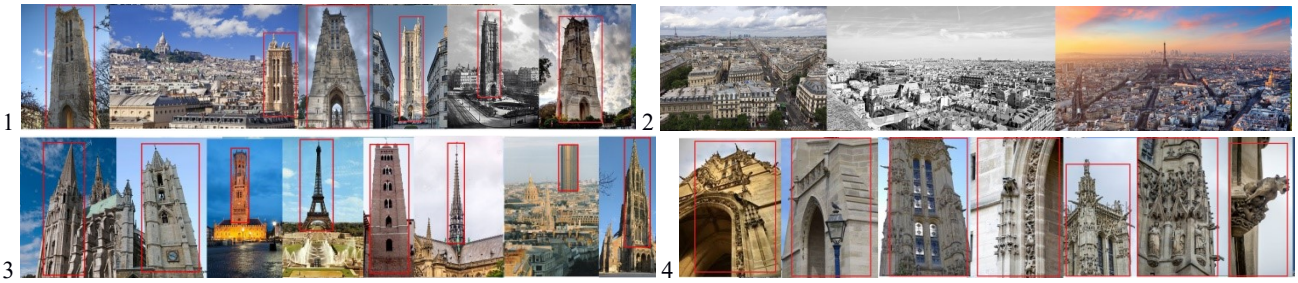

Figure 3. A selection of the pictures from the datasets: (1) Tour Saint Jacuqes; (2) landscape; (3) negative matching; (4) tower parts.

Dataset	Duration	Year	Director	Type	Film	Colour	Archive
La tour Saint Jacques	9min 47s	1967	J. Sanger	documentary		Black and white	Ina.fr
Études sur Paris	76min	1928	A. Sauvage	documentary	16 mm	Black and white	CNC and VOD
Paris, Roman d'une Ville	49min	1991	S. Neumann	documentary	16 mm	Black and white	Forum des Images
Paris 2ème partie	4min 44s	1935	G. Auger	documentary	16 mm	Black and white	Forum des Images
Passant par Paris	13min 39s	1955	P. Perrier	fiction	8 mm	Black and white	Forum des Images
Vue Panoramique sur Paris	2min	1954	A. Lartigue	documentary	16 mm	Black and white	Forum des Images
Un film sur Paris	45min	1926	C. Lambert, J. Levesque	documentary		Black and white	Lobster
La nouvelle babylone	24s	1929	L. Trauberg, G. Kozintsev	historical		Black and white	Lobster
Paris, 1946	13min	1946	J.C. Bernard	documentary		Colour	Lobster
La grande roue	4min 20s	1913		documentary		Black and white	Lobster
Paris et ses monuments	7s	1912	Pathe	documentary		Black and white	Lobster

Table 2. Test dataset description.

5. RESULTS AND DISCUSSION

5.1 Networks validation and evaluation

The first network trained is the RCNN – labelled by Luminoth as an accurate network – and in a first training (RUN A) only positive matches were used, i.e. included only images of the Saint Jacques tower, with integral or partial views (dataset 1 and 4). The results on the training set and on the validation sets containing the tower (valid1 and valid4) are represented in terms of sensitivity in Figure 4 and for two reference probability thresholds equal to 0.5 and 0.9. The trends show a rapid convergence of the network, and a very high training accuracy, as well as for the validation set valid1 which includes only complete images of the tower. The specificity becomes more limited (around 0.8) for valid4, since the presence of partial views of the tower is not always detected. Moreover, as was to be expected, by increasing the probability threshold the network becomes more selective and the sensitivity tends to decrease, especially in the case of valid4.

To complete the network quality analysis the validation behaviour of validation sets not including the Saint Jacques tower was analysed. Figure 5 shows the specificity trend for valid2 and valid3. The trend of valid2 is quite good (around 90%), which contains images in contexts similar to those in

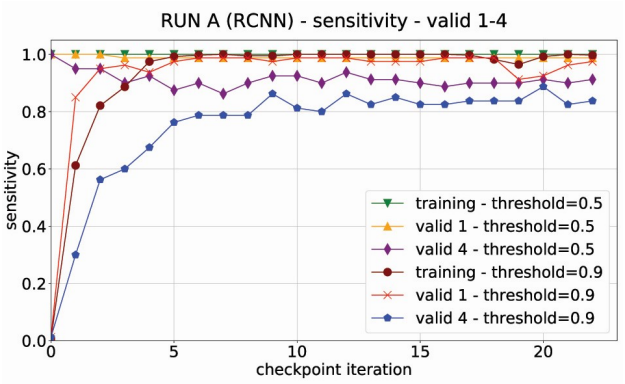

Figure 4. Sensitivity trend of RUN A with valid 1 and 4.

which the tower is located. For the set valid3 which is characterized by images of towers similar to the tower of Saint Jacques, as expected, the negative match is more misleading and brings the specificity to rather modest levels (around 50%): therefore the network is not able to distinguish with great accuracy the real tower from other similar towers. As expected, by increasing the threshold the network becomes more selective and therefore the problem of false positives is, at least in part, mitigated.

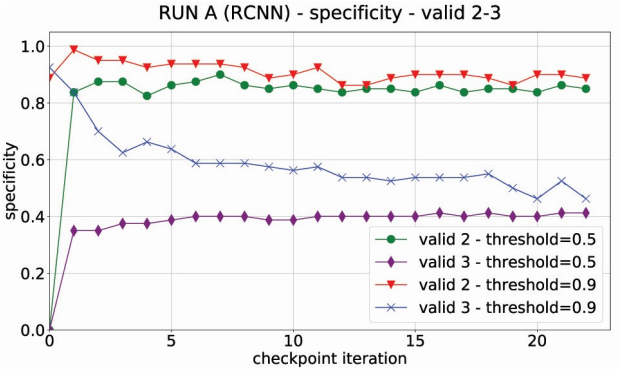

Figure 5. Specificity trend od RUN A with valid 2 and 3.

To try to improve the network's performance, a second training (RUN B) was performed in which the negative matching was already included in the training phase as an additional category. Analysing the network's sensitivity – shown in Figure 6 – it is noted that, as expected, the network became more selective with respect to the previous training and therefore the recognition of the true positives slightly worsened. However, with regard to the specificity – shown in Figure 7 – the problem of false positives appears completely solved. For this reason the advantages of RUN B training outweigh the disadvantages. However, depending on the case in question, it could be decided to always favour sensitivity so that RUN A has slightly superior performance.

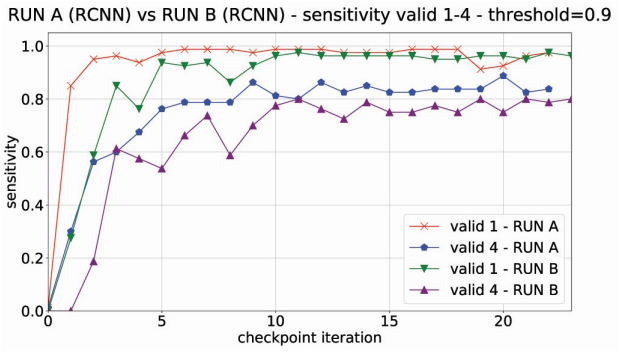

Figure 6. Sensitivity trend of RUN A and RUN B with valid 1 and 4, and threshold 0.9.

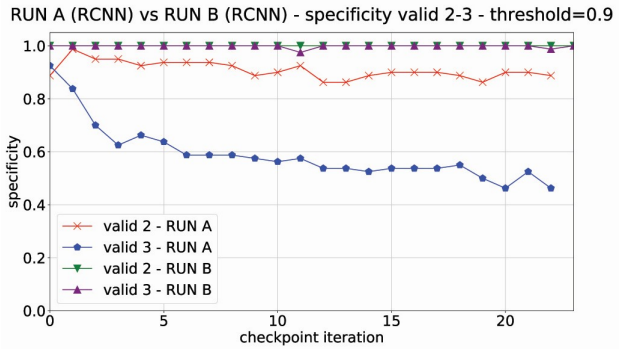

Figure 7. Specificity trend of RUN A and RUN B with valid 2 and 3, and threshold 0.9.

To more fully analyse the influence of the probability threshold on the results, in Figures 8 and 9 the sensitivity and specificity trends for the two performed trainings were analysed. The sensitivity is expected to decrease with the threshold whereas the specificity should increase. In the figures, two reasonable thresholds to be set, 0.8 or 0.9 are represented.

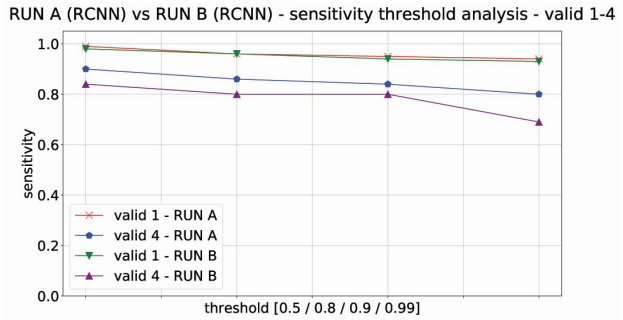

Figure 8. Sensitivity threshold analysis for RUN A and RUN B with valid 1 and 4.

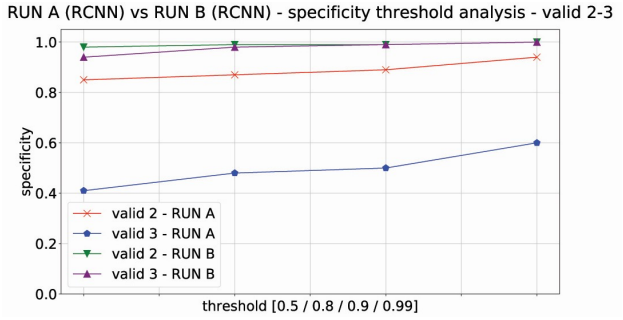

Figure 9. Specificity threshold analysis for RUN A and RUN B with valid 2 and 3.

As further training RUN C, the SSD network was implemented – labelled by Luminoth as Fast – considering the training with the negative matching activated. Comparing the results between the RCNN and SSD networks (as shown in Figure 10), it can be seen that the latter guarantees better sensitivity (around 10% improvement) on positive cases. On the other hand, the specificity only slightly deteriorates.

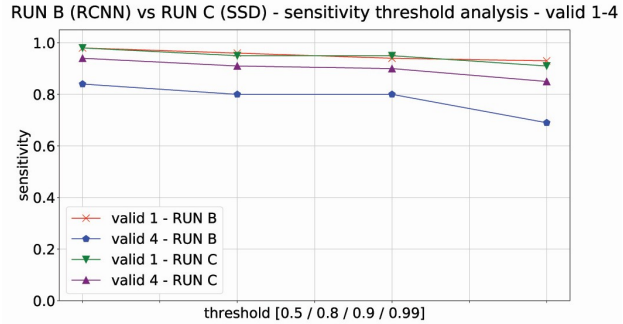

Figure 10. Sensitivity threshold analysis for RUN B and RUN C with valid 1 and 4.

In conclusion, the performance of three Neural Networks evaluated on four validation sets were analysed, in order to highlight the behaviour of the networks in four typical reference conditions (as shown in Figure 11). The addition of the negative matching used in RUN B and RUN C constitutes in principle an improvement in the quality of the results. As regards an effective comparison between RUN B and RUN C it is

necessary to evaluate the network with respect to datasets that present a realistic distribution of positives and negatives.

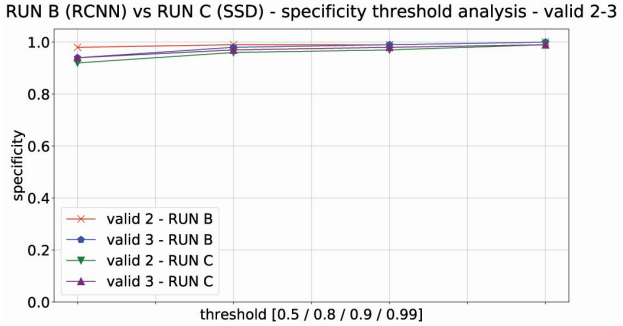

Figure 11. Specificity threshold analysis for RUN B and RUN C with valid 2 and 3.

5.2 Hardware comparison

The use of High Performance Computing (HPC) architectures is definitely advantageous in the context of Machine Learning and GPUs (Graphical Processing Units) are currently particularly convenient devices for applications of this type. By way of comparison, the computing times for processing an image of the dataset in the training phase are shown in Table 3.

NVIDIA 630M	NVIDIA K80s	NVIDIA P100
30 s/image	1 s/image	0.5 s/image

Table 3. Hardware comparison.

The elapsed times are reported for the case of a common laptop with an NVIDIA 630M entry-level graphics card, and two types of HPC computing nodes featuring two NVIDIA Tesla GPUs, namely K80s and P100, respectively. The difference in timing is very marked. For a complete training the times pass from several tens of days to less than 24h. In the massive inference phase, the use of HPC platforms can become a fundamental requirement. The use of GPUs designed for computing also gives an advantage in terms of reliability.

5.3 Network inference on historical videos

The trained RUN B and RUN C networks were tested on the set of videos described in Table 2. The results were obtained by re-sampling the videos at 2 frame per second (fps). In the cases investigated in this work, it has been verified that the results do not change significantly increasing further the fps. Clearly working at low fps allows a massive reduction in the computing times of this inference phase.
From the graph in Figure 13, it is worth noting that the sensitivity values are lower than the corresponding validation ones. By manually analysing the cases of FN, it turns out that the vast majority of the not detected towers are very low quality images, and in several cases drawings. As for the specificity, the values are high for both networks. In particular, for the RUN B network the values are very close to the unity ideal value. However, high specificity values can be due to the high number of negative values (N) which is much larger than the positive corresponding ones (P). In terms of absolute values the SSD network presents a significant amount of FP. To give a quantitative assessment of this, the precision is also represented in the plot. It turns out that the SSD network has a very low precision (less than 0.25) except for threshold values (greater than 0.99) where the sensitivity is dramatically low. In addition to these indicators, it is worth recalling that the SSD network is computationally much lighter. Concluding, both networks have their own advantages and the final choice is a matter of the specific requirements of the research including the available computational resources.

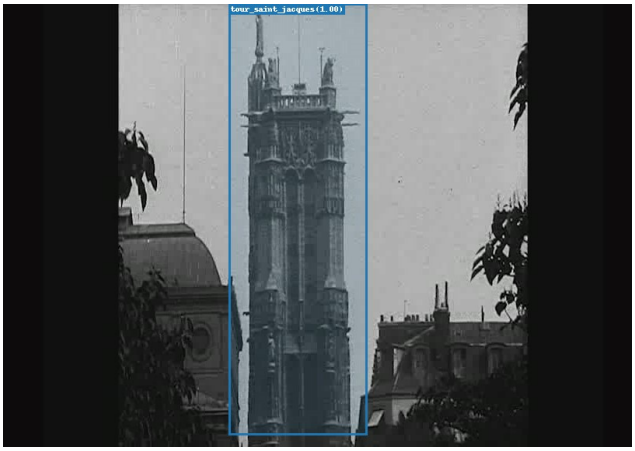

Figure 12. An example of the Tour Saint Jacques detected in the footage.

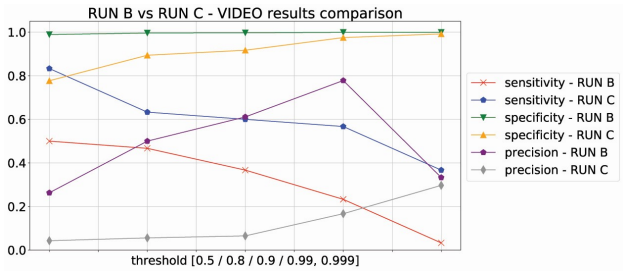

Figure 13. Results of network inference on historical videos.

5.4 Photogrammetry: processing and evaluation

The algorithm, after recognised the architectural heritage in the footage, automatically selects and extracts frames suitable to be process with photogrammetry. Among footage with high probability, the film chosen is “Études sur Paris” from the CNC-VOD archive (Figure 14). The tower appears in the video shot with the Tilting type of camera motion and presents the following characteristics: Gauge 16 mm; Focal Length 25 mm; Digital format Resolution 480x360 pixels.

Figure 14. A selection of frames from the film footage “Études sur Paris”.

To the 16 selected frame the SfM pipeline was applied and the 3D reconstruction succeeded, as shown in Figure 15.

Figure 15. Results from SfM pipeline: (1) Feature detection and extraction, (2) Feature matching and geometric verification, (3) Structure and motion reconstruction.

For the precision analysis, all values of Final Cost (FC) were used for the calculation of the Mean and the Standard Deviation and reported in the following graphs to analyse the trend of the data and the comparison with the benchmark (Figure 16).

Moreover, the minimum and the maximum values were highlighted and transformed in centimetre with the Ground Sample Distance (GSD) calculation, considering a distance of 15 m (GSD benchmark=1.2 [cm/px], GSD case study=1,43) and reported in Table 4.

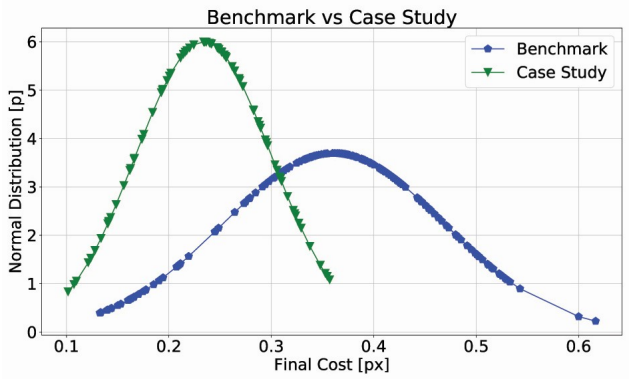

Figure 16. Comparison of Normal Distribution of the Final Cost value between benchmark and case study.

Case	Mean	St Dev	Min FC	Max FC
Benchmark	0.36 [px]	0.10 [px]	0.13 [px]	0.60 [px]
Case study	0.23 [px]	0.06 [px]	0.10[px]	0.35 [px]
Benchmark	0.1 [cm]	0.8 [cm]	0.1 [cm]	0.8 [cm]
Case study	0.33 [cm]	0.08 [cm]	0.14 [cm]	0.5 [cm]

Table 4. Mean, Standard Deviation, Min and Max values of Final Cost.

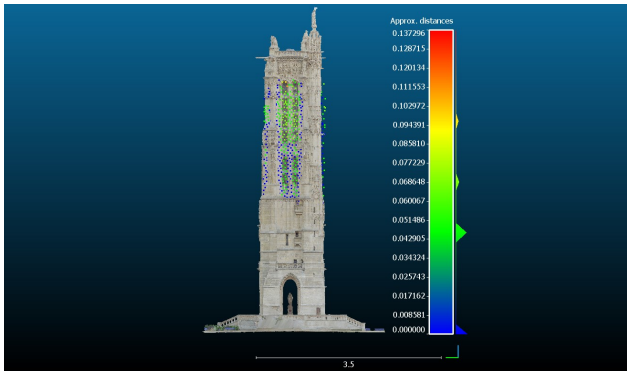

Figure 17. Distances Comparison between obtained point cloud and existing 3D model.

From graphs comes out that also for the case study the trend of the curves follows the Gaussian Distribution, like the case of the benchmark. It is apparent from the tables that, comparing the two results, it can be seen that the differences between values of the case study and the benchmark are not significative.
Moreover, the point cloud obtained from the process, even if with low density, was compared with the 3D model by Iconem. The comparison (Figure 17) showed that the distances calculated between the mesh of the model and the point cloud resulted are less of the order of one pixel.

6. CONCLUSIONS AND PERSPECTIVES

The importance and originality of the experimental work presented here are that it explores innovative ways to reduce human effort and to increase overall efficiency. ML potentially allows a drastic decrease in the time required to search for monuments within historical archive video materials.
This paper discussed how to effectively train state-of-the-art Neural Networks for researching historical monuments. In particular, the Tour Saint Jacques was chosen as a case study. This tower still exists but has undergone many transformations. Thus, a large quantity of material was available to test the algorithm and to obtain a metric comparison and test the potentialities of the approach. The quality of the results is encouraging both in terms of saving human time and in terms of results achieved according to the appropriate metrics of the case in question. This investigation can then be inserted as the first stage of a photogrammetric pipeline and allows the identification of frames that can be used for 3D reconstruction. The quality of the reconstruction of historical video was also evaluated according to a previously defined and useful benchmark.
The findings should make an important contribution to the field of Cultural Heritage studies because they provide a new tool for the research of historical material in archives. The tool can be of great use for both architectural historians and geomatics experts, to study the evolution of cities and buildings.
The strategy described in this work can be applied to different historical monuments. In particular, it would be interesting to apply the procedure to destroyed monuments for which the 3D reconstruction from historical videos is the only possible option. For a monument that no longer exists, training datasets can be found in historical material, and in this case it would be interesting to evaluate the network performance under these conditions. A further field of development could be simplifying the use of ML in view of a possible non-expert user. The development of an intuitive interface that allows the automation of more complex phases of the process would be a great help. Another important development could be the automatic determination of camera movements in the time intervals in which the object was detected. In principle this would allow the automation of the entire photogrammetric pipeline and could therefore be a particularly interesting line of investigation.

ACKNOWLEDGEMENTS

The authors express thankfulness to the archive Lobster Films for sharing footage used in this research, to CNC, Forum des Images and Ina.fr, and to ICONEM for kindly making available the model of the Tour Saint Jacques. The authors also acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support.
This work was supported and funded by the GAMHer project (Geomatics data Acquisition and Management for landscape

and built Heritage in a European perspective), a 3-year project financed under the Italian PRIN 2015 framework (Progetti di Ricerca di Rilevante Interesse Nazionale).

REFERENCES

Abadi, M., et al., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. http://tensorflow.org (30 June 2019).

Belhi, A., Bouras, A. and Foufou, S., 2018. Leveraging Known Data for Missing Label Prediction in Cultural Heritage Context. In: *Applied Sciences*, Vol. 8(10), 1768, pp. 1-19, doi.org/10.3390/app8101768.

Caraceni, S., Carpenè, M., D'Antonio, M., Fiameni, G., Guidazzoli, A., Imboden, S., Liguori, M. C., Montanari, M., Trotta, G., Scipione, G. and Hanegreefs, D, 2017. I-media-cities, a searchable platform on moving images with automatic and manual annotations. In: 23rd International Conference on Virtual System & Multimedia, pp. 1-8, doi.org/10.1109/VSMM.2017.8346274.

COLMAP, Johannes L. Schoenberger, 2019. COLMAP - Structure-From-Motion and Multi-View Stereo. https://github.com/colmap/colmap (30 June 2019).

Condorelli, F. and Rinaudo, F., 2019. Benchmark of metric quality assessment in photogrammetric reconstruction for historical film footage. In: *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., Vol. XLII-2/W11, pp. 443-448, doi.org/10.5194/isprs-archives-XLII-2-W11-443-2019.

Dutta, A., Gupta, A. and Zissermann, A., 2016. VGG Image Annotator (VIA), http://www.robots.ox.ac.uk/vgg/software/via (30 June 2019).

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, doi.org/10.1109/CVPR.2016.90.

Hu, B., Lu, Z., Li, H. and Chen, Q., 2014. Convolutional Neural Network Architectures for Matching Natural Language Sentences. In: *Advances in Neural Information Processing Systems, NIPS 2014*, Vol. 27, pp. 1-9, arXiv.org/1503.03244v1.

Kalal, Z., Matas, J. and Mikolajczyk, K., 2015. P-n learning: Bootstrapping binary classifiers by structural constraints. In: CVPR, 2015, pp. 1-8.

Krizhevsky, A., Sutskever, I. and Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. In: *NIPS, 2012*., Vol. 1, pp. 1097-1105, doi.org/10.1145/3065386.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C. and Reed, S., 2016. SSD: Single shot multibox detector. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science*, Vol. 9905, pp. 21-37, doi.org/10.1007/978-3-319-46448-0_2.

Llamas, J., Lerones, P. M., Medina, R., Zalama, E. and Gómez-García-Bermejo, J., 2017. Classification of Architectural Heritage Images Using Deep Learning Techniques. In: *Appl. Sci. 2017*, Vol. 7(10), 992, pp. 1-25, doi.org/10.3390/app7100992.

Meurgey, J., 1926. Histoire de la paroisse Saint-Jacques de-la-Boucherie. Paris, Champion, Paris, pp. 347.

O'Connell, L., 2001. Afterlives of the Tour Saint-Jacques: Plotting the Perceptual History of an Urban Fragment. In: *Journal of the Society of Architectural Historians*, Vol. 60, pp.450-473.

Palma, V., 2019. Towards deep learning for architecture: a monument recognition mobile app. In: *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., Vol. XLII-2/W9, pp. 551-556, doi.org/10.5194/isprs-archives-XLII-2-W9-551-2019.

Radenović, F., Tolias, G. and Chum, O., 2016. CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples. In: European Conference on Computer Vision ECCV 2016, pp. 1-17, doi.org/10.1007/978-3-319-46448-0_1.

Ren, S., He, K., Girshick, R. and Sun, J., 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, 6, pp. 1-13, doi.org/10.1109/TPAMI.2016.2577031.

Rey, J., Tayler, I., Descoins, A., Rodriguez, G., Azzinnari, A., et al., 2017. Open source computer vision toolkit. https://github.com/tryolabs/luminoth (30 June 2019).

Saini, A., Gupta, T., Kumar, G., Kumar Gupta, A., Panwar, M. and Mittal, A., 2017. Image based Indian Monument Recognition using Convoluted Neural Networks. In: International Conference on Big Data, IoT and Data Science, pp. 1-5, doi.org/10.1109/BID.2017.8336587.

Schönberger, J. L. and Frahm, J. M., 2016. Structure-from-motion revisited. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Vol. 2016, pp. 4104-4113.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. In: ICLR 2015, pp. 1-14, arxiv.org/abs/1409.1556.

Szegedy, C., Liu, X., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In: CVPR, 2015, pp. 1-12, arxiv.org/abs/1409.4842.

Tolias, G., Sicre, R. and Jégou, H., 2016. Particular Object Retrieval With Integral Max-Pooling of CNN Activations. In: ICL 2016, pp.1-12, hal-01842218.

Zanella, R., Fiameni, R. and Rorro, M., 2018. A performance study of Machine and deep learning frameworks on Cineca HPC systems. In: *Advances in Parallel Computing*, Vol 33, pp. 1-15, 978-1-61499-881-5.