An approach for precise 2D/3D semantic annotation of spatially-oriented images for in situ visualization applications

A. Manuel, E. Gattet, L. De Luca
UMR 3495 MAP
CNRS/MCC
Marseille, France
(adeline.manuel, eloi.gattet, livio.deluca)@map.archi.fr

P. Veron
UMR CNRS 7296 LSIS
CNRS
Aix-en-Provence, France
philippe.veron@ensam.eu

Abstract—**Thanks to nowadays technologies, innovative tools afford to increase our knowledge of historic monuments, in the field of preservation and valuation of cultural heritage. These tools are aimed to help experts to create, enrich and share information on historical buildings.**

Among the various documentary sources, photographs contain a high level of details about shapes and colors. With the development of image analysis and image-based-modeling techniques, large sets of images can be spatially oriented towards a digital mock-up. For these reasons, digital photographs prove to be an easy to use, affordable and flexible support, for heritage documentation.

This article presents, in a first step, an approach for 2D/3D semantic annotations in a set of spatially-oriented photographs (whose positions and orientations in space are automatically estimated). In a second step, we will focus on a method for displaying those annotations on new images acquired by mobile devices in situ.

Firstly, an automated image-based reconstruction method produces 3D information (specifically 3D coordinates) by processing a large images set. Then, images are semantically annotated and a process uses the previously generated 3D information inherent to images for the annotations transfer. As a consequence, this protocol provides a simple way to finely annotate a large quantity of images at once instead of one by one. As those images annotations are directly inherent to 3D information, they can be stored as 3D files. To bring up on screen the information related to a building, the user takes a picture in situ. An image processing method allows estimating the orientation parameters of this new photograph inside the already oriented large images base. Then the annotations can be precisely projected on the oriented picture and send back to the user.

In this way a continuity of information could be established from the initial acquisition to the in situ visualization.

Index Terms—**Semantic annotations, photogrammetry, image processing, dense image matching**

I. INTRODUCTION

Nowadays in the field of documentation of the architectural heritage, innovative tools afford to increase our knowledge of historic monuments, thanks to new technologies. These tools are aimed to help experts to create, enrich and share information on historical buildings.

Among the various iconographic sources, photographs constitute an important value for architectural studies. They testify to the state of preservation of a building at a specific time. As the digital imaging technologies have been significantly improved in recent years, the production and the processing of huge amount of high-resolution digital photographs becomes easy today. Furthermore, photographs contain a high level of information about shapes (geometry, dimensions, visual appearance, etc). All of these reasons suggest that photographs are a very promising support for digital documentation.

As a technical support photographs can be, partially or entirely, semantically described with keywords or ontologies. Some analyzes (typify surfaces, observing the state of preservation, etc) can be directly perform on them. As well with the development of automated image-based-modeling techniques, large sets of images can be spatially oriented towards a digital 3D scene.

Thus, photography appears as a great way to document the morphology and the state of preservation of buildings. However an exhaustive documentation of a building needs a large quantity of images.

This paper has been divided into seven parts; Section 2 presents some works about the annotation of images; Section 3 broaches the general approach of this work; Section 4, 5 and 6 describe the computer implementation of the approach. Finally, section 7 will evaluate the system and bring some perspectives for further developments.

II. RELATED WORK

In the domain of cultural heritage, the process of annotation on iconographic sources and more specifically on photographs helps the comprehension of a building by informing semantic information.

Three main methods are currently available to annotate a 2D source: manual, automatic, or semi-automatic. Manual annotations are defined by the user on images one by one by using either ontologies [1] or keywords [2]. Automatic

annotations are based on an analysis of image content by means of a segmentation followed by shape recognition [3]. Semi-automatic annotations merge manual methods and automatic ones and the user takes part in the validation of keywords [4-5]. These three methods only use 2D information.

Regarding 3D models annotations, information can be attached to points [6], segments, surfaces [7] or objects [8] in the digital mock-up.

Recently, researchers have shown an increased interest in the use of 3D information in the image annotation process. Phototourism [9] uses a "light" point cloud (SIFT points) to transfer annotations between images. But only rectangles can be transferred. Other works [10] take advantage of the 3D model. Annotations are attached to a segmented digital mock-up and the images are oriented towards this model. By using the point of view of an image, annotations of the 3D model can be projected on the image. Still with this process, annotations cannot be defined directly on images.

In the light of these works, the process of images annotations could be significantly improved by connecting iconographic sources to a 3D representation of the building. However, the use of the relation between the spatially-oriented set of images and the 3D model is just at the beginning of its possibilities.

Visualizing annotations on a picture taken in situ means to combine a virtual object with real world information. R. Azuma [11] introduced augmented reality (AR) techniques as systems following three characteristics:

1) Combines real and virtual
2) Interactive in real time
3) Registered in 3-D

To fit a 3D model on the real world picture, the three orientations and three position parameters, representing the camera localization towards the 3D world, need to be known. Two main approaches exist: either with a 2D marker set in the real world and detected in a video stream or image, or using a combination of sensors such as compass, gyroscope, GPS.

2D markers, such as simple tags [12] or templates, allows, via a pattern extraction, to compute the deformation of the marker. This allows the software to find the relative orientation between the camera and the real world. As the marker can be finely detected, the fitting of the 3D model is precise. But, it is inconceivable to cover a whole historical building with markers.

Markerless solutions use the mobile device's built-in GPS to get a real world to 3D position and compass to find the orientation. Nevertheless, in normal conditions, GPS data are relatively imprecise and prove to be insufficient to finely fit the virtual image on the real world picture.

Taken together, those observations show that real time AR systems are either precise or markerless. As the images annotation process allows to define precise annotations, they need to be finely lined up over the images, in the aim for the user to understand the architectural details.

However, as the real time visualization is not the predominant factor in some application context (such as the accurate annotation of surfaces regions for analysis preservation purposes), an image-based orientation process such as special resection or structure from motion should be used.

III. MAIN APPROACH

The aim of this research is to develop a process for automatically distributing 2D/3D semantic annotations in a set of spatially-oriented photographs and to display those annotations on new images acquired by mobile devices (e.g. tablets, smartphones, etc).

Our approach consists in five main steps (Fig.1).

First of all, hundreds of photographs are acquired in order to cover all views of the building (Fig.1 step 1). Quantity depends on the building's size. This large set of images is then processed with a method for automated orientation enabling to generate point clouds inherent to images (Fig.1 step 2). This processing step is described in section IV. The large set of orientated images is stored as an image stock-reference base for further orientations.
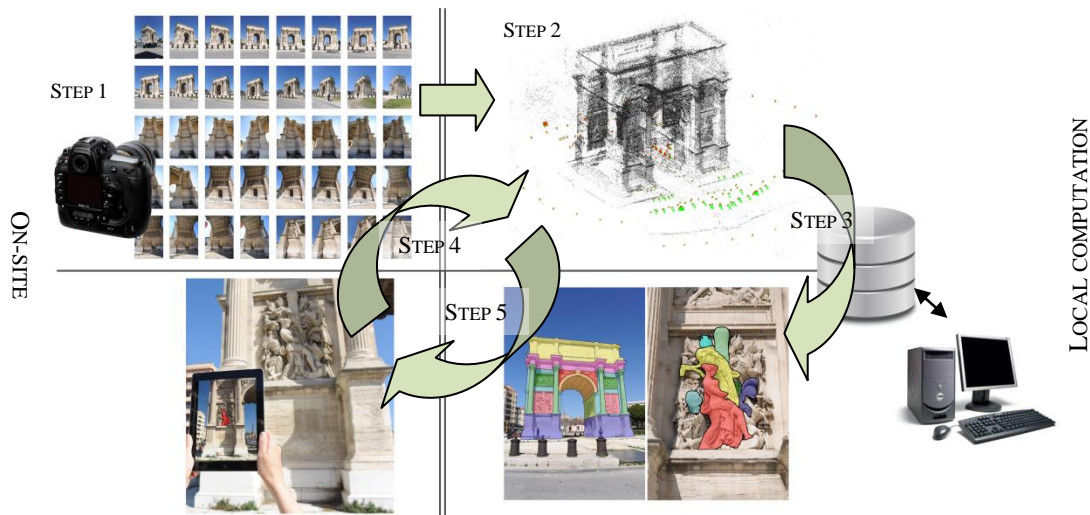

Fig. 1: Main approach: Step 1: acquisition of photographs, Step 2: Automated 3D modeling, Step 3: 2D/3D semantic annotation on photographs, Step 4: orientation of the in situ acquired image inside the already oriented images base, Step 5: projection of semantic annotation on the in situ acquired image

Then, images are precisely segmented and semantically annotated (Fig.1 step 3). In view of the large quantity of acquired images for an exhaustive documentation of a building, a manual annotation of all images one by one is not conceivable. For this reason, a process for distributing/propagating annotations towards the set of images (using image's geometric and projective properties estimated in the previous step) assists the definition of semantic annotation by using 3D information (specifically, 3D coordinates from the previously generated point cloud). As those images' annotations are directly inherent to 3D information, they can be stored as 3D point-based representations. The principle of annotations transfer is described in section V.

Finally, so as to visualize in situ the information related to a building, a new picture is acquired with a mobile device. This picture is then oriented inside the already oriented image database (Fig.1 step 4). Basing on the estimated orientation of this new picture and on the segmentation of the first set of images, semantic annotations can be precisely projected on the new picture (Fig.1 step 5). The in situ visualization of annotations is described in section VI.

This paper will mainly focus on geometric issues of the described approach, time and memory consuming issues will not be resolved in this article.

IV. ACQUISITION AND IMAGE PROCESSING

The core of our approach is based on a set of tools and procedures concerning the on-going development of an automated image-based 3D reconstruction method [13-14]. The process consists of an automated calibration and orientation of images, a very dense multi-view correlation and a point cloud generation.

A. Acquisition of photographs

The experimentations were carried out on the Porte d'Aix historical building in Marseille. This place is small enough to allow a quick, dense and full photographic acquisition and wide enough to provide different levels of segmentation for the semantic annotation.

250 photographs were taken with a professional digital camera and three different focal lengths. The goal was to get as most points of view as possible, all around the place, while following the acquisition protocol needed for a correct multi-view stereo correlation.

B. Automated orientation of images

The photographs are computed with the SIFT [15] algorithm, adapted by A. Vedaldi [16] in order to extract common tie points (Fig.2).

As the images were taken without a known-relation between each others, the SIFT detectors are first searched on all pairs of images with a low resolution. Using the results of this first approach, a second higher resolution tie point extraction is conducted on the images.
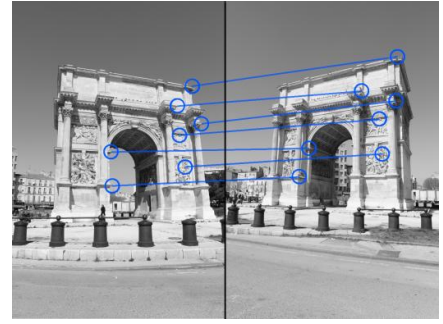

Fig. 2: Detection of tie points

The second processing step creates the global calibration and orientation of the first set of images by estimating internal and external parameters of the camera's geometric model. The calibration can be made once for all, so the mobile device's built-in camera is calibrated at the same time.

Then an initial orientation computation is automatically made by first orienting the two photographs which have the most common points. After that, the software automatically adds one by one all the other images and proceeds a bundle adjustment to refine each image orientation (Fig.3). 250 photographs' computing of orientation led on a recent double quad core CPU took approximately 2.5 hours.
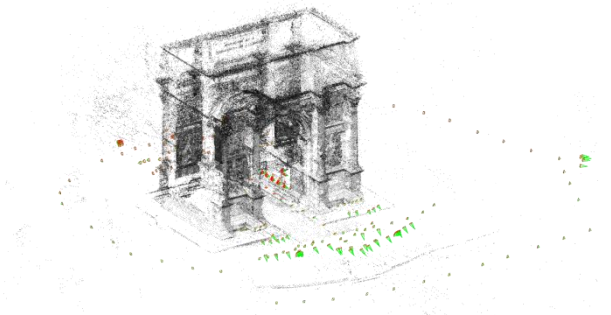

Fig. 3: SIFT points and images orientation

Now that the images are oriented in relation with each other, a dense matching-base surface reconstruction automatic method [13-14] generates, on each most relevant image, a depth map. This operation follows a pyramidal scale level approach (Fig.4), which allows in the end to find for each pixel of an image the corresponding depth. Dense matching on 20 different points of view took approximately 3 hours.

Fig. 4: Pyramidal approach: example of results during the multi-scale matching

Finally each depth map is converted into a 3D metric point cloud by projecting each pixel of the image in space according to the image orientation parameters and pixels depth. Thus a dense point cloud (up to one 3D point for one pixel) is created for each computed image. The superposition of all point clouds allows the creation of the complete point cloud of the building (Fig.5).

Fig. 5: 3D automated image-based reconstruction

While the point clouds are generated, the 3D coordinates of each pixels are stored in a specific TIFF file, called XYZ file.

A TIFF file is an image file that stores color information of each pixel on a matrix of three layers attached each to a primary color (Red, Green, and Blue). Instead of storing color information of pixels, XYZ files use this structure to store the calculated 3D coordinates of pixels. X coordinates are stored on red layer, Y coordinates on green layer and Z coordinates on blue layer. As all pixels of the image have 3D coordinates, each XYZ file has the same size as the attached image.

To read the 3D coordinates of a pixel placed at the row i and the column j in the image, the value of each layer of the XYZ file of the image must be read at the same position. 3D coordinates of a pixel in the image can be so easily extracted from the XYZ file (Fig.6).
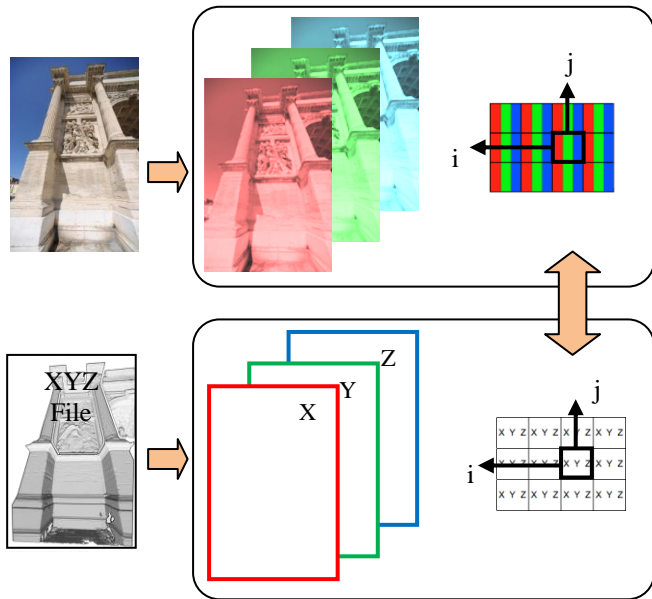

Fig. 6: Structure of a TIFF file and of a XYZ file

Thus an XYZ file allows an inherent relation between each pixel of images and the associated 3D point.

V. ANNOTATIONS

Having the set of oriented photographs and the XYZ files, the process of annotation can be set up.

Looking at the number of needed photographs for an exhaustive documentation, an automatic transfer of annotation towards the set of images must be instated. The adopted methodology for annotations' propagation between images is based on the idea that 3D information could serve as a support for transferring a 2D annotation.

This propagation consists in three steps (Fig.7):
- Definition of the annotation on one image
- Research of 3D coordinates of the annotated area
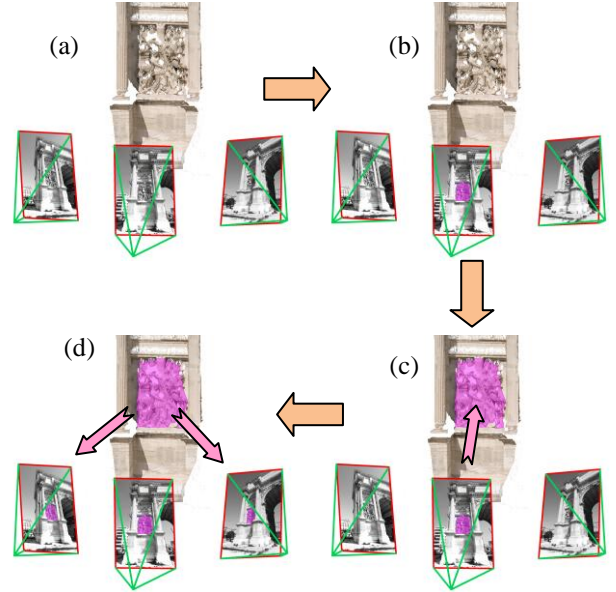- Projection on the other images of the set

Fig. 7: Steps of propagation: (a) oriented set of images and 3D coordinates, (b) annotation on one image, (c) research of 3D coordinates and (d) propagation on other images

With this process, annotations are simultaneously defined by areas on images and by a set of 3D coordinates in space. They are called hybrid annotations.

A. Definition of annotations

Starting from an annotation defined on one image, the corresponding area is used to create a black and white mask. This mask is same-sized as the image and the white areas correspond to the annotated area (Fig.8).
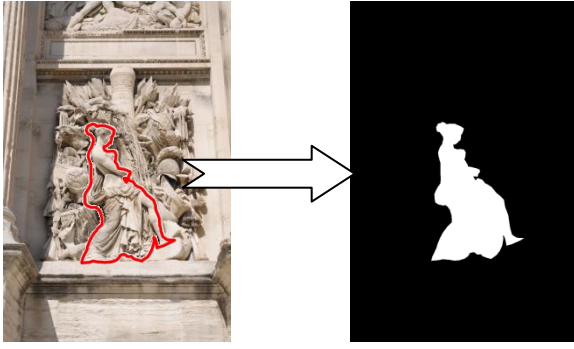
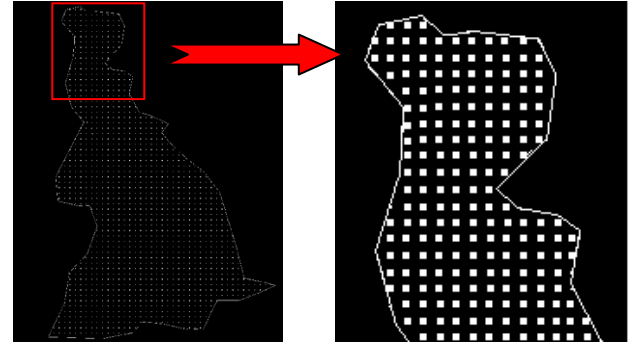Fig. 8: Annotated area (left) and deducted mask (right)

Fig 11: Sampling of pixels in the mask

B. 3D coordinates of annotations

In the mask, white areas are the interesting ones. In order to extract the 3D coordinates of these areas, the position (row and column in the mask) of these pixels must be known. All positions of white pixels are therefore searched. Thus, a list is created, uniting all couples of positions i (row) and j (column). Knowing the concerned positions in the mask, 3D coordinates of the white area can be read at these positions in the XYZ file attached to the image (Fig.9).
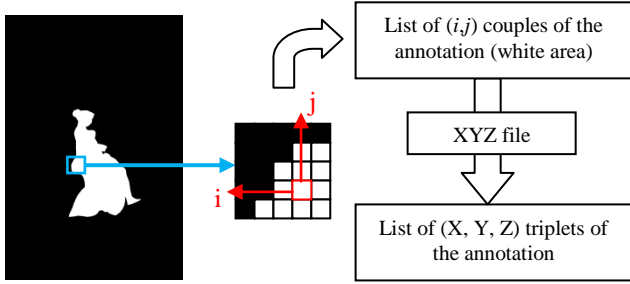

Fig. 9: Search of 3D coordinates of the white areas

The annotated area is so defined in space by its 3D coordinates (Fig.10).

Fig. 10: 3D point-based representation of the annotated area

In this step, a sampling of white pixels can be used in order to limit the number of annotation coordinates. This sampling contains only the pixels of the white areas' contour and of a grid pattern inside the white area (Fig.11).

C. Projection on other images

The 3D coordinates of the annotation must be now retrieved in the XYZ files of the other images of the set.

So as to transfer the annotation on another image of the set, the principle is to compare all 3D coordinates of the annotation (according to a specific image resolution and zoom level) with all 3D coordinates contained in the XYZ file attached to the image. With this comparison, provided by a binary search, positions (row and column) of 3D coordinates in the XYZ file, corresponding to one of the annotation triplet, are extracted (Fig.12).
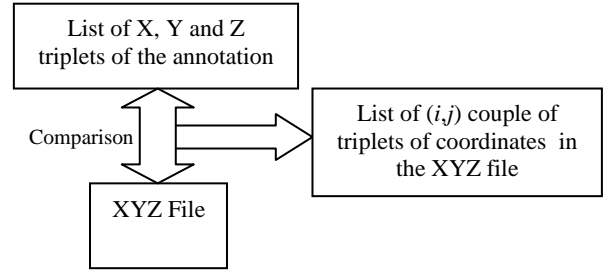

Fig. 12: Comparison of the two lists of 3D coordinates to research the wanted positions in the XYZ file

A new mask, same-sized as the XYZ file and, so, as the image, is constructed. In this mask, white pixels are placed at the previously detected positions and all other pixels are black. A dilation and an erosion of the mask permit to fill holes that can appear, especially if a sampling has been used (Fig.13). On six-million-pixels' images, the construction of an image's mask lasts around 3.2 seconds without sampling and around 3 seconds with a 20px grid-pattern sampling with the actual implementation. Detected areas on masks created with sampling are less precise than detected areas on masks created without sampling.
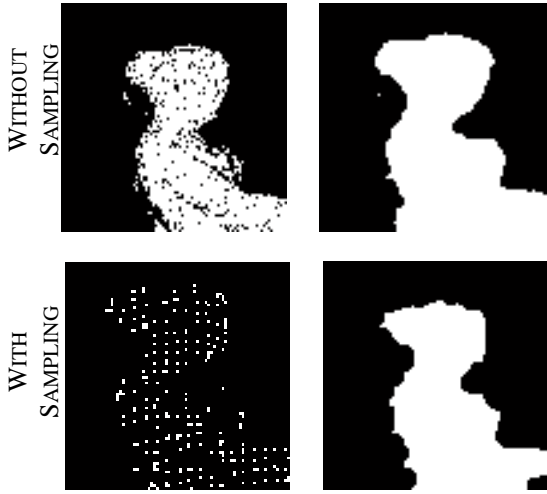
Fig. 13: Transferred annotation details (left) and masks after dilation and erosion (right)

The detected area can be displayed on the image as a colored half-transparent layer (Fig.14).

If the drawn area (corresponding to the annotation) does not appears in one of the other images, the created mask for this image will be only composed of black pixels.

At this point, annotations are represented as both masks and 3D representation. Annotations' metadata and semantic information can be defined and stored in an attached separate file (of which format still needs to be chosen) but this is not currently implemented.
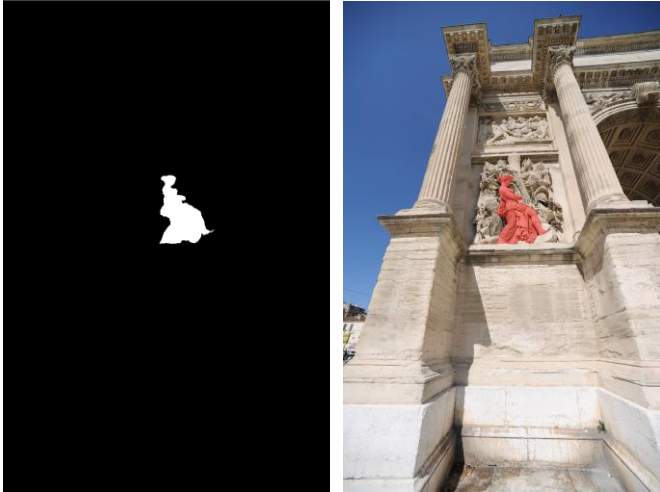

Fig. 14: Final mask (left) and visualisation of the area in another image (right)

D. Multi-view annotations

The presented method for the propagation of annotation towards a set of spatially-oriented images only allows the transfer of points existing on the annotated image. But, in general, an object needs different views to be wholly selected.

For this reason, a multi-view annotation system, using also the propagation of annotation, is necessary. This multi-view system is implemented by using the three steps of propagation, described in sections V.A., V.B. and V.C.

First of all, steps described in sections V.A. and V.B. are implemented on each view to be annotated. Thus, for each view, a list of 3D coordinates of the annotation is extracted. Then, all lists of 3D coordinates are assembled to form only one list of 3D coordinates of the annotation. Finally, this last list is used in the step described in section V.C. to project the entire annotation on all images (Fig.15).
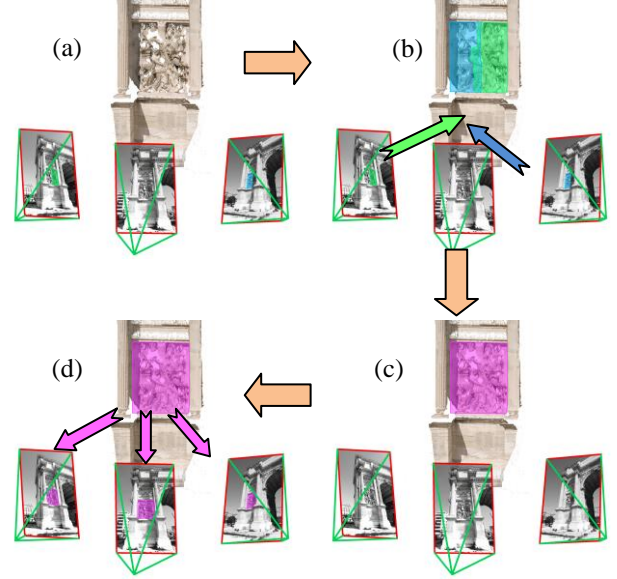

Fig. 15: Steps of multi-view annotations: (a) definition of the annotation from two images, (b) research of 3D coordinates of the defined areas, (c) grouping of the two lists of 3D coordinates and (d) propagation of the entire annotation on all images

In this way, the enrichment of annotations can be performed from different viewpoints.

VI. IN SITU VISUALIZATION

This section describes how a user can take a picture with a mobile device in situ and visualize as a non-real time augmented reality way the different annotations made on the first set of photographs.

A. New image orientation

First, a set of hybrid annotations were made on the previously oriented images and stored as 3D files (see section V). Then a picture is taken in situ with a mobile device. This picture is afterwards oriented towards the images base. The described process is automatic and can be launched by a server receiving the image. For this step, we compare the new picture with all of the photographs in the base (or a sub-set of them) in order to extract common feature points (as described in section IV.B). As the light changes depending on the time of the day and the weather (Fig.16), the tie-points matching must be made on the full scaled pictures to bypass the differences of shadows and lights. With these parameters, tie-point extraction step takes up to 30 seconds, on a double quad core computer.

Even though this full scaled tie-points extraction takes a long time, it is robust enough to find, in almost all cases,

enough points for a precise orientation.The main goal was to provide a robust solution, no matter the time consuming.

Fig. 16: Example of diachrony issues

Then the in-situ picture is oriented towards the images base. For this step, we compute the new image's orientation parameters using a 3Dto2D spatial resection procedure based on our bundle adjustment method: the previously known orientations (and related 3D coordinates) are frozen so their positions remain the same. The new image's orientation is computed in about 2 minutes.

B. Projection of annotations

Once the new picture is oriented, the related internal and external camera parameters are used for displaying the 3D point-based representation of the annotation towards the in-situ image point of view (Fig.17).
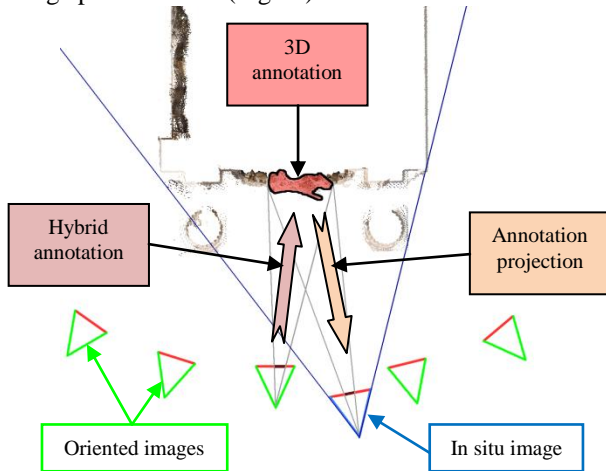

Fig. 17: Annotation reading on an in situ acquired images

This shows that a fine and detailed surface's annotation can be precisely projected on a picture taken anytime in situ (Fig.18). We made a first implementation using Maya scripts, the image projection lasts a few seconds.

Fig. 18: Results for the projection of annotations on an in situ acquired image

As a point cloud is a discrete system, if the picture taken in situ contains more pixels than points available, the annotation will appear discontinuously (Fig.19). A meshed model, represented by a surface (or the 3Dto2D projection of the region contours), could prevent this kind of issues.

Fig. 19: Scale resolution issues

VII. CONCLUSION

This work has described a process for automatic annotations transfer towards a set of spatially-oriented images and the visualization of these annotations on an in situ acquired image. The complete on-site visualization process is not finished as we did not focus on the image transfer protocol between the mobile device and a server. But results shown that in a matter of 3 minutes, we can finely project precise annotations on a new image, orientated towards a large 2D/3D database. 70% of computation time is used in the image's orientation. Experimentations show encouraging results and

some applications could be already considered: monitoring the state of conservation of historic buildings (e.g. detailed semantic annotation of degradation phenomena); assisting the historic studies (e.g. representing building's stratigraphic layers in situ); enhancing touristic visits by giving access to detailed documentation concerning the observed artifact in situ. Our approach validates the subject's geometric point of view, but further informatics implementation requires an in-depth analysis of several issues that we'll take into account for our future development.

First of all, in the process of annotation, the use of an automated segmentation of images or of implicit point cloud (implicit because contained in XYZ files) could help the selection or the transfer of annotations. Indeed, the definition of annotation can be come down to simply select parts of the segmentation and the transfer can be come down to detect parts of the segmentation.

Second, currently, in order to orientate the in situ acquired image, a comparison with all images in the set of already oriented images is implemented. So as to speed up this step, the use of images GPS data (that could be acquired at the same time as the image) could limit the number of images to compare with.

A further complete computer implementation of the process should provide:
- an optimized processing for in situ applications
- a parallel computing for larger images base (e.g. using GPU-based implementation of the SIFT algorithm could divide by 7 the tie-point research time [17])
- a cloud computing system in order to process the in situ acquired image and to communicate results to the mobile device

As a conclusion, at this moment, our approach focuses only on the in situ reading of a pre-segmented and annotated 3D scene. An important interest would be to be able to directly annotate the in situ acquired image. For that, the generation of a "virtual" XYZ file for this image (created from the orientation of this image and XYZ files of all other images) could allow to use the annotations' transfer process.

Finally, as a building aspect evolves in time (conservation state, restauration, etc), the introduction of the temporal dimension in our approach (for managing spatio-temporal transformations) should be expected.

REFERENCES

[1] K. Petridis, D. Anastasopoulos, C. Saathoff , N. Timmermann, I. Kompatsiaris, and S. Staab, M-OntoMat-Annotizer: Image annotation linking ontologies and multimedia low-level features, KES 2006 10th Intnl. conf. on knowledge based, intelligent information and engineering systems, 2006.

[2] C. Halaschek-Wiener , G. Jennifer , S. Andrew , G. Michael, P. Bijan , and H. Jim, PhotoStuff -- An Image Annotation Tool for the Semantic Web, 2005.

[3] J. Shotton., J. Winn , C. Rother, and A. Criminisi, TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context, IJCV(81), No. 1, 2009.

[4] S. Barrat and S. Tabbone, Classification et extension automatique d'annotations d'images en utilisant un réseau Bayesien, 2009.

[5] J. Li and J.Z. Wang, Real-Time Computerized Annotation of Pictures, Proc. ACM Multimedia. pp. 911–920, 2006.

[6] J. Hunter and C.H. Yu, Assessing the Value of Semantic Annotation for 3D Museum Artifacts, 2011

[7] M. Attene, F. Robbiano, M. Spagnuolo and B. Falcidieno, Part-based Annotation of Virtual 3D Shapes, 2009

[8] S. Havemann, V. Settgast , R. Berndt , O. Eide and D.W. Fellner, The Arrigo Showcase Reloaded – Towards a sustainable link between 3D and semantics, 2008

[9] N. Snavely, S. Seitz , and R. Szeliski, Photo Tourism: Exploring Photo Collections in 3D, 2006.

[10] C. Stefani, C. Busayarat, J. Lombardo and L. De Luca, A database of spatialized and semantically-enriched iconographic sources for the documentation of cultural heritage buildings. Proceedings of IEEE VSMM 2012 - 18th International Conference on Virtual Systems and Multimedia. Milan, Italy, 2-5 September 2012.

[11] R.T. Azuma, A survey of augmented reality, Teleoperators and Virtual Environments 6, 4, pp. 355–385, August 1997.

[12] H. Kato and M. Billinghurst, Marker tracking and HMD calibration for a video-based augmented reality conferencing system, Proceedings of the 2nd International Workshop on Augmented Reality (IWAR 99). October 1999

[13] M. Pierrot-Desseiligny and I. Clery, Apero, an open source bundle adjustment software for automatic calibration and orientation of a set of images, Proceedings of the ISPRS Commission V Symposium, Image Engineering and Vision Metrology, Trento, Italy, 2-4 March 2011

[14] M. Pierrot-Desseiligny, L. De Luca and F. Remondino, Automated image-based procedures for accurate artifacts 3D modeling and orthoimage generation, Geoinformatics FCE CTU Journal, vol. 6, pp. 291-299, Prague, Czech Republic, 2011

[15] D.G. Lowe, Object recognition from local scale-invariant features. Proceedings of the International Conference on Computer Vision, 1999

[16] A. Vedaldi, An open implementation of the SIFT detector and descriptor, UCLA CSD Tech. Report 070012, 2006

[17] S. Heymann, K. Maller, A. Smolic, B. Froehlich, and T. Wiegand, SIFT implementation and optimization for general-purpose GPU, in Proceedings of the International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, 2007.,