3D Visualization through Planar Pattern Based Augmented Reality

Charalabos IOANNIDIS and Styliani VERYKOKOU, Greece

Key words: Augmented reality, Photogrammetry, Computer vision, Features detection, Image matching.

SUMMARY

Augmented reality is the scientific field that aims at an enhanced sense of the real world. Although it was first presented in the 1960s, in recent years it has actually begun to have practical applications in various fields and today arouses the interest of many researchers and scientists, as well as companies.
The purpose of this paper is the presentation of an augmented reality application that allows the visualization of the three-dimensional anaglyph of a region. The application was developed through the use of methods and algorithms of photogrammetry and computer vision and is based on the recognition of an orthoimage of a region that is augmented with its digital terrain model on a computer screen. The SURF algorithm is used for the extraction of interest points in a photograph of the orthoimage and in every real world frame. The matched features between these images are geometrically verified by the RANSAC algorithm. The Levenberg-Marquardt optimization algorithm is used for the calculation of the homography between the matched images, pattern recognition is conducted and the camera exterior orientation for each frame is estimated. The latter, in combination with the camera intrinsic parameters, which are computed through a camera calibration procedure, leads to the augmentation of the real world scenes.
The application indicates the very promising capabilities of augmented reality for the realistic visualization of the topography of an area and the detailed observation of the anaglyph, without the need of a three-dimensional printing of the terrain model.

3D Visualization through Planar Pattern Based Augmented Reality

Charalabos IOANNIDIS and Styliani VERYKOKOU, Greece

1. INTRODUCTION

Augmented reality (AR) is a rapidly evolving technology, which enriches reality with computer generated information. Its objective is the enhancement of people's perception of reality through the integration of synthetic information in the real environment, which has the dominant role. Ideally, the objects that complement the real world are presented as coexisting in the same place with it. However, visual information is not the only kind of complementary information that is added in the actual scene. Augmented reality can potentially enhance all five senses and especially - in addition to vision - hearing and touch, in spite of the fact that its predominant use is the addition of visual information in the real world.

An AR system combines real and virtual objects in a real environment, allows real-time interaction and registers virtual objects in the three-dimensional space (Azuma, 1997; Azuma et al., 2001). The basic components of an AR system include a display, a computer system, a camera or another optical instrument, appropriate software, the real scene and the virtual objects and – depending on the application – various sensors (GPS, compass, accelerometer, gyroscope), markers or patterns for recognition, a content server, web services and possibly more (Kipper and Rampolla, 2013). AR displays can be divided into three basic categories: head-worn displays (HWDs), handheld displays (HDs) and spatial displays. HWDs include head-mounted displays, helmet-mounted displays, head-mounted projective displays, virtual retinal displays, AR glasses and AR contact lenses. Smartphones, tablet PCs and less used personal digital assistants are characterized as HDs, as they contain a small screen on which the users can view the augmented scene. Spatial displays may be divided into screen-based video see-through displays, which require an ordinary PC and off-the-shelf hardware components, spatial optical see-through displays, which use spatial optical combiners, transparent screens or optical holograms, and projection-based spatial displays, which project images on physical surfaces of objects (Bimber and Raskar, 2005).

The term "augmented reality" was coined in 1992 by Tom Caudell and David Mizell (Dixon, 2010) and, since then, a lot of applications have been developed in various fields, including medicine, education, the army, entertainment, sports, art, culture, archaeology, tourism, navigation, commerce, advertising, architecture, interior design and task support. AR applications can be classified as either indoor or outdoor. They can also be classified as those that give the users freedom of movement and those which assume that the users stay in the same position. Moreover, they may be divided into applications which add information that allow for a better understanding of the environment and those that create a fantasy environment, allowing the users to see objects that do not really exist (Kipper and Rampolla, 2013). According to another categorization of the applications, they are divided into those that overlay information which is not part of the natural world, those which depict additional information so that it is not distinguished from the real environment and those which overlay

information that exists in the natural environment, but is not directly observable (Goldiez, 2004).

Depending on the methodology followed in order to achieve the augmentation of the real world, AR may be classified into four main categories (Kipper and Rampolla, 2013):

- Pattern-based augmented reality recognizes a pattern, which may be either a marker placed on the real scene (marker-based AR) or another picture of the actual scene (markerless AR), and augments it with virtual objects;
- Outline augmented reality is based on the recognition of a part of the body or the entire body, which is then augmented with a synthetic object;
- Location-based augmented reality uses GPS or triangulation location information, which – in combination with information from an accelerometer and a digital compass – allows integration of virtual objects in the right position on the image of the real world;
- Surface augmented reality is accomplished using screens, walls, or floors that respond to the touch of objects or people and provides them with virtual information in real time.

1.1 Augmented Reality Applications in the Field of Surveying

Some AR applications are related to surveyor science. An important aspect is the possibility for virtual reconstruction of half-ruined buildings, statues or archaeological sites (e.g., at the archaeological site of Olympia, Greece, Vlahakis et al., 2002). Visitors who are suitably equipped have the ability to see three-dimensional monuments as they were in antiquity, rather than the present ruins. Moreover, AR can be used for the visualization of constructions during their design phase, since, through this technology, three-dimensional models of planned buildings or other projects can be superimposed on the place where their construction is planned. Also, high-accuracy three-dimensional photo-models derived from close range photogrammetry may be integrated into the real world (Portalés et al., 2009). Cartographic issues related to the technology of AR are also being investigated resulting in applications that combine methods of cartographic representation with this technology (Koussoulakou et al., 2001).

Furthermore, AR can be used in navigation, in order to improve the effectiveness of simple applications of this kind. For instance, with the use of AR navigation applications targeted at drivers, the latter follow the route depicted in the image of the real world through appropriate graphics and do not need to consult a map, which may be more distracting when they are driving (Leventi, 2013). The combination of AR and location-based services can result in significant applications, which address not only drivers, but also pedestrians. One of the first AR navigation systems for pedestrians was the wearable computer system “Map-in-hat”, with a see-through display, a digital compass, a differential GPS and navigation software (Thomas et al., 1998). Today, such systems overlay instructions on the real world, for finding points of interest.

Location-based AR applications focus not only on navigation but also on providing information for places seen by the users. Such applications may make use of a GPS sensor, an accelerometer, a gyroscope or a compass, as well as internet services. The first one was the

Real-World Wide Web Browser (Kooper and MacIntyre, 2003), which overlays data from the web on the real world through an HMD and the overlaid information is updated depending on the users' location and orientation. Other applications like MARA from Nokia, as well as the AR browsers Wikitude, Layar and Junaio superimpose information about any given surroundings on a live video of the real world on mobile devices.

In this paper, an AR application for visualizing the three-dimensional anaglyph of a region through the screen of a PC is introduced. The aim was the development of an application that gives the users the possibility of examining the anaglyph of a specific area without the need of a three-dimensional print-out of the Digital Terrain Model (DTM). The basic concept is the depiction of the DTM instead of a printed orthoimage of the same area on a computer screen, at the same position as the orthoimage, with the same orientation and size and with the proper perspective, each time the orthoimage is located in the field of view of the computer camera. The application complements the real environment and does not replace it. The latter, which consists of the room – or any indoor or outdoor environment – where the computer is located, including the orthoimage as well as the people who are there, is not entirely hidden. The only additional object is the DTM, which is superimposed on the orthoimage. Thus, it is an augmented reality application, rather than a virtual reality one, as the users are not immersed in a completely synthetic environment and are given the opportunity to interact with an augmented world.

2. METHODOLOGY

A planar pattern based markerless AR application, based on the recognition of a planar object, which is a printed orthoimage, was developed. The area depicted in the orthoimage is about 18km x 12km around the artificial lake of the river Ladonas in the Peloponnese, in southern Greece. The initial data includes the following:

- a pattern image, which is a photograph of the printed orthoimage,
- a DTM of the region depicted in the orthoimage, on which the latter is draped for a realistic representation of the anaglyph, and
- the interior orientation of the camera, which captures the real world (the computer built-in camera or the external camera linked to it).

The exterior orientation of every video frame is computed using 6-DOF visual tracking – unless the pattern is not recognized – by analyzing features detected in every frame and establishing correspondences between video frames and positions in the three-dimensional space. Finally, the DTM is drawn properly on a computer window whereby the background is each video frame.

2.1 Camera Calibration

The term camera calibration refers to all the measuring and computational procedures required for the determination of the interior orientation parameters of a camera, that is, the pixel coordinates of the principal point (x_0, y_0), the camera constant in pixels in x and y direction (c_x, c_y) and the coefficients of lens distortion polynomials. As far as the distortion is concerned, the Brown-Conrady model (Brown, 1971) is used as shown in equation (1).

$$\begin{bmatrix} x_{ideal} \\ y_{ideal} \end{bmatrix} = \left(1 + k_1 \cdot r^2 + k_2 \cdot r^4 + k_3 \cdot r^6\right) \cdot \begin{bmatrix} x_d \\ y_d \end{bmatrix} + \begin{bmatrix} 2 \cdot p_1 \cdot x_d \cdot y_d + p_2 \cdot \left(r^2 + 2 \cdot x_d^2\right) \\ p_1 \cdot \left(r^2 + 2 \cdot y_d^2\right) + 2 \cdot p_2 \cdot x_d \cdot y_d \end{bmatrix} \quad (1)$$

$$\text{where} \quad \begin{Bmatrix} x_d = x - x_0, \quad y_d = y - y_0, \quad r = \sqrt{\left(x - x_0\right)^2 + \left(y - y_0\right)^2} \\ k_1, k_2, k_3 \ \text{ the coefficients of the radial distortion polynomial, and} \\ p_1, p_2 \ \text{ the coefficients of the tangential distortion polynomial} \end{Bmatrix}$$

The calibration was done by taking pictures of a planar chessboard pattern shown at several different orientations. The methodology followed is based on Zhang's and Bouguet's methods (Zhang, 2000; Bouguet, 2013), and is fully automated. The initial data includes the images of the chessboard pattern and the number of its internal corners in the two perpendicular directions. The procedure followed for the computation of the calibration parameters consists of:

- an initial processing of each image,
- a check whether the chessboard pattern can be recognized in each image, followed by the detection of the internal corners of the chessboard if that check is positive,
- the computation of the object coordinates of the internal corners of the chessboard,
- the estimation of the initial interior orientation parameters of the camera,
- the computation of the approximate exterior orientation parameters of the camera for each image, if the chessboard pattern was detected, and
- the final computation of the camera interior parameters, as well as the camera exterior parameters for each image, using the Levenberg-Marquardt optimization algorithm (Levenberg, 1944; Marquardt, 1963), with the criterion of minimizing the reprojection error.

The detection of the internal corners of the chessboard in each image was held using morphological operations and image processing, a contour detection algorithm (Suzuki and Abe, 1985) and some additional processes and checks. As far as the initial interior orientation parameters of the camera are concerned, the principal point is considered to be at the center of the image, the estimation of the camera constant is based on the extraction of the vanishing points, while the distortion coefficients are set to zero. The approximate camera exterior orientation for each chessboard picture was computed based on the methodology described in section 2.6.

2.2 Definition of the Coordinates of the Corners of the Pattern Object

The origin of the object coordinate system is meant to be located at the center of the pattern object, to wit the orthoimage, X and Y axes lie on the object plane and Z axis is perpendicular to it. In this system, the coordinates of the four corners of the pattern object are defined. Z coordinate is assumed to be zero, while X and Y coordinates are derived from the normalized width and height of the pattern image (equation (2)), being within the range of [-1, 1].

$$\text{normalized\_width} = \frac{\text{width}}{\max(\text{width}, \text{height})}, \quad \text{normalized\_height} = \frac{\text{height}}{\max(\text{width}, \text{height})} \quad (2)$$

2.3 Features Extraction and Description in the Pattern Image and in each Video Frame

A very important step for the augmentation of the real world scenes is the digital matching between the pattern image and each video frame. In the application, feature-based matching was implemented. This is conducted in two basic steps:

– the detection and description of interest points independently in each image, which was accomplished using the SURF (Speeded-Up Robust Features) algorithm (Bay et al., 2006), and
– the extraction of correspondences, which is the main process of image matching.

Thus, features detection and description is applied once in the pattern image and in every video frame, after they have been converted to greyscale, as SURF does not use color information. This algorithm detects scale and rotation invariant feature points, while skew, anisotropic scaling and perspective effects are also covered to some degree.

SURF divides the scale space into octaves, each of which is divided into scale levels. The algorithm uses integral images (Viola and Jones, 2001) in order to achieve fast convolution with box type filters, which consist of rectangular regions. The entry of an integral image at a location **x**=(x,y) is the sum of all pixels in the input image within a rectangular region formed by the origin and **x**. The algorithm detects interest points located in blob-like structures of the image, based on the determinant of an approximation of the Hessian Matrix. Its elements are the convolution of box type filters, which approximate the Gaussian second order derivatives in directions x, y and xy (Figure 1), with the image in point **x**. Blob response at location **x** and at scale s that corresponds to the dimensions of the filters is given by the determinant of the approximated Hessian Matrix, which is computed for every pixel, with different filter sizes. This process results in the acquisition of blob response maps for the scale levels of each octave.

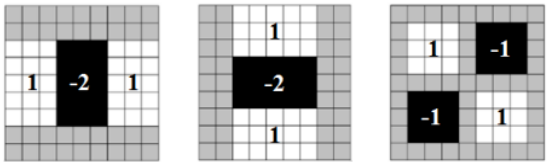

Figure 1. 9x9 filters that approximate the Gaussian second order derivatives in directions x (left), y (center) and xy (right). Grey regions have zero value (source: Bay et al., 2006).

The pixels with a Hessian determinant below a specific threshold are rejected, the image location and scale of the feature points are computed through a non-maximal suppression in a 3x3x3 neighborhood in scale space (Neubeck and Van Gool, 2006) and finally, the feature points are located in the image and over scales with sub-pixel and sub-scale accuracy, by interpolating the determinant in scale and image space with the method proposed by Brown and Lowe (2002). Furthermore, the algorithm returns the sign of the trace of the Hessian Matrix for fast indexing during the matching stage.

Figure 2. Haar wavelet filters in x (left) and y direction (right) (source: Bay et al., 2006).

The detection of feature points is followed by the extraction of their descriptors. Firstly, the dominant orientation of each interest point is estimated by computing the Haar wavelet responses in x and y direction (Figure 2) for some points in its neighborhood, and by summing the responses within a sliding orientation window, in order to compute a local orientation vector for each location of the window, the longest of which demonstrates the dominant orientation of the interest point (Figure 3). Subsequently, a square region is defined around each interest point, with the size depending on its scale, and is oriented along the dominant orientation. The descriptor vector of length 64 of each interest point is based on the sum of the Haar wavelet responses in x and y direction and of their absolute values for 25 points of each of the 16 subregions of the square region and indicates the underlying intensity structure of the region.

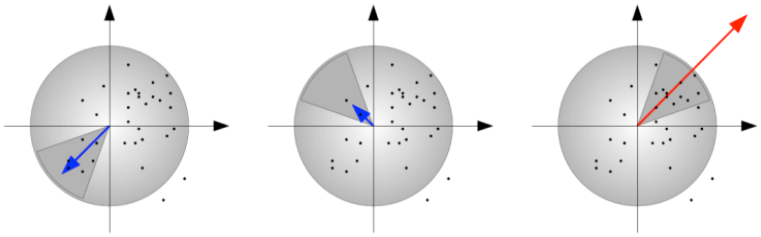

Figure 3. Orientation assignment (source: ftp://tracking.mat.ucsb.edu/projects/external_docs/tracking/intro_to_matching.pdf).

2.4 Matching of Feature Points between the Pattern Image and each Video Frame - Outliers Removal

At the stage of finding correspondences between the pattern image and every video frame, only feature points that have the same type of contrast are compared. The matching criterion is the Euclidean distance between the descriptor vectors of the feature points. However, the minimization of the latter does not suffice and thus it is combined with other techniques in order to reject incorrect correspondences, known as outliers.

In the presented application, a cross-check test was implemented. According to this, the descriptor vector of each feature point in the pattern image is compared with the descriptor vector of every feature point in the video frame and some candidate correspondences are returned. These are additionally checked through reverse comparison. Thus, two feature points, i and j, are matched if the nearest neighbor of the descriptor of point i in the pattern image is the descriptor of point j in the video frame and reversely if the nearest neighbor of the descriptor of point j in the video frame is the descriptor of point i in the pattern image. However, after cross testing, many outliers still remain. Therefore, a maximum accepted distance is defined as a threshold and the correspondences are rejected if the Euclidean distance between the descriptors of the matched feature points is above this threshold. In this way, a significant number of outliers is removed, although a few incorrect matches are not detected. These are rejected using the RANSAC algorithm (Fischler and Bolles, 1981). The latter is applied if at least five matches are detected. Otherwise, the real world scene is not augmented, as it is considered that the orthoimage cannot be recognized in the frame.

RANSAC (RANdom SAmple Consensus) calculates the parameters of a mathematical model using a data set, which may contain many errors, and relies on the use of minimum data. In the case of the application, the data set consists of the correspondences that were not rejected and the model is the geometric relation between the pattern image and the image of the orthoimage in the video frame. The above relation is considered to be the two-dimensional projective transformation, also known as homography in the two-dimensional projective space (section 2.5), as the latter can map points in two different images of the same plane, which may have been taken under different orientations and from a different position.

The iterative procedure followed by RANSAC for the removal of outliers can be summarized as follows. Firstly, a sample of four matches, which is the minimum number of matches required to calculate the homography, is randomly chosen from all matches. The parameters of the homography are estimated using the random sample. The number of valid matches (inliers) for the above solution is calculated and this shows the quality of the computed geometric relation. If the number of inliers is greater than a threshold, the model is accepted and the algorithm terminates with success, having calculated the inliers and thus having rejected the outliers. Otherwise, if the minimum number of matches was not found, and if the above steps were repeated N times, where N is the maximum number of iterations, the algorithm terminates with failure. Otherwise, these steps are repeated again.

The rejection of outliers by RANSAC yields very satisfactory results and, through the described procedure, the correspondences between the pattern image and each video frame which were not rejected consist only of inliers in the vast majority of cases.

2.5 Estimation of the Homography between the Pattern Image and each Video Frame

Homography in 2D projective space is expressed by an invertible 3x3 matrix with the use of homogenous coordinates. This kind of coordinates is used in projective geometry. One of their advantages is the possibility they offer for a linear solution for problems that, when expressed in Cartesian coordinates, are nonlinear. One of their important properties is the fact that the homogeneous coordinates of a point represent the same point when they are multiplied by a non-zero constant. Furthermore, n+1 homogeneous coordinates are required in order for a point to be expressed in n-dimensional projective space $\mathbf{P^n}$. The conversion from a Cartesian representation of a point in Euclidean space $\mathbf{R^n}$ to homogeneous coordinates in $\mathbf{P^n}$ is done by adding an additional unit coordinate, while the inverse conversion is done by dividing the first n coordinates of the point with the (n+1)th coordinate.

The relation between the points of the orthoimage in the video frame and the corresponding points in the pattern image is given by equation (3), where x_{frame}, y_{frame} are the pixel coordinates of a point in a video frame, $x_{pattern}$, $y_{pattern}$ are the pixel coordinates of the same point in the pattern image, s is any non-zero constant and h_{ij} are the elements of the homography matrix **H**.

$$s \cdot \begin{bmatrix} x_{frame} \\ y_{frame} \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \cdot \begin{bmatrix} x_{pattern} \\ y_{pattern} \\ 1 \end{bmatrix} \quad (3)$$

Homography does not change if the homography matrix is multiplied by a non-zero constant. Consequently, only the ratio of the elements of the matrix is important, and as among its nine elements only eight different ratios can be formed, planar homography has eight degrees of freedom (Grammatikopoulos, 2007). Thus, the element h_{33} of the matrix can be considered equal to 1, as is assumed in the application.

The homography matrix is calculated by RANSAC through a linear method, with the use of only four matches, without using all the inliers detected by the algorithm. Thus, the initial homography estimation made by RANSAC is refined using the set of all the inliers, via a nonlinear optimization using the Levenberg-Marquardt algorithm, in order to minimize the reprojection error.

The recognition of the orthoimage in every video frame is accomplished if a minimum number of five inliers was detected by RANSAC. The pixel coordinates of the four corners of the orthoimage are calculated using the computed homography matrix and the pixel coordinates of the four corners of the pattern image. These are transformed to homogeneous coordinates and are multiplied by the homography matrix, in order to obtain the homogeneous coordinates of the corners of the orthoimage in the video frame, which are then transformed into Cartesian pixel coordinates.

2.6 Estimation of Camera Exterior Orientation for Every Video Frame

The estimation of the camera 6-DOF pose for every video frame is done using the camera interior orientation parameters, the pixel coordinates of the corners of the orthoimage in the video frame and their corresponding object coordinates. The results of the above computation are the translation of the object coordinate system into the projection center, which is the origin of the three-dimensional Cartesian camera system, and the rotation of the object coordinate system into the camera system.

The mathematical model used is the projection transformation, which is expressed by equation (4). In this equation, **K** is the matrix with the camera intrinsic parameters, also known as camera matrix, the joint rotation-translation matrix [**R**|**t**] is the matrix of extrinsic parameters, X, Y, Z are the object coordinates of a point, x, y are the pixel coordinates of that point and λ is a scale factor.

$$\lambda \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} c_x & 0 & x_0 \\ 0 & c_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{K}} \cdot \underbrace{\begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix}}_{[\mathbf{R}|\mathbf{t}]} \cdot \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (4)$$

The elements of the joint rotation-translation matrix are computed linearly according to equation (5), after the undistortion of the image coordinates of the four corners of the orthoimage in the video frame and the computation of the two-dimensional homography **H** that relates the X, Y object coordinates of the orthoimage with the corresponding undistorted image coordinates.

$$\begin{aligned} \mathbf{r_1} &= \lambda \cdot \mathbf{K}^{-1} \cdot \mathbf{h_1} \\ \mathbf{r_2} &= \lambda \cdot \mathbf{K}^{-1} \cdot \mathbf{h_2} \\ \mathbf{r_3} &= \mathbf{r_1} \times \mathbf{r_2} \\ \mathbf{t} &= \lambda \cdot \mathbf{K}^{-1} \cdot \mathbf{h_3} \end{aligned} \quad \text{where} \quad \left\{ \begin{aligned} \mathbf{h_1} &= \begin{bmatrix} h_{11} & h_{21} & h_{31} \end{bmatrix}^T \\ \mathbf{h_2} &= \begin{bmatrix} h_{12} & h_{22} & h_{32} \end{bmatrix}^T \\ \mathbf{h_3} &= \begin{bmatrix} h_{13} & h_{23} & h_{33} \end{bmatrix}^T \\ \mathbf{r_1} &= \begin{bmatrix} r_{11} & r_{21} & r_{31} \end{bmatrix}^T \\ \mathbf{r_2} &= \begin{bmatrix} r_{12} & r_{22} & r_{32} \end{bmatrix}^T \\ \mathbf{r_3} &= \begin{bmatrix} r_{13} & r_{23} & r_{33} \end{bmatrix}^T \end{aligned} \right. \quad \text{where } \mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \quad \text{and } \lambda = \frac{1}{\left\| \mathbf{K}^{-1} \cdot \mathbf{h_1} \right\|} \tag{5}$$

However, due to noise in data, the computed rotation matrix $\mathbf{R}=[\mathbf{r_1}\ \mathbf{r_2}\ \mathbf{r_3}]$ may not generally satisfy the properties of a rotation matrix. Thus, it is "coerced" to satisfy the orthogonality condition $\mathbf{R}\mathbf{R}^T=\mathbf{R}^T\mathbf{R}=\mathbf{I}$ by calculating its singular value decomposition, thus factoring it into two orthonormal matrices **U** and **V** and a middle matrix **W** with the singular values of **R** in its diagonal, as shown in equation (6).

$$\mathbf{R} = \mathbf{U} \cdot \mathbf{W} \cdot \mathbf{V}^T \tag{6}$$

The best approximating matrix is then given by equation (7) (Zhang, 2000; McGlone et al., 2004; Bradski and Kaehler, 2008).

$$\hat{\mathbf{R}} = \mathbf{U} \cdot \mathbf{V}^T \tag{7}$$

Afterwards, the rotation matrix is transformed to a three-dimensional rotation vector, using the Rodrigues rotation formula (Bradski and Kaehler, 2008). This vector indicates the direction of the rotation axis and its magnitude is equal to the magnitude of the rotation. This axis-angle representation is more compact than a rotation matrix, so it is more suitable for optimization procedures.

The penultimate step of the estimation of camera pose for every video frame is the Levenberg-Marquardt optimization, in order to refine the translation and rotation vectors by reducing the reprojection error. Finally, the rotation vector is converted back into a 3x3 rotation matrix using the Rodrigues formula (Bradski and Kaehler, 2008), and thus the result of this process is the joint rotation-translation matrix of the camera extrinsic parameters for each video frame.

2.7 Rendering of the Augmented Scene

Having calculated the camera interior orientation and the camera exterior orientation for a video frame, the DTM can be drawn at the right position, with the proper scale, orientation and perspective in the scene of the real world. In particular, it is drawn at the position where the orthoimage is found, with the same dimensions (length and width) and orientation as this.

The first step of this process is the rendering of the video frame on a computer window, so that it forms its background. The second and the most important step is the rendering of the DTM on that window. This process is described below.

The geometry of the DTM is defined by a set of X, Y, Z coordinates for each of the vertices in its local coordinate system. These coordinates are transformed into the object coordinate system, by being normalized into the range of [-1, 1]. Furthermore, a matrix that converts these normalized model coordinates from the object coordinate system into the camera system is formed. This kind of transformation is known as viewing transformation and is expressed by equation (8), using the elements of the computed joint rotation-translation matrix for the specific video frame.

$$\begin{bmatrix} X_{CAMERA} \\ Y_{CAMERA} \\ Z_{CAMERA} \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} X_{MODEL} \\ Y_{MODEL} \\ Z_{MODEL} \\ 1 \end{bmatrix} \tag{8}$$

In the above equation, $[X_{MODEL}\ Y_{MODEL}\ Z_{MODEL}\ 1]^T$ is the vector with the homogeneous coordinates of the vertices of the DTM in the object coordinate system, while $[X_{CAMERA}\ Y_{CAMERA}\ Z_{CAMERA}\ 1]^T$ is the vector with the corresponding coordinates in the camera coordinate system.

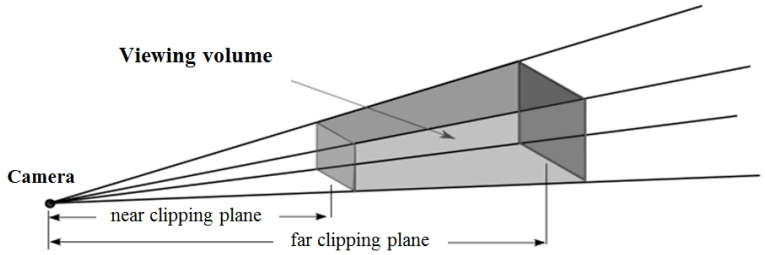

Figure 4. Viewing volume (source: http://math.hws.edu/graphicsnotes/c3/s5.html).

Furthermore, the viewing volume of the camera has to be specified. The latter determines how the DTM is projected into the scene and which of its parts are clipped, so as not to be drawn in the final scene. For this reason, perspective projection is used, because of the realistic rendering of the DTM in the scene. The viewing volume for a perspective projection is a truncated pyramid, as shown in Figure 4. This transformation converts the homogeneous camera coordinates of the vertices of the DTM into homogeneous clip coordinates $[X_{CLIP}\ Y_{CLIP}\ Z_{CLIP}\ W_{CLIP}]^T$ through equation (9), using the intrinsic parameters of the camera, the dimensions (width and height) of the video frame and the distances of the near and far clipping planes from the projection center. These distances may be given any value, provided that the camera coordinates of the vertices of the DTM are within the near and far clipping planes.

$$\begin{bmatrix} X_{CLIP} \\ Y_{CLIP} \\ Z_{CLIP} \\ W_{CLIP} \end{bmatrix} = \begin{bmatrix} \frac{2 \cdot c_x}{width} & 0 & 1 - \frac{2 \cdot x_0}{width} & 0 \\ 0 & \frac{2 \cdot c_y}{height} & -1 + \frac{2 \cdot y_0}{height} & 0 \\ 0 & 0 & \frac{near + far}{near - far} & \frac{-2 \cdot far \cdot near}{far - near} \\ 0 & 0 & -1 & 0 \end{bmatrix} \cdot \begin{bmatrix} X_{CAMERA} \\ Y_{CAMERA} \\ Z_{CAMERA} \\ 1 \end{bmatrix} \quad (9)$$

Subsequently, the homogeneous clip coordinates are converted into normalized Cartesian coordinates (X_N, Y_N, Z_N), by division with W_{CLIP}, and then X_N, Y_N are transformed into window coordinates in pixels by viewport transformation, according to equations (10) and (11), where width and height are the dimensions of the video frame in pixels, which are considered to be equal to the dimensions of the window. Moreover, a depth-range transformation is applied according to equation (12), for the acquisition of depth information.

$$x = \frac{width}{2} \cdot X_N + \frac{width}{2} \quad (10)$$

$$y = \frac{height}{2} \cdot Y_N + \frac{height}{2} \quad (11)$$

$$z = \frac{1}{2} \cdot Z_N + \frac{1}{2} \quad (12)$$

Also, a texture coordinate corresponds to each vertex of the DTM and texture mapping is applied in order for the DTM to be drawn with the orthoimage as a texture. Finally, the DTM is drawn appropriately on the defined window, whereby the background is the video frame. However, if the orthoimage is not recognized in the frame, only the real scene is drawn on the window.

3. APPLICATION

The application that has been developed is intended for computers running Microsoft Windows. It uses the computer camera in order to capture frames of the real world and augment them in almost real time, as the extraction and the processing of information from each frame and the proper rendering of the augmented scene may last from 50 to 400 milliseconds, depending on the capabilities of the computer as well as the characteristics of each frame. Furthermore, the users can insert a video file, in order to augment it, as well as watch a preview of the application with a default video file that is augmented. Additionally, the possibility of inserting the camera intrinsic parameters is provided in order to ensure that the camera pose will be computed correctly and that the DTM will be rendered with the right perspective. The camera of a laptop computer as well as the camera of a mobile phone were calibrated and the intrinsic parameters were loaded in the application so that they can be used as default values in case the users do not insert the camera interior orientation, as well as for the right augmentation of the default video captured by the calibrated camera of the mobile phone.

3.1 Development of the Application

The application was developed in the C++ programming language. The OpenCV (Open source Computer Vision) library was used for the calibration of the cameras and the computation of the camera exterior orientation for each video frame. Additionally, OpenGL (Open Graphics Library) was used for the correct rendering of the DTM and the drawing of the augmented scene. Finally, library GLM: An Alias Wavefront OBJ file Library was used in order to load the DTM, which is an Alias Wavefront object file.

During the development of the application, particular attention was paid to the combination of the OpenCV library with the OpenGL application programming interface. Specifically, OpenCV assumes a left-handed camera coordinate system, and the origin of the image coordinate system is the upper left pixel. On the other hand, OpenGL assumes a right-handed camera system and a lower left origin for window coordinates in pixels. Thus, the appropriate conversions were made for the proper combination of the results obtained using OpenCV and OpenGL and the consequent right augmentation of the real world scenes.

3.2 Results

In this section, some intermediate results, as well as the final result of the application, are presented. Figure 5(b) shows the detected interest points in the pattern image, which is illustrated in Figure 5(a), while Figure 5(d) shows the detected feature points in a random real world frame that is shown in Figure 5(c).

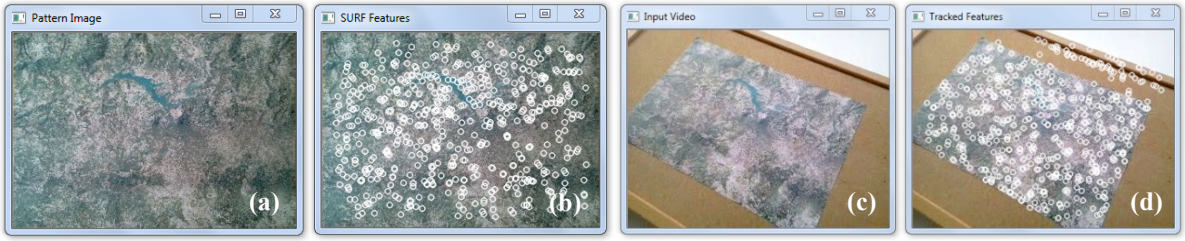

Figure 5. Pattern image (a), detected feature points in the pattern image (b), a random video frame (c), detected feature points in the video frame (d).

Figure 6 indicates the rejection of wrong matches via the RANSAC algorithm. In the window "Matches", the matches are shown between a random video frame and the pattern image, after the implementation of the cross-check test and the removal of correspondences between the feature points if the distance between their descriptor vectors is greater than the defined threshold. The window "Inliers" depicts the inliers returned by RANSAC. Three geometrically incorrect correspondences that are rejected by RANSAC are marked on the window "Matches".

Figure 7 shows the recognition of the orthoimage in a random video frame. The red line connects its four corners. Figures 8 and 9 show some random augmented scenes of the default video file and the live feed of a computer camera. Finally, Figure 10 depicts the main steps of the iterative procedure that leads to the augmentation of a real scene.

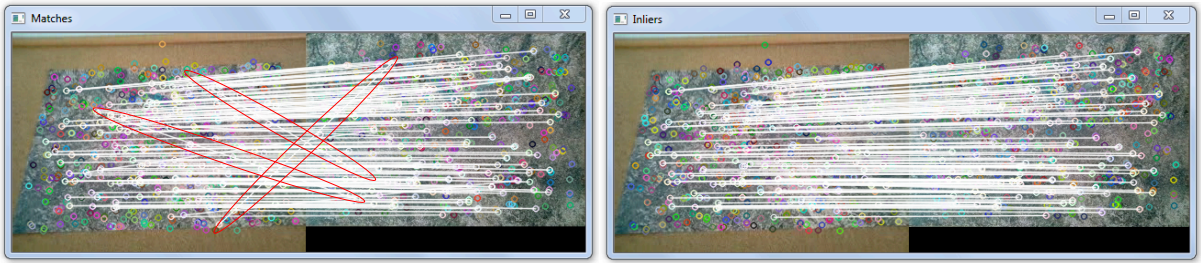

Figure 6. Matches between a random video frame and the pattern image before the rejection of the outliers by RANSAC (left) and inliers returned by RANSAC between the same random video frame and the pattern image (right).

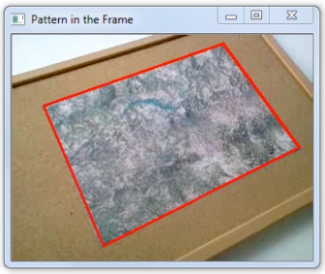

Figure 7. Recognition of the orthoimage in a random video frame.

Figure 8. Several augmented scenes captured by the camera of a mobile phone.

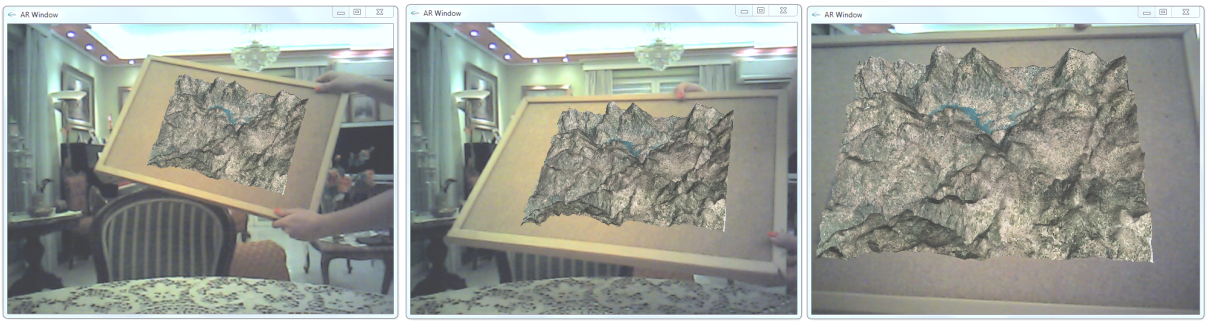

Figure 9. Augmented scenes captured by a laptop computer built-in camera.

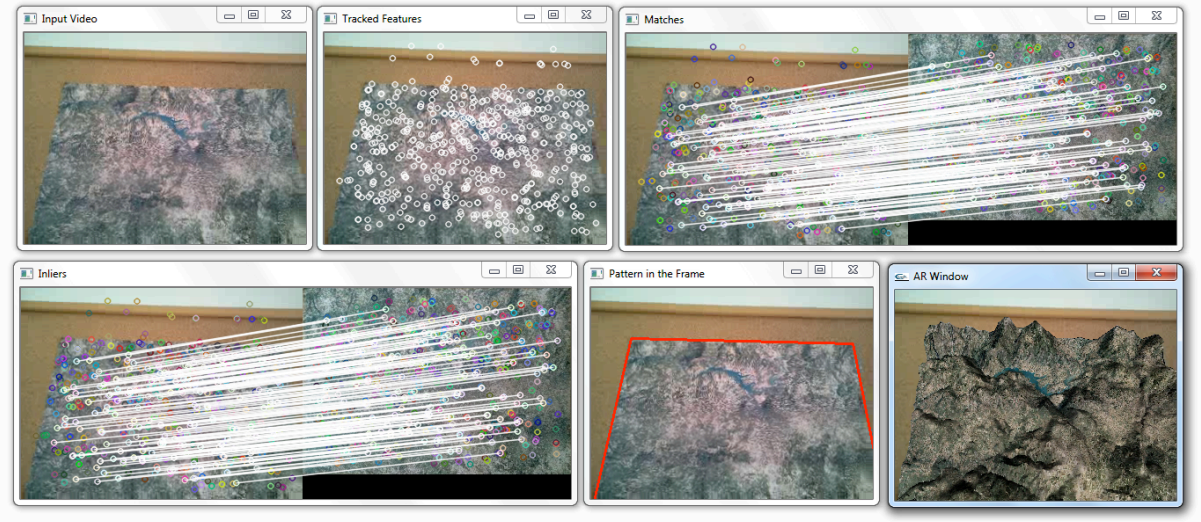

Figure 10. The procedure followed for the augmentation of a real world frame illustrated with pictures.

4. CONCLUSIONS

The results of the application are considered to be satisfactory. It successfully recognizes the orthoimage when it is placed in the field of view of the camera, even if only a part of it is captured in the frame. Furthermore, the recognition is accomplished regardless of the lighting conditions, the orientation of the pattern object and its size. Proper recognition, which mainly depends on the features detection and description algorithm, combined with the accuracy of the results of the calibration is the key element that determines the success of the augmentation of the real world and the realism of the augmented scenes.

The presented procedure may be used for visualization purposes in various fields, such as:

– in cartography, as it provides a realistic as well as innovative way of representation of the three-dimensional topography of an area;
– in school education, as it could give school students the opportunity of examining the anaglyph of regions depicted in their geography books through the use of a computer screen;
– in tourism, as it may be used by tourists for three-dimensional observation of regions of particular interest shown in paper tourist maps.

A possible extension of the application could include the import of different DTMs and pattern images and the automatic recognition of a specific orthoimage or another kind of map that is seen by the camera, so that it can be augmented with the proper DTM.

REFERENCES

Azuma, R. (1997). A Survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments, v*ol. 6, no. 4, pp. 355-385. MIT Press, USA.

Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S. and MacIntyre, B. (2001). Recent Advances in Augmented Reality. *IEEE Computer Graphics and Applications*, vol. 21, no. 6, pp. 34-47. IEEE Computer Society, USA.

Bay, H., Tuytelaars, T. and Van Gool, L. (2006). Speeded-Up Robust Features (SURF). *Proceedings of the 9th European Conference on Computer Vision*, vol. 3951, part 1, pp. 404-417. Springer-Verlag, Berlin - Heidelberg.

Bimber, O. and Raskar, R. (2005). *Spatial Augmented Reality: Merging Real and Virtual Worlds*. A K Peters Ltd, Wellesley, MA.

Bouguet, J. Y. (2013). *Camera Calibration Toolbox for Matlab* [Internet]. Available from: http://www.vision.caltech.edu/bouguetj/calib_doc. California Institute of Technology. [Accessed 23 November 2013].

Bradski, G. and Kaehler, A. (2008). *Learning OpenCV*. O'Reilly Media, USA.

Brown, D.C. (1971). Close-Range Camera Calibration. *Photogrammetric Engineering*, vol. 37, no. 8, pp. 855-866. American Society of Photogrammetry, Washington.

Brown, M. and Lowe, D. (2002). Invariant features from interest point groups. *Proceedings of the 13th British Machine Vision Conference*, pp. 253-262. BMVA, Cardiff, UK.

Dixon, D. (2010). *Augmented Reality Goes Mobile* [Internet]. Available from: http://www.manifest-tech.com/society/augmented_reality.htm. [Accessed 22 November 2013].

Fischler, M. and Bolles, R. (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, vol. 24, no. 6, pp. 381-395. ACM Press.

Goldiez, B. (2004). *Techniques for Assessing and Imroving Performance in Navigation and Wayfinding Using Mobile Augmented Reality*, Ph.D. Dissertation. University of Central Florida, College of Arts and Sciences, Orlando.

Grammatikopoulos, L. (2007). *Geometric Information from Single Images in Photogrammetry and Computer Vision*, PhD Thesis. National Technical University of Athens, School of Rural and Surveying Engineering, Athens, Greece.

Kipper, G. and Rampolla, J. (2013). *Augmented Reality: An Emerging Technologies Guide to AR*. Syngress, USA.

Kooper, R. and MacIntyre, B. (2003). Browsing the Real-World Wide Web Browser: Maintaining Awareness of Virtual Information in an AR Information Space. *International Journal of Human - Computer Interaction,* vol. 16, no. 3, pp. 425-446. CSC Journals, Kuala Lumpur, Malaysia.

Koussoulakou, A., Patias, P., Sechidis, L and Stylianidis, E. (2001). Desktop Cartographic Augmented Reality: 3D Mapping and Inverse Photogrammetry in Convergence.

Proceedings of the 20th International Cartographic Association Conference. pp, 2506-2513. International Cartographic Association (ICA).

Levenberg, K. (1944). A Method for the Solution of Certain Non-Linear Problems in Least Squares. *Quarterly of Applied Mathematics*, vol. 2, no. 2, pp. 164-168. American Mathematical Society, Boston, MA.

Leventi, I. (2013). *Location Based Services (LBS) and Augmented Reality (AR),* Master's Thesis. National Technical University of Athens, School of Rural and Surveying Engineering, Athens, Greece.

Marquardt, D. W. (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431-441. Society for Industrial and Applied Mathematics, USA.

McGlone, J. C., Mikhail, E. M. and Bethel, J. (2004). *Manual of Photogrammetry*, 5th ed. American Society for Photogrammetry and Remote Sensing, USA.

Neubeck, A. and Van Gool, L. (2006). Efficient Non-Maximum Suppression. *Proceedings of the 18th International Conference on Pattern Recognition*, vol. 3, pp. 850-855. IEEE Computer Society, USA.

Portalés, C., Lerma, J. L. and Navarro, S. (2009). Augmented reality and photogrammetry: A synergy to visualize physical and virtual city environments. *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, issue 1, pp. 134-142. Elsevier.

Suzuki, S. and Abe, K. (1985). Topological Structural Analysis of Digitized Binary Images by Border Following, *Computer Vision, Graphics and Image Processing*, vol. 30, no. 1, pp. 32-46. Academic Press Professional Inc., San Diego, CA.

Thomas, B., Victor, D., Wayne, P., David, H. and Bernard, G. (1998). A Wearable Computer System with Augmented Reality to Support Terrestrial Navigation, *Proceedings of the 2nd IEEE International Symposium on Wearable Computers*, pp. 168-171. IEEE Computer Society, USA.

Viola, P. and Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 511-518. IEEE Computer Society, USA.

Vlahakis, V., Ioannidis, N., Karigiannis, J., Tsotros, M., Gounaris, M., Stricker, D., Gleue, T., Daehne, P. and Almeida, L. (2002). Archeoguide: An Augmented Reality Guide for Archeological Sites. *IEEE Computer Graphics and Applications*, vol. 22, no. 5, pp. 52-60. IEEE Computer Society, USA.

Zhang, Z. (2000). A Flexible New Technique for Camera Calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no 11, pp. 1330-1334. IEEE Computer Society, USA.

BIOGRAPHICAL NOTES

Charalabos IOANNIDIS
Associate Professor at the Lab. of Photogrammetry, School of Rural and Surveying Engineering, National Technical University of Athens (NTUA), Greece, teaching photogrammetry and cadastre. Until 1996 he worked at private sector.
1992-1996: Co-chairman of Commission VI-WG2 'Computer Assisted Teaching' in ISPRS.
1997-2001: Member of the Directing Council of Hellenic Mapping and Cadastral Organization and Deputy Project Manager of the Hellenic Cadastre.
2010-2014: Chair of Working Group 3.2 'Technical Aspects of SIM' of FIG Com 3.
His research interests focus on terrestrial and satellite photogrammetry, aerial triangulations, digital orthophotos, applications of digital photogrammetry on the cadastre and GIS. He has authored more than 140 papers in the above fields, and has given lectures in related seminars both in Greece and abroad.

Styliani VERYKOKOU
Surveyor Engineer, postgraduate student at School of Rural and Surveying Engineering, National Technical University of Athens (NTUA), Greece.
She was awarded three scholarships by the National Institute of Scholarships of Greece as a result of her performance in her studies.
Her research interests lie in the fields of photogrammetry, computer vision, cartography, remote sensing and augmented reality.

CONTACTS

Ass. Prof. **Charalabos Ioannidis**
National Technical University of Athens (NTUA)
9 Iroon Polytechniou St.
Athens
GREECE
Tel. +302107722686
Fax +302107722677
Email: cioannid@survey.ntua.gr

Styliani Verykokou
National Technical University of Athens (NTUA)
20 Theonos St., 11743
Athens
GREECE
Tel. + 302109220703
Email: st.verykokou@gmail.com