A MATCH-MOVING METHOD COMBINING AI AND SFM ALGORITHMS IN HISTORICAL FILM FOOTAGE

F.Condorelli [1*], F. Rinaudo [1], F. Salvadore [2], S. Tagliaventi [2]

[1] DAD, Department of Architecture and Design, Politecnico di Torino, Italy - (francesca.condorelli, fulvio.rinaudo)@polito.it
[2] CINECA – HPC Department, Rome, Italy - (f.salvadore, s.tagliaventi)@cineca.it

Commission II, WG II/8

KEY WORDS: Object Detection, Neural Networks, Camera Tracking, Photogrammetry, Cultural Heritage, Metric Quality Assessment

ABSTRACT:

Searching for suitable material for photogrammetry is a key part in the documentation of Cultural Heritage. Photogrammetry can be used to produce a metrically certified 3D model. Material contained in historical film footage archives is especially useful for documentation when the heritage has been lost. In this research an innovative match-moving method is proposed that aims to exploit Artificial Intelligence and SfM algorithms to identify the frames extracted from a film footage in which the lost monument appears and that are suitable to be processed with photogrammetry for its 3D reconstruction. First of all the identification and tracking of the heritage in the videos was performed training an object detection Neural Network. Then the frames detected were automatically extracted with the coordinates of the bounding boxes that contain the monument. The camera motions were identified by selecting only the shots taken from multiple points of view of the same scene and analysing the evolution of the bounding boxes position over time. A further check of the material was necessary to select only sequences and to eliminate single frames and images from different historic periods. After this process, only the correct frames were automatically selected and processed with photogrammetry and the quality of the obtained 3D model was assessed. The method experimented in this research represents a powerful tool in the field of Cultural Heritage because it makes the selection of suitable material for photogrammetry automatic. Moreover it offers important insights that could be extended to other sectors.

1. INTRODUCTION

The rapid development and diffusion of ways to shoot videos used in a wide variety of applications, both professional and amateur, is being increasingly documented. Consequently there is now a large quantity of data available and this has led to an increased interest in using this material, especially in the field of Cultural Heritage. Photogrammetry plays a key role in this context since it represents a powerful and stable technique to document Heritage and to extract metric information.
The objective of this work is to apply a new match-moving method to document lost monuments by exploiting the metric potential of historical archive material.
Match-moving is a technique used to track the movements of a camera in a 3D space using the images that it acquires while moving. This method is widely used in computer vision, the film making industry and video editing because it allows the real scene to be matched with virtual creations such as visual effects. Structure from Motion (SfM) is the main part of this process which allows the extraction of the 3D information from the scene.
In this paper historical films were chosen to experiment the method because in many cases they are the only remaining traces of Cultural Heritage that have been lost or changed over time. By using photogrammetry with historical film footage, it is possible to process the data and reconstruct the heritage virtually. However, its realisation is technically demanding as it is difficult to find historical data which is suitable for processing. After identifying the monument on film, to be documented, it is necessary to understand whether the selected images are suitable for photogrammetric processing in order to virtually reconstruct the building. The duration of the film and the way in which the video was shot are determining. Generally it is very rare to find historical footage with long tracks and camera movements filmed from different angles on the same building. Both of these are necessary to create normal or converging views, which are required for photogrammetry.

This study seeks to select suitable data within a large quantity of unorganized and low-quality material, which will be used both for the search of a historical monument and its photogrammetric reconstruction. To do that a match-moving workflow is proposed with the specific aim of identifying the frames extracted from the film footage in which the lost monument appears, and that are suitable for processing with photogrammetry for 3D reconstruction.
The remaining part of the paper proceeds as follows: the first part begins by giving an overview of the recent match-moving algorithms and their applications; it will then go on to illustrate the proposed workflow highlighting the innovation and the differences with the previous existing methods. The third part presents the case studies on which the algorithm was implemented and the fourth part is concerned with the discussion of the results. Conclusions and future perspectives of the method experimented here will be provided in the final section.

2. EXISTING MATCH-MOVING METHODS

As introduced, the technology behind camera tracking is based on the SfM procedure since the determination of the position of the camera and the field of view is done by analyzing the film

* Corresponding author

shot and extrapolating 3D data from the original 2D imagery (Condell and Moore, 2006; Zhang et al., 2009; Ingwer et al., 2015).

According to previous studies (Lee et al., 2006), camera match-moving approaches can be classified into two categories: feature-based and model-based approaches. The first one uses appropriate feature points, i.e. a method based on a 3D plane tracking technique, which allows the estimation of the homographies induced by a 3D plane between successive image pairs (Lourakis and Argyros, 2005); a practical realtime camera tracking system which involves an offline process for space abstraction using features and an online step for feature matching (Dong et al., 2009); and a non-consecutive feature tracking framework to match interrupted tracks distributed in different subsequences or even in different videos (Zhang et al., 2016). The second one uses a known geometric object in the given environment, i.e. the development of a real-time marker-based camera tracking method working in unknown environments using a known marker for the fast detection and tracking of feature points (Lee et al., 2006). Other technique uses depth information to evaluate the camera pose and trajectory (Luo et al., 2016).

Different tools and software exist to recreate the path of the camera and they vary in price, usage, functionality and user interface. Among the commercial software, the most used are Boujou (Vicon Motion Systems Ltd UK) that works through a frame by frame comparison to track the camera; and SynthEyes (Andersson Technologies LLC) that can determine how the real camera moved during the shoot, what the camera's field of view (focal length) was, and where various locations were in 3-D space. An example of open source software is ACTS (Zhang et al., 2009), an automatic camera tracking system which supports to track camera motion but limited only on two types, pure rotation and free-moving. Another restriction is that it works only with long sequences.

Some research has been carried out on match-moving for analysis of buildings (Dağlar et al., 2011) and for the use of video sequences as a source of metric data from the filmed architectures (Mancini et al., 2013). However, a precise pipeline to virtually reconstruct lost Cultural Heritage has not yet been established in the architectural field.

Two important themes emerge from the studies discussed so far. Most of them are dated and there is a lack of experimentation of the more efficient techniques of Artificial Intelligence, which are very useful for such tasks. Moreover, no open source software exists, with the exception of one case that presents some limitations especially concerning the inability to handle short tracks and tilting and trucking camera motion types, both of which are very common in historical film footage.

3. INNOVATION OF THE PROPOSED WORKFLOW

This paper proposes an innovative and open source match-moving method which combines Artificial Intelligence (AI), in particular Deep Learning (DL), with the Structure-from-Motion (SfM) open source algorithms on historical film footage to document lost Cultural Heritage.

3.1 Standard method VS proposed method

In Section 2 previous studies on existing methodologies regarding camera motion estimation were discussed and the most important open issues were highlighted. However, overall the match-moving process is similar in every software and consists in the following steps, as shown in Figure 1: feature identification and tracking, camera tracking and 3D modelling (Haji et al., 2016; Dağlar et al., 2011).

- Feature tracking consists in finding the position of points of interest in the footage calculating their motion vectors frame by frame.
- Camera tracking finds the motion of the camera in 3D space extracting with SfM its characteristics (orientation, position and focal length).
- 3D modelling is performed with the use of SfM to reconstruct a 3D scene.

The standard match-moving process was strongly modified by the authors in order to boost it for a more efficient use in Heritage. Indeed, the objective to reach is different from the original match-moving process which aims to correctly insert an object in a video. In this research the purpose is extracting images from historical videos in a way which is suitable for the photogrammetric procedure. To do that, it is necessary to know how the camera moved to shoot the video because this dramatically influences the results of the photogrammetric reconstruction. With this aim, the innovations introduced in the proposed workflow are: (i) the use of Artificial Intelligence object detection algorithms as feature tracking method that, as it emerged from the state of the art study, is a development respect on previous studies; (ii) the algorithm of camera tracking that represents the originality of this paper (highlighted in red in Figure 1) and it will be explained in much detail in section 3.3; (iii) the open source SfM algorithms and the metric quality evaluation that certifies the quality of the 3D reconstruction.

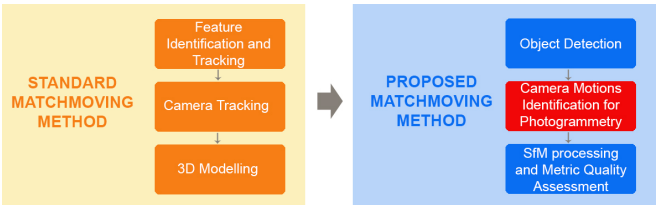

Figure 1. Workflow of the standard match-moving method compared with the method proposed in this paper.

3.2 Object detection with Neural Networks

The first step of the workflow was to identify and track features from the video. This was performed using an object detection Neural Network trained to automatically recognise the monument in the film footage.

In object detection the searched object is detected segmenting a region of interest, is classified putting a bounding box around it and assigning a label with the name of the corresponding class. To track the object in the video sequence, the evolution of the position of the bounding boxes over time is analysed in order to precisely locate the object.

Object detection is a good solution in applications like monument recognition in film footage because it allows the tracking of the object also when the image is noisy, the camera is not stable and the object is with a complex structure (Parekh et al., 2014).

The choice of the Neural Network, the training and the validation phases were widely explained in a previous paper of the authors (Condorelli et al., 2019) in which a workflow capable of automatically detecting architectural heritage in film footage was detailed. This workflow allows to extract the frames containing the architecture but they were subsequently filtered to be processed with photogrammetry in a manual way. In the next section it will be explained how this step was automatized.

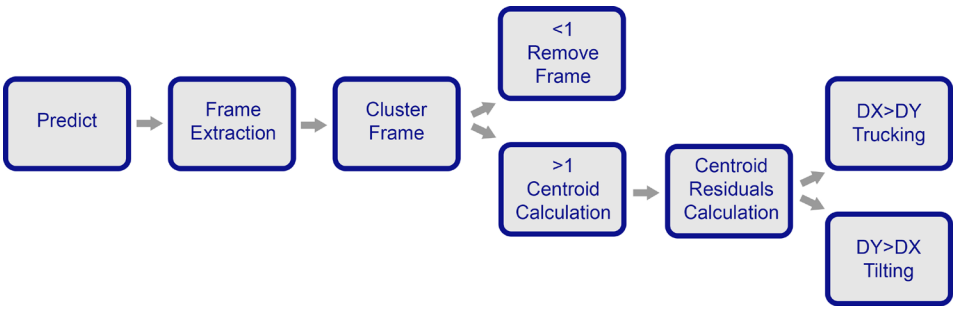

Figure 2. Workflow of the second step of the proposed match-moving method: the tilting and trucking camera motions identification suitable for photogrammetry.

3.3 Camera motions identification for photogrammetry

In the second stage of the workflow, the frames suitable to be processed with photogrammetry are selected from all the frames detected by the Neural Network. The selection is performed according to the camera motions within the scene of the video. Only the shots taken from multiple points of view of the same scene are suitable for the photogrammetric process and the tilting and trucking camera motions have demonstrated to be more effective to perform this kind of frame selection (Condorelli and Rinaudo, 2019).

The algorithm to determine the camera motion starting from the results of the Neural Network object detection is sketched in Figure 2 and detailed below:

1. *Predict*: the object detection algorithm ends with the predict step that outputs the list of frames where the searched monument appears. Each frame is uniquely identified by the video name and the time appearance within the video. The object detection also returns the coordinates of the position of the bounding boxes in the frame and the probability score of the presence of the monument in the video. Considering a predefined probability threshold, positive frames but with a lower score than the threshold are removed from the list of predicted items.
2. *Image extraction*: from the results of the previous step, the frames are extracted from the video and saved as separate images. It is worth noting that the photogrammetric procedure may start from this set of images, but high failure rates are expected when using images extracted from different videos (with different features and qualities), scenes and camera motions at the same time. This usually requires manual intervention and decisions to achieve a final successful photogrammetric reconstruction. In order to automate a successful procedure, a further elaboration of extracted frames is proposed.
3. *Frame clustering*: the extracted frames are grouped in two different splitting criteria, both aim to detect a change of the scene in the video. In particular, a new frame cluster is created if at least one of the following criteria is met:
    a. The first one is time-based: frames which are consecutive in the time-line belong to the same group.
    b. The second one relies on a structural similarity comparison (Wang et al., 2004; Avanaki 2009). Looping over the selected frames, a similarity score between 0 and 1 is evaluated for each frame compared to the previous one. If the score is less than a predefined threshold Ts, a new group is created for the analyzed frame.
4. *Cluster cleaning*: since the intention is to perform photogrammetry on each cluster of frames, clusters with only one image are marked as invalid. The use of frames belonging to the same cluster of intervals brings a higher success rate of the photogrammetric process. In fact, putting together single frames taken from different time intervals or from different videos can certainly help to recover more information about the lost heritage but at the same time there is a higher risk that the photogrammetric process fails because they are dated in different historical time periods.
5. *Bounding boxes centroid calculation*: for each frame in a valid cluster, the centroid *[xc,yc]* of the bounding box is computed in order to analyse the position change of the object.
6. *Centroid residuals calculation*: for each frame cluster, the cumulative residuals between the frame centroids are evaluated using the first and last frame centroids of the cluster:

$$Dx = xc(last\ frame) - xc(first\ frame)$$
$$Dy = yc(last\ frame) - yc(first\ frame)$$

7. *Camera motion estimation*: These residuals are used to guess the camera motion. Clearly, the detection of centroid movements is not sufficient to accurately evaluate the camera motion. However, simple assumptions can lead to results which are satisfactory for the success of the entire proposed procedure. In particular, when $|Dx| > |Dy|$ the trucking camera motion is expected while camera tilting corresponds to the $|Dx| < |Dy|$ case. This simple assumption may lead to wrong results for common cases. If both Dx and Dy are very small, this could correspond to a fixed camera poorly anchored to the terrain (a tripod was not used for most historical film footage). On the other hand, if Dx and Dy are not small but close to each other, camera motion guessing is more questionable. For these reasons, the proposed algorithm uses can be summarized in the following Figure 3.

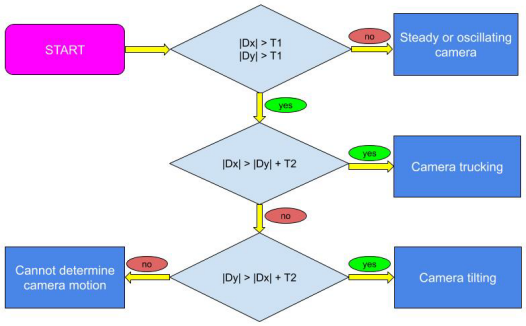

Figure 3. Workflow of the camera motion estimation algorithm.

where T1 and T2 are two thresholds. In the end, the devised algorithm allows the user to distinguish among four camera motion categories, namely: “steady or oscillating camera”, “camera trucking”, “camera tilting” and “cannot determine camera motion”. This simple categorization is however useful in view of the final purpose of the algorithm that is to detect frame clusters useful for the photogrammetry.

As discussed above, some of the steps described lead to automatic data filtering. Being an automatic procedure means that the filtering can lead to the elimination of potentially useful data for the final process. Decreasing the introduced thresholds limits the loss of data, but puts at risk the success of the automatic procedure, at least in some cases. Depending on the context of use, it is necessary to evaluate the choices to make. Overall, the choice to go towards an automatic procedure leads to a dramatic improvement in the efficiency of the process both in terms of time and in terms of simplicity of execution. In these cases, a certain penalty in terms of the ability to exploit the single data should be acceptable.

3.4 3D modelling: SfM processing and metric quality assessment

The last step concerned the photogrammetric reconstruction of the Heritage with open source algorithms and the metric quality assessment of the model.
As reference software for photogrammetric processing was chosen COLMAP (Schönberger et al., 2016), an opensource Structure-from-Motion and Multi-View Stereo (MVS) algorithm implementation, developed by ETH of Zurich, (COLMAP, Johannes L. Schoenberger, 2019). During the process specific feature points are manually selected in order to guarantee their presence in the final point cloud. This step will be very useful during the evaluation of the metric quality of the results. The detection and extraction of these feature points is performed importing in COLMAP a text file in which the image coordinates (x, y) expressed in pixel and the scale and orientation information are indicated (Condorelli et al., 2019).
The results of the 3D reconstruction of the Heritage were compared with a benchmark specifically created to evaluate the metric quality of the model according to the type of camera motion used (Condorelli and Rinaudo, 2019). The assessment was completed with the scale of the model through the feature points selected during the photogrammetric process and the comparison with existing material from which extract metric information: a point cloud, if present or historical drawings, for example. In both cases the presence of specific feature points in both point cloud resulted from the process and the existing material is necessary for the metric comparison and scale.

4. RESULTS AND DISCUSSION

4.1 Case studies and materials

In order to test the workflow, two case-studies in Paris were chosen: the UNESCO Heritage Tour Saint Jacques and the pavilions of Les Halles of Baltard. These case studies represent two different situations of Heritage because the tower was transformed over time but still exists (Figure 4) and the pavilions were destroyed in 1971 (Figure 5). Thus, it is possible to compare the different results obtained from the implementation of the workflow to the two case studies. The methodology and the quality of the results were analysed, with particular focus on the camera tracking and the effect of the source of the images used for the evaluation of the metric quality phases.

After a deep consultation of historical archives in Paris (Lobster, Ina.fr, CNC, Forum des Images, Les Documents Cinematographique), numerous video documents were collected by the authors. In this historical film footage both the monuments were shot in different situations and with different techniques and motions of the camera. Therefore, they represent good case studies and were used to test the proposed match-moving method. In addition, historical photographs, drawings, images and design projects have been collected from archives, as well as a 3D model, from a recent photogrammetric survey of the existing tower carried out by Iconem in 2015. Together these materials were used in the last phase of metric evaluation of the results.

Figure 4. Tour Saint Jacques in a historical photograph by Merville and in the actual state.

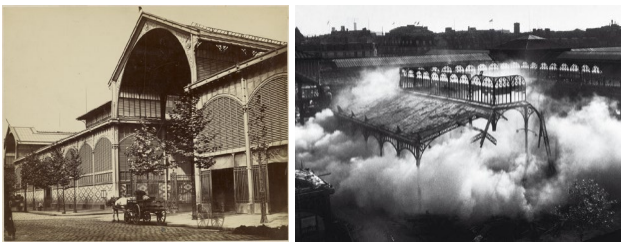

Figure 5. Les Halles of Baltard before and during demolitions in 1971.

4.2 Results of object detection with Neural Networks

The object detection algorithm achieved the recognition of the correct frame containing the searched monument, both in the case of the tower and the pavilions. Despite the presence of both false positives and false negatives in the final results, the evaluation of the performance of the Neural Networks for the automatic detection of the monuments has been substantially successful in terms of saving time and efficiency for the end user over a manual search. A full explanation and discussion of quantitative performance evaluation is provided in Condorelli et al., 2019.

The processing of the videos was performed thanks to the use of the High-Performance Computing resources by CINECA that benefits greatly from the use of Graphical Processing Units (GPUs).

In Figure 6 and Figure 7 some of the results of the detection are shown.

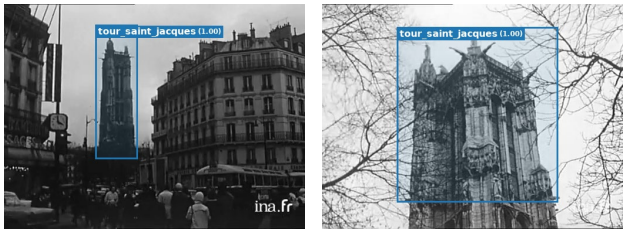

Figure 6. Example of frames with the Tour Saint Jacques correctly detected in the film footage.

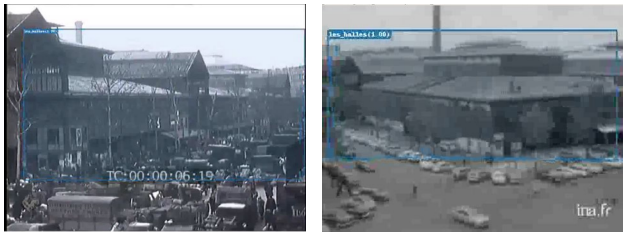

Figure 7. Example of frames with Les Halles correctly detected in the film footage.

4.3 Results of camera motions identification

In order to automatically select the frame sets to be used for photogrammetry, the algorithm for camera motion identification described in Section 3.3 was employed. The procedure was experimented on three different film footage, namely: "Tour Saint Jacques" from Ina.fr archive, "Études sur Paris" from the CNC-VOD archive and "La Destruction des Halles de Paris" from Les Documents Cinematographiques archive.
As introduced, the algorithm requires a proper selection of some parameters. First of all, for the algorithm step 3 (frame clustering), the second splitting criterion requires to adopt a structural similarity threshold Ts. From visual inspection of the extracted frames, it turned out that setting Ts=0.1 is an effective choice to detect a change of the scene in the great majority of the analysed cases. As for the residual thresholds T1 and T2, a selection is not straightforward and a tuning based on tentative results seems to be a viable strategy.
In Table 1, the accuracy results for three different choices of (T1,T2), namely (0,0), (10,10), (20,20) are shown. In order to evaluate the accuracy, for each frame cluster the identified camera motion is compared against the real camera motion considering the 4 output categories, i.e. "steady or oscillating camera", "camera trucking", "camera tilting" and "cannot determine camera motion". The accuracy is evaluated as the ratio between the number of correctly identified frame cluster motions and the total number of analysed frame clusters.

To allow a more detailed investigation of the results, the accuracy values are reported for each video separately. Moreover, for each video, two accuracy values are provided. The first one (TOT) is based on all the detected frame clusters while the second one (NNTP) only includes frame clusters which really represent the searched object. Indeed, since the camera motion algorithm is applied to the results of Neural Network (NN) object detection, some extracted frames are False Positive NN results, i.e. they do not correspond to the searched object. Since NN False Positives may correspond to other objects, or also to completely wrong image detections, it is expected that the camera motion algorithm will work more smoothly when filtering out these bad cases. In any case, summarizing, the TOT accuracy summarizes both NN and camera motion estimation accuracies, while NNTP accuracy more strictly refers to the camera motion algorithm accuracy.

From Table 1 results, it turns out that considering T1=T2=0, the TOT accuracy is poor (below 30%) whereas the NNTP accuracy is optimal for two of the videos but very poor for the third one. Increasing the thresholds to T1=T2=10, the average accuracies are both greater or equal to 80%. Setting T1=T2=20, there is an accuracy degradation, especially for one of the videos. Concluding, the intermediate setup T1=T2=10 seems to be the best choice. All in all, it is clear that the number of analyzed cases is not enough to demonstrate the effectiveness of the algorithm and of the entire workflow in a general context. However, the discussed preliminary results are encouraging. Not only the overall accuracy values are satisfactory, but the most suitable videos for photogrammetry are correctly identified and this means that, at least for the considered videos, the described workflow allows the user to reach the final goal which is identifying videos for photogrammetric reconstruction.

4.4 3D modelling: SfM processing and metric quality assessment

From the previous analysis concerning camera motion identification the frames correctly selected by the algorithm and suitable for photogrammetry appeared in two different films. One is "Études sur Paris", dated 1928, that was shot with the tilting camera motion and contains sequences of the Tour Saint Jacques (Figure 8). The technical features of the film are: gauge 16 mm; focal length 25 mm; digital format resolution 480x360 pixels; black and white.

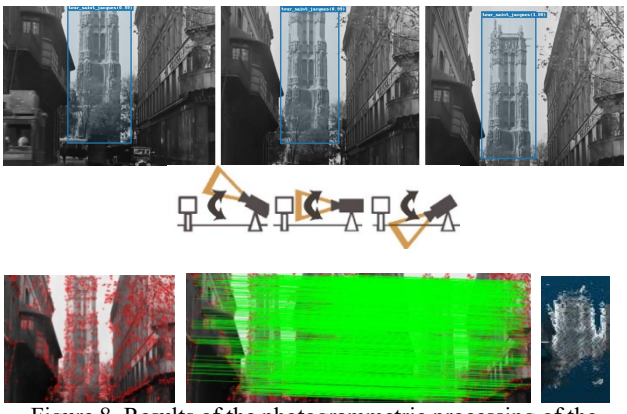
Figure 8. Results of the photogrammetric processing of the frames from the film footage "Études sur Paris" shot with tilting camera motion.

Film	N° cluster	Thresholds 0.0 0.0		Thresholds 10.0 10.0		Thresholds 20.0 20.0	
		TOT [%]	NNTP [%]	TOT [%]	NNTP [%]	TOT [%]	NNTP [%]
Tour Saint Jacques	57	13	12	72	60	77	68
Études sur Paris	112	26	100	72	100	73	100
La Destruction des Halles de Paris	11	50	100	92	80	66	20

Table 1. Accuracy values expressed in percentage according to different clusters and thresholds.

The second film is "La Destruction des Halles de Paris", dated 1971, and took Les Halles with trucking camera motion (Figure 9). The technical features of the film are: gauge 35 mm; focal length 35 mm; digital format resolution 492x360 pixels; black and white.

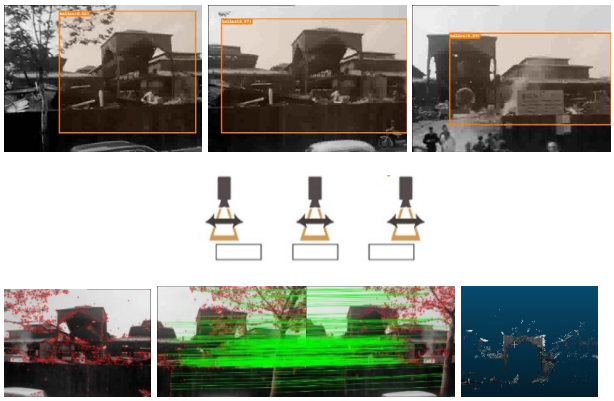

Figure 9. Results of the photogrammetric processing of the frames from the film footage "La Destruction des Halles de Paris" shot with trucking camera motion.

As explained in section 3, in order to certify the quality of the 3D models resulted from the photogrammetric process, a comparison with a benchmark previously created was performed. Firstly the Mean and the Standard Deviation of the values of residuals of the reprojection errors computed in all images, were calculated for both case studies and compared with the same values calculated for the benchmark.

For the tower case studies was considered a distance of 15 m and for the pavilion of 120 m and compared with the corresponding camera motions benchmark. Considering that the GSD for the tilting benchmark was 1.2 cm/px, the GSD for the tower is 1.43 cm/px; while for the trucking benchmark the GSD is 11.2 cm/px and the GSD for Les Halles of 23.6 cm/px. All the results are reported in Tables 2 and 3.

From Table 2 it is clear that comparing the two results, the differences between the values of the case of the tower and the benchmark are not significant. On the other hand, in Table 3 the values in pixels are almost the same for Les Halles, but after transformation into centimetres by the GSD calculation, the values are higher than the benchmark. These differences could result from the approximation of the focal length of the camera used for filming and influenced by the taking distance. Therefore, an error margin must be taken into account in this evaluation.

Secondly, the trend of all the values of residuals was analysed to check if the curve follows the Gaussian distribution like for the benchmark for both case studies. The results are shown in Figure 10 and Figure 11.

As the last step of the metric quality evaluation the scale and the analysis of the point clouds obtained from the photogrammetric process were performed in two different ways.

For the case of the tower the resulted sparse point cloud was compared with the existing dense point cloud calculating distances between the two models. The results showed that they are less than one pixel (Figure 12).

Case	Mean		St Dev		Min Residual		Max Residual	
	px	cm	px	cm	px	cm	px	cm
Bench-mark	0.36	0.10	0.10	0.80	0.13	0.10	0.60	0.80
Case study	0.23	0.33	0.06	0.08	0.10	0.14	0.35	0.50

Table 2. Mean, Standard Deviation, Min and Max values of residuals for the Tour Saint Jacques compared with the tilting benchmark.

Case	Mean		St Dev		Min Residual		Max Residual	
	px	cm	px	cm	px	cm	px	cm
Bench-mark	0.47	1.10	5.20	0.77	0.13	8.70	1.40	0.10
Case study	0.50	4.20	11.8	0.79	0.10	18.6	2.30	0.18

Table 3. Mean, Standard Deviation, Min and Max values of residuals for Les Halles compared with the trucking benchmark.

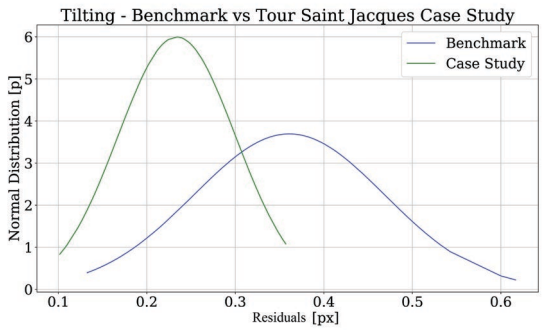

Figure 10. Trends of residual values for the Tour Saint Jacques compared with the tilting benchmark.

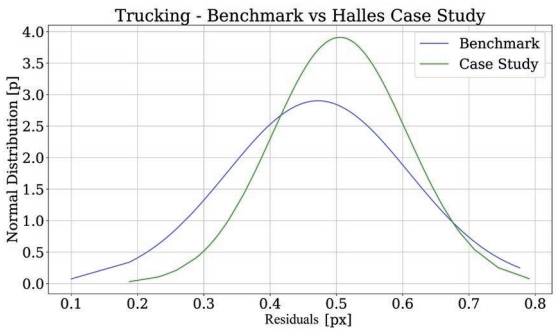

Figure 11. Trends of residual values for Les Halles compared with the trucking benchmark.

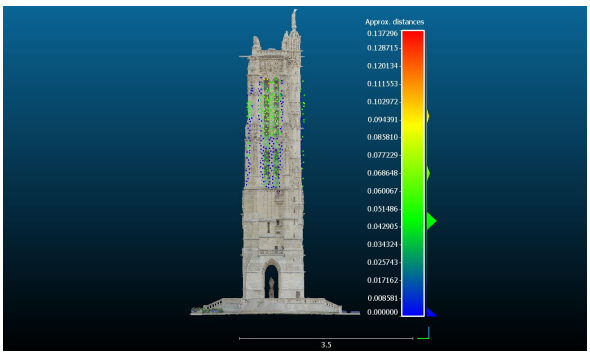

Figure 12. Metric comparison between point cloud of the tower obtained from film footage and from a recent survey.

For the case of Les Halles specific feature points corresponding to points of interest in project drawing (Figure 13) were manually selected during the photogrammetric process. Thanks to the presence of these feature points in the final point cloud (Figure 14), it was possible to scale the model using the distance AB (highlighted in green in Figure 13 and 14). Moreover, a metric evaluation was performed calculating residuals between the distances from drawing, considered as the reference, and the final point cloud. The results are shown in Table 4 and it is noted that, although some limitations due to the quality and the reliability of the project drawings, probably different from the real building, the values of residuals are acceptable.

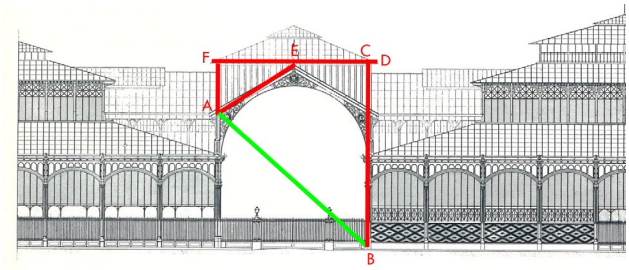

Figure 13. Original drawing of Les Halles by Baltard.

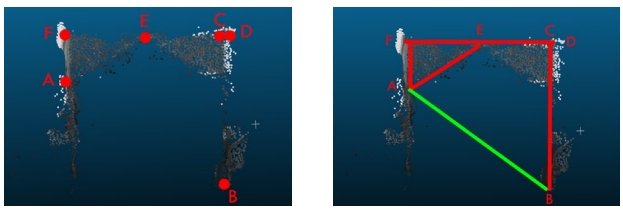

Figure 14. Feature points and distances in the final point cloud of the pavilion of Les Halles.

Distance	Drawing [m]	Point Cloud [m]	Residuals [m]
BC	19.50	18.87	0.63
DF	17	17.19	-0.19
AF	5	4.59	0.41
AE	9.50	9.58	-0.08

Table 4. Residuals calculation between distances from the drawing and the point cloud of Les Halles.

5. CONCLUSIONS AND PERSPECTIVES

The present study was designed to experiment the use of a new match-moving workflow devised to automatically select suitable frames for photogrammetric processing in film footage, in order to document lost heritage.

In the first stage of the workflow, the use of object detection was demonstrated as a good solution for the automatic recognition of architectural heritage in historical videos especially because it can extract the coordinates of the bounding boxes which locate the monument in the frame.

Following the evolution of NN bounding boxes, the second stage of the workflow identifies the camera motions in significant categories. In particular, this strategy is used to detect tilting and trucking camera motions, which are suitable for photogrammetry and are very common in historical videos. Due to its structure, the algorithm can work even when the track is very short. The video frames are first grouped in frame clusters according to image similarity criteria. Then, for each frame cluster, the camera motion category is evaluated. This is particularly useful in view of photogrammetry where each frame cluster can be used separately to increase the procedure positive completion. Indeed photogrammetric reconstruction is expected to fail when images from different scenes or videos are used in the same process.

The proposed algorithm includes some parameters that, properly selected, allow the attainment of an overall accuracy of the camera motion identification up to 80%. This accuracy is achieved by minimizing the misinterpretation of camera oscillations, due to the low quality of the camera used to shoot historical film footage and the absence of a tripod, and by properly setting the image similarity criterion used to group the frames.

The findings of this study are very encouraging. Actually there are intrinsic challenges when working on historical videos such as the lack of important information about the camera, the quality of the film used to shoot the videos, and the unavailability of a precise metric reference when the monument is lost. The preliminary results shown in this paper prove that the presented automatic workflow can be effective even in these critical conditions.

There are many aspects which can be further investigated such as adopting other types of Neural Network model as well as tuning and improving the camera motion algorithm. From the end-user perspective, in order to fully exploit the automatic procedure, developing a graphic interface for the entire pipeline would be important to improve the usability of the workflow.

There are several areas where this study makes an original contribution, notably in the field of Cultural Heritage as it automates a task that is manual, and boosts the photogrammetric process by selecting suitable material for the application. This could offer some important insights into the management and organization of historical information and the protection of the memory of the past. Moreover, the same methodology could be applied to recent videos and allow access and recording of information concerning daily life. For this reason, further research could evaluate the effectiveness of the experimental methodology in other fields, such as UAV, structural analysis and computer vision, for example.

ACKNOWLEDGEMENTS

The authors express thankfulness to the archive Lobster Films for sharing footage used in this research, to CNC, Forum des Images, Ina.fr, Les Documents Cinematographiques, 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 has been carried out under the GAMHer project: Geomatics Data Acquisition and Management for Landscape and Built Heritage in a European Perspective, PRIN: Progetti di Ricerca di Rilevante Interesse Nazionale – Bando 2015, Prot. 2015HJLS7E.

REFERENCES

ACTS, Zhang, G., Dong, Z., Jia, J., Wan, L., Wong, T., Bao, H., 2009. ACTS: Automatic Camera Tracking System. http://www.zjucvg.net/acts/acts.html (1 May 2020).

Avanaki, A.N., 2009: Exact global histogram specification optimized for structural similarity. *Optical Review*, 16, 613-621, arXiv.org/0901.0065, doi.org/10.1007/s10043-009-0119-z.

Boujou, Vicon Motion Systems Ltd UK. http://www.vicon.com/software/boujou/ (1 May 2020).

Chapel, M.N., Bouwmans, T., 2020: Moving Objects Detection with a Moving Camera: A Comprehensive Review. *Computer Vision and Pattern Recognition*, arxiv.org/abs/2001.05238.

COLMAP, Johannes L. Schoenberger, 2019. COLMAP - Structure-From-Motion and Multi-View Stereo. https://github.com/colmap/colmap (1 May 2020).

Condell, J., Moore, G., 2006: Software and Methods for Motion Capture and Tracking in Animation. *Proceedings of the 2006 International Conference on Computer Graphics & Virtual Reality, CGVR 2006*, CSREA Press 2006, 3-9.

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

Condorelli, F., Rinaudo, F., Salvadore, F., and Tagliaventi, S., 2019: Architectural heritage recognition in historical film footage using Neural Networks. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W15, 343–350, doi.org/10.5194/isprs-archives-XLII-2-W15-343-2019, 2019.

Condorelli, F., Higuchi, R., Nasu, S., Rinaudo, F., and Sugawara, H., 2019: Improving performance of feature extraction in SfM algorithms for 3D sparse point cloud. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLII-2/W17, 101–106, doi.org/10.5194/isprs-archives-XLII-2-W17-101-2019.

Dağlar, O., Tong, T., 2011: A Method on Using Video in Architectural Design Process: Matchmoving. *Respecting Fragile Places: 29th eCAADe Conference Proceedings*, 339-348.

Delis, P., Zacharek, M., Wierzbicki, D., Grochala, A., 2017: Point Cloud derived from video frames: accuracy assessment in relation to terrestrial laser scanning and digital camera data. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* XLII-2/W3, 217-223, doi.org/10.5194/isprs-archives-XLII-2-W3-217-2017.

Dong, Z., Zhang, G., Jia, J., Bao, H., 2009: Keyframe-Based Real-Time Camera Tracking. *IEEE International Conference on Computer Vision (ICCV)*, 118,1538 - 1545 doi.org.0.1109/ICCV.2009.5459273.

Haji, K., Sharif, A. P., Rabbani, I. A., 2016: An Overview of Matchmoving using Structure from Motion Methods. *Proceedings of the 2016 Symposium on Digital Production*, 45-54, doi.10.1145/2947688.2947697.

Ingwer, P., Gassen, F., Püst, S., Duhn, M., Schälicke, M., Müller, K., Ruhm, H., Rettig, J., Hasche, E., Fischer, E., Creutzburg, R., 2015: Practical usefulness of structure from motion (SfM) point clouds obtained from different consumer cameras. *Proc. SPIE 9411, Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications 2015*, 941102, doi: 10.1117/12.2074892.

Lee, B., Park, J., Young Sung, M., 2006: Vision-Based Real-Time Camera Matchmoving with a Known Marker. *Proceedings of the 5th international conference on Entertainment Computing*, 193-204, doi.10.1007/11872320_23.

Lourakis, M., Argyros, A., 2005: Camera Matchmoving in Unprepared, Unknown Environments. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, doi.org10.1109/CVPR.2005.96.

Luo, A., Chen, S., Tseng, K., 2016. A real-time camera match-moving method for virtual-real synthesis image composition using temporal depth fusion. *International Conference on Optoelectronics and Image Processing (ICOIP)*, 35-39, doi.org. 10.1109/OPTIP.2016.7528515.

Mancini, M. F., Moscarelli, A., Mulla, E., 2013: From video sequence as a database for the generation of 3D models to video as a tool for architecture communication. *Heritage 2013 Monitoring Conservation Management*.

Parekh, H., Thakore, D., Jaliya, K., 2014: A Survey on Object Detection and Tracking Methods. *International Journal of Innovative Research in Computer and Communication*, 2(2), 2979-2978.

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

SynthEyes, Andersson Technologies LLC. https://www.ssontech.com/ (1 May 2020).

Toda, T., Masuyama, G., Umeda, K., 2016: Detecting Moving Objects Using Optical Flow with a Moving Stereo Camera. *MOBIQUITOUS 2016: Adjunct Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing Networking and Services*, 35–40, doi.org/10.1145/3004010.3004016.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13, 600-612, doi.org/10.1109/TIP.2003.81986.

Zhang, G., Qin, X., Hua, W., Wong, T., Heng P., Bao, H., 2007: Robust Metric Reconstruction from Challenging Video Sequences. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-8.

Zhang, G., Dong, Z., Jia, J., Wan, L., Wong, T., Bao, H., 2009: Refilming with Depth-Inferred Videos. *IEEE Transactions on Visualization and Computer Graphics*, 15(5), 828-840.

Zhang, G., Liu, H., Dong, Z., Jia, J., Wong, T., Bao, H., 2016: Efficient Non-Consecutive Feature Tracking for Robust Structure-From-Motion. *IEEE Transactions on image processing*, 25(12), 422-435, doi.org.10.1109/TIP.2016.2607425.