Recovering the Full Pose from a Single Keyframe

Pierre Fite Georgel* CAMP - TU München
Selim Benhimane METAIO
Jürgen Sotke CAMP - TU München
Nassir Navab CAMP - TU München

Abstract

Photo-based Augmentation is a growing field in particular for Industrial Augmented Reality (IAR) applications. Registration is at the core of every photo-based AR software. This alignment of the image to the 3D model coordinate system is usually achieved with fiducial markers. When a single keyframe is used, the unknown baseline length has to be estimated in order to superimpose virtual models onto the image. In this paper, we develop an automatic algorithm to augment the relative pose, estimated using a single keyframe, into a full pose that will permit superimposition. This is performed by propagating known 2D-3D correspondences to the target image using perspectively corrected template matching and followed by a refinement of the estimated full pose that combines geometric and photometric information. The performance and the stability of the proposed method is extensively demonstrated on synthetic data and its applicability is shown within an industrial AR software for Visual Inspection and Documentation.

1. Introduction

Augmented Reality (AR) [2] has progressed over the last two decades and starts to be applied in different industrial applications over the recent years. Traditional AR focuses on augmenting the reality with virtual data. The connection to reality can be direct with see-through head mounted displays [16]or indirect using video-stream with video see-through [17] head-mount or tablet pc [15].These powerful methods however do not always integrate well in industrial workflow. Therefore a new trend appeared where instead of having continuous augmentation only single images are augmented. One could call this Photo-based AR.
Stricker and Navab [18] augment industrial pictures with CAD information to help architectural planning. For the registration, the user manually selects corresponding points between the image and the model. Appel et al. [1] use photographs of plants to redesign a CAD model and improve documentation. They use correspondences between technical drawings and images to perform the registration. Pentenrieder et al. [13] are using photo-based AR for factory planning. They use pictures of a factory to verify if this factory can be used to fabricate another component for example the new model of a car. They use fiducial markers to register the images to the CAD reference system. Georgel et al. [8], use natural landmarks to perform the registration and use AR to find discrepancies between the CAD model and the actual plant.
In this paper, we present a new method to automatically register an image to a CAD model using a *single* keyframe because pose estimation using CAD Model [8] is rarely automatic and hardly handles missing or wrong data [4]. We use state of the art computer vision techniques to estimate a relative pose from the keyframe to the image to register. Unfortunately, this relative pose lacks one degree of freedom, the length of the baseline preventing it from being used for augmentation. We estimate the length of this baseline automatically by using template matching to propagate 2D-3D correspondences established within the keyframe. The template matching searches photometrically correct correspondences while varying the baseline length. The parameterization used for the warping handles all perspective distortions. Then, using this initial estimate, we perform a nonlinear full pose refinement combining geometric and photometric information. Our method automatically estimates a full pose which can be used for direct augmentation.

2. Related Work

Keyframes are commonly applied in tracking for AR applications to increase efficiency and reliability. Keyframes are still images that have been pre-registered to the model. Vacchetti et al. [19] for example uses several keyframes and local bundle adjustment to obtain a full pose for the current frame. Similarly, Chia et al. [3] use two or three keyframes simultaneously to estimate the full pose (i.e. in regard to the model and not the relative pose to another frame) of the current image. Stricker and Navab [18] use their previously registered image to estimate the pose and the change in focal length/zoom of the camera. They solve the scale problem by interactively selecting corresponding point between the image and the model. Platonov et al. [14] uses a set

*e-mail: pierre.georgel@gmail.com

of pre-registered keyframes to estimate a relative pose for the current frame. The extension to a full pose is then automatically computed using the 3D model. The real-scale is estimated based on the depth of triangulated feature points projected onto the model. Georgel et al. [7] also use a single keyframe to estimate the relative pose of the target image and extend it to a full pose by matching reconstructed planes from the feature points to planar structures in the model.
The novelty of our approach lies in the fact that it only uses a single keyframe and limited 3D information: a sparse set of points. Actually we only require one 2D-3D correspondence from the keyframe such a low amount of data from the CAD model would cause all the previous method to fail. Additionally, it refines the estimated nonlinear parameters using both photometric and geometric information from the images.

3. Theoretical Background

The motion between two calibrated cameras (i.e. cameras with known internal parameters) is described by the essential matrix $\mathbf{E}$, which relates points $\mathbf{p}$ from the keyframe to point $\mathbf{q}$ on the epipolar lines $\mathbf{l}$ in the target image $\mathbf{q}^\top \mathbf{K}_t^{-\top} \mathbf{E} \mathbf{K}_s^{-1} \mathbf{p} = 0$, where $\mathbf{K}_s$ (resp. $\mathbf{K}_t$) is the matrix of intrinsic parameters for the keyframe (resp. target). This matrix can be decomposed [10] into the product of a skew matrix and an orthogonal matrix as follows:

$$\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}. \tag{1}$$

Each matrix $\mathbf{E}$ leads to four possible decompositions. This ambiguity is solved using the points correspondences, because only the physically correct set of rotation and translation triangulates the image points in front of the cameras. Hence, we will from now on suppose that we have access to the correct decomposition. From (1) we can already grasp the problem we solve within this paper: $\mathbf{t}$ can only be determined up to scale. Therefore, we suppose that $\mathbf{t}$ is a unit vector for which we have to find the correct scaling.
The initial essential matrix is usually estimated using the 8-point algorithm by [10] and is then refined using a nonlinear scheme which minimizes a quadratic geometric cost function, the so-called re-projection error:

$$\mathscr{C}_{\mathscr{g}}(\mathbf{M}_i, \mathbf{R}, \mathbf{t}) = \sum_{i=1}^{n} \begin{array}{l} \|\mathbf{K}_s \mathbf{w}(\mathbf{M}_i) - \mathbf{p}_i\|^2 \\ + \quad \|\mathbf{K}_t \mathbf{w}(\mathbf{R}\mathbf{M}_i + \mathbf{t}) - \mathbf{q}_i\|^2 \end{array}, \tag{2}$$

with $\mathbf{w}\left([x, y, z]^\top\right) = [x/z, y/z, 1]^\top$ being the warping function, $\mathbf{M}_i$ the 3D points triangulation of the correspondences $(\mathbf{p}_i, \mathbf{q}_i)$ and n the number of point correspondences. Note that this method does not estimate the true scale of the observed structure because this cost is invariant to changes in scales ($\forall s \neq 0, \quad \mathscr{C}_{\mathscr{g}}(s\mathbf{M}_i, \mathbf{R}, s\mathbf{t}) = \mathscr{C}_{\mathscr{g}}(\mathbf{M}_i, \mathbf{R}, \mathbf{t})$).
In this the paper we will therefore focus on the estimation of the unknown translation length/scale s which is necessary to augment the target image. The standard method to recover the scale s is to manually set a known 3D point or a known 3D distance. These two methods will be briefly described in the next two subsections.

3.1. Scale from a 3D point

Using a given correspondence between a point $\mathbf{q}$ in the target image and a 3D point $\mathbf{M}$ which satisfies $\mathbf{R}\mathbf{M} + s\mathbf{t} \propto \mathbf{m}'$ with $\mathbf{m}' = \mathbf{K}_t^{-1}\mathbf{q}$, the translation scale can be deduced as follows:

$$s = -\frac{\left([\mathbf{m}']_\times \mathbf{t}\right)^\top [\mathbf{m}']_\times \mathbf{R}\mathbf{M}}{\left\|[\mathbf{m}']_\times \mathbf{t}\right\|^2} \tag{3}$$

3.2. Scale from a 3D distance

The standard approach is to use a known 3D distance. The norm of a 3D reconstructed (using the unit translation) segment visible in both image is estimated and then the ratio between the obtained norm and the known 3D distance gives the scale s. Since $\mathbf{R}\mathbf{M} + \mathbf{t} \propto \mathbf{m}'$ we can deduce:

$$\left[\mathbf{m}'\right]_\times (z\mathbf{R}\mathbf{m} + s\mathbf{t}) = \mathbf{0} \Rightarrow z\left[\mathbf{m}'\right]_\times \mathbf{R}\mathbf{m} = -s\left[\mathbf{m}'\right]_\times \mathbf{t}, \tag{4}$$

with $\mathbf{m} = \mathbf{w}(\mathbf{M})$.
By defining $\mathbf{a} = [\mathbf{m}']_\times \mathbf{R}\mathbf{m}$ and $\mathbf{b} = [\mathbf{m}']_\times \mathbf{t}$, we can express the depth of a 3D point as:

$$z = -s\frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|^2} \tag{5}$$

Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be the triangulation of the two pairs $(\mathbf{p}_1, \mathbf{q}_1)$ and $(\mathbf{p}_2, \mathbf{q}_2)$ using $\mathbf{R}$ and $\mathbf{t}$. Since $s > 0$ this leads to:

$$\|\mathbf{M}_1 - \mathbf{M}_2\| = \|z_1\mathbf{m}_1 - z_2\mathbf{m}_2\| = s\left\|\frac{\mathbf{a}_1^\top \mathbf{b}_1}{\|\mathbf{a}_1\|^2}\mathbf{m}_1 - \frac{\mathbf{a}_2^\top \mathbf{b}_2}{\|\mathbf{a}_2\|^2}\mathbf{m}_2\right\|. \tag{6}$$

Knowing the $\|\mathbf{M}_1 - \mathbf{M}_2\|$ one can find the scale s of the translation.

Both of these approaches are neither automatic nor make use of information linked to the original keyframe established during its registration. In the next section, we will introduce our method that makes use of such embedded information.

4. Our Method

We henceforth assume that we have access to a number of correspondences in the keyframe with the model expressed as $(\mathbf{c}, \mathbf{C})$. These pairs of 2D point $\mathbf{c}$ and 3D point $\mathbf{C}$ are usually estimated during the registration of the keyframe. Let $\mathbf{l}$ be the epipolar line induced by the point $\mathbf{c}$ in the target image. All points $\mathbf{c}'$ on $\mathbf{l}$ correspond to a unique

scale s. Similar to equation (3), this bijective relation is deduced using $\mathbf{RC}+s\mathbf{t} \propto \mathbf{K}_t^{-1}\mathbf{c}' = \mathbf{m}'$ as follows:

$$\forall \mathbf{c}' \in \mathbf{l},\ \left[\mathbf{m}'\right]_\times (\mathbf{RC}+s\mathbf{t}) = \mathbf{0} \Rightarrow s = -\frac{\left([\mathbf{m}']_\times \mathbf{t}\right)^\top [\mathbf{m}']_\times \mathbf{RC}}{\left\|[\mathbf{m}']_\times \mathbf{t}\right\|^2}, \tag{7}$$

this relation is true for all points that satisfy $[\mathbf{m}']_\times \mathbf{t} \neq \mathbf{0}$. This would only happen for points on the epipole which anyway cannot be reconstructed.
Furthermore, if we suppose that $\mathbf{C}$ is locally planar and that $\mathbf{n}$ ($\|\mathbf{n}\| = 1$) is a normal vector to this plane (which can be obtained from a CAD model), each point $\mathbf{C}$ induces a set of homographies

$$\mathbf{H}(s, \pi_\mathbf{C}) = \mathbf{R} - s\frac{\mathbf{t}\mathbf{n}^\top}{d}, \tag{8}$$

between the keyframe and the image to register, with $\pi_\mathbf{C} = \left[\mathbf{n}^\top, d\right]^\top$ the plane around $\mathbf{C}$ and d being the distance between the point $\mathscr{X}$ and camera center of the keyframe.
For each homography, we have a one to one mapping between neighbors of $\mathbf{c}$ and the neighbors of $\mathbf{c}'$. Therefore, it is possible to define an intensity based criterion to match $\mathbf{c}$ to the right $\mathbf{c}'$. Our template matching score $f(s)$ is defined as follows:

$$f(s) = \mathrm{SM}\left(\mathscr{S}, \mathbf{H}^{-1}(s, \pi_\mathbf{C})(\mathscr{T})\right), \tag{9}$$

with $\mathscr{S}$ and $\mathscr{T}$ being two image patches and SM being any similarity measure. The template search can be then expressed as an extremum search on f. This is made possible because (7) guarantees a unique s for each points of $\mathbf{l}$. So finding the scale s can be summarized as computing f for each $\mathbf{c}' \in \mathbf{l}$ and looking for the extremum of the function. A schematic of the search is shown in Figure 1. This discrete search is then refined by minimizing a nonlinear cost that combines both geometric and photometric information as expressed in the following section.

4.1. Nonlinear Refinement

If the propagated correspondences $\left(\mathbf{c}'_j, \mathbf{C}_j\right)$ were added to the geometric cost (2), it would stay optimal because it was optimal for $(\mathbf{M}_i, \mathbf{R}, \mathbf{t})$ and the propagation $\left(\mathbf{c}'_j, \mathbf{C}_j\right)$ have been selected to verify $\mathbf{K}_t \mathbf{w}(\mathbf{RC}_i + s\mathbf{t}) = \mathbf{c}'_i$. But the selected scale s is not optimal for (9) because it was sampled over the epipolar lines. Therefore it needs to be refined.

Using a similar approach as [6] which combines geometric and photometric information we create hybrid cost function that will estimate the full pose.
First we define a photometric cost function based on the template matching results by:

$$\mathscr{C}_\mathscr{P}(\mathbf{R}, \mathbf{t}) = \sum_{j=1}^{m} \sum_{\mathbf{X}}^{\mathscr{N}_{\mathbf{C}_j}} \left\| \mathscr{S}(\mathbf{K}_s \mathbf{w}(\mathbf{X})) - \mathscr{T}(\mathbf{K}_t \mathbf{w}(\mathbf{RX}+\mathbf{t})) \right\|^2, \tag{10}$$

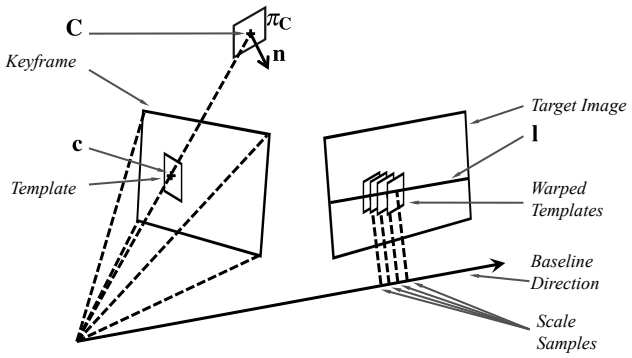

Figure 1. Scale from one propagated 2D-3D correspondence: the 3D point $\mathbf{C}$ projects on $\mathbf{c}$ in the *keyframe* and $\mathbf{c}$ maps to the epipolar line $\mathbf{l}$ in the *target image*. The template matching is performed between the *template* around $\mathbf{c}$ and *warped templates* on $\mathbf{l}$. The warp is parametrized using the plan $\pi_\mathbf{C}$ and the *scale samples*.

with $\mathscr{N}_{\mathbf{C}_j}$ being the 3D neighborhood of $\mathbf{C}_j$ defined by the plane $\pi_{\mathbf{C}_j}$.
This leads to the following least square minimization problem:

$$\arg\min_{\mathbf{M}_i, \mathbf{R}, \mathbf{t}} \mathscr{C}_\mathscr{G}(\mathbf{M}_i, \mathbf{R}, \mathbf{t}) + \mathscr{C}_\mathscr{P}(\mathbf{R}, \mathbf{t}). \tag{11}$$

We would like to emphasize two important facts. First that such a cost function will be only convex if it is well initialized. Therefore, it is mandatory to perform the scale search using the template matching. And second that the problem's formulation does not have any gauge freedom [12] since we are estimating a full pose using real 3D data. This is one of the main difference to [6] where they have to force the scale in order to obtain a stable minimization. In the next section implementation details will be given and the performance of the approach will be evaluated.

5. Implementation Details

The experiments were all run using Matlab and then later integrated into our IAR software. The initial rotation $\mathbf{R}$, translation $\mathbf{t}$ (of unit norm) and structure $\mathbf{M}_i$ is oriented (e.g. the points are triangulated in front of the camera) and is supposed to be optimal for (2). Since we are oriented we only have to consider positive scale s. The method is split in three steps. First, the template matching is performed (i.e. discrete search). The epipolar line is sampled every 5 pixel. The template size is 32×32 pixels. The similarity measure used for (9) is the NCC (we search for a maximum) which handles changes in illumination and the associated threshold $\tau = 0.8$. Second, an initial scale is selected. We choose the best scale from the set of scales estimated from the template matching. This is performed by applying the cost (9) to all 2D-3D correspondences using all the obtained scales and by choosing the scale that maximized the sum of the score (2D-3D correspondences

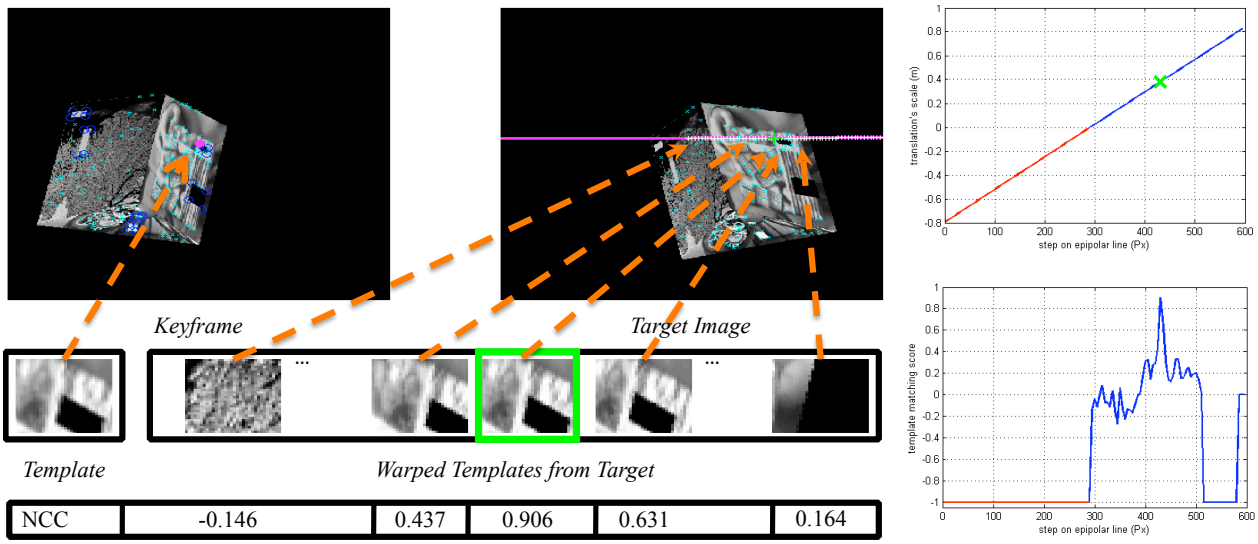

Figure 2. Template Matching Scheme: cyan crosses are Harris corner; blue circles 2D-3D correspondences; pink dot currently used 2D-3D correspondences which maps to pink epipolar line; and white crosses correspond to the samples. Samples relate to warped templates and the relative NCC scores at the bottom. The upper graph represents the evolution of the scale (negative scales in red are not considered) and the lower graph represents the NCC scores over all samples. (Object in green corresponds to the true scale)

with score lower than τ are discarded). This provides an initial estimate for the refinement and a set of matched 2D-3D correspondences. Finally, the full pose is refined, and we estimate the full pose based on equation (11). For the nonlinear minimization we normalize the gray scale intensity information between zero and one to give similar importance to $\mathscr{D}_{\mathscr{G}}$ and $\mathscr{D}_{\mathscr{P}}$. The algorithm is summarized in 1 and a schematic of the process is described in figure 2.

Input: $\mathbf{R}$, $\mathbf{t}$, the image point correspondences $(\mathbf{p}_i, \mathbf{q}_i, \mathbf{M}_i)$, the 3D-2D correspondences $(\mathbf{C}_j, \mathbf{c}_j)$ and τ
Output: $\widetilde{\mathbf{R}}$, $\widetilde{\mathbf{t}}$ and $\widetilde{\mathbf{M}}_i$
$\mathbf{F} \leftarrow \mathbf{K}_i^{-\top}[\mathbf{t}]_\times \mathbf{R}\mathbf{K}_s^{-1}$; $f_{total} \leftarrow 0$; $s_{best} \leftarrow 0$;
foreach $\mathbf{c}_j$ **do**
  $f_{max} \leftarrow 0$; $current \leftarrow 0$;
  foreach $\mathbf{c}' \in \mathscr{I}_{\mathscr{G}} \cup \mathbf{Fc}_j$ **do**
    Compute s from (7) ;
    if $(s > 0)\&(f_{max} > f(s, \mathbf{c}_j, \mathbf{c}'))$ **then**
      $f_{max} \leftarrow f(s)$; $s_{max} \leftarrow s$;
    end
  end
  if $f_{max} > \tau$ **then**
    foreach $\mathbf{c}_{j'}$ **do**
      if $f(s_{max}, \mathbf{c}_j, \mathbf{c}') > \tau$ **then**
        $current += f(s_{max}, \mathbf{c}_j, \mathbf{c}')$;
      end
    end
    if $current > f_{total}$ **then**
      $s_{best} \leftarrow s_{max}$; $f_{total} \leftarrow current$;
    end
  end
end
if $s_{best} > 0$ **then**
  $\mathbf{t} \leftarrow s_{best}\mathbf{t}$; $\mathbf{M}_i \leftarrow s_{best}\mathbf{M}_i$ % Bring to scale ;
  Estimate $\left(\widetilde{\mathbf{R}}, \widetilde{\mathbf{t}}, \widetilde{\mathbf{M}}_i\right)$ with$\{\mathbf{C}_j\}$ using (11);
end

Algorithm 1: Propagation of 2D-3D correspondences from a keyframe to target image to obtain a full pose.

6. Results

All the experiments presented in this section are designed to evaluate the precision and stability of the presented method. First we will focus on synthetic simulations and then demonstrate the use of the method in industrial AR application.

6.1. Synthetic Simulations

The experiments are based on a synthetic model which is formed of 3 textured planes. The motion between the images is perfectly known and is our ground truth data. Harris corners [9] were detected on the keyframe and then propagated to the target image using the true motion. At most 154 Harris corners are used during the experiments for visibility reasons (this number can be smaller). On each plane of the synthetic model eight 2D-3D correspondences are marked, each correspondence is link to a normal vector. The number of 2D-3D correspondences used is at most 24 (if not specified otherwise) depending on their visibility. Examples of generated views are visible in figure 2. The changes in depth of the viewpoints was almost constant across all experiments (if not specified otherwise) in order to have a comparable reprojection error between different experiments.

The error of an estimation is measured using the true (i.e. noiseless) 2D-3D correspondences and the estimated full pose. We project the 3D points with the estimated full pose and measure the distances to the 2D points obtained from the ground truth, the mean of this distance will be our error model, this is often called the mean target reprojection er-

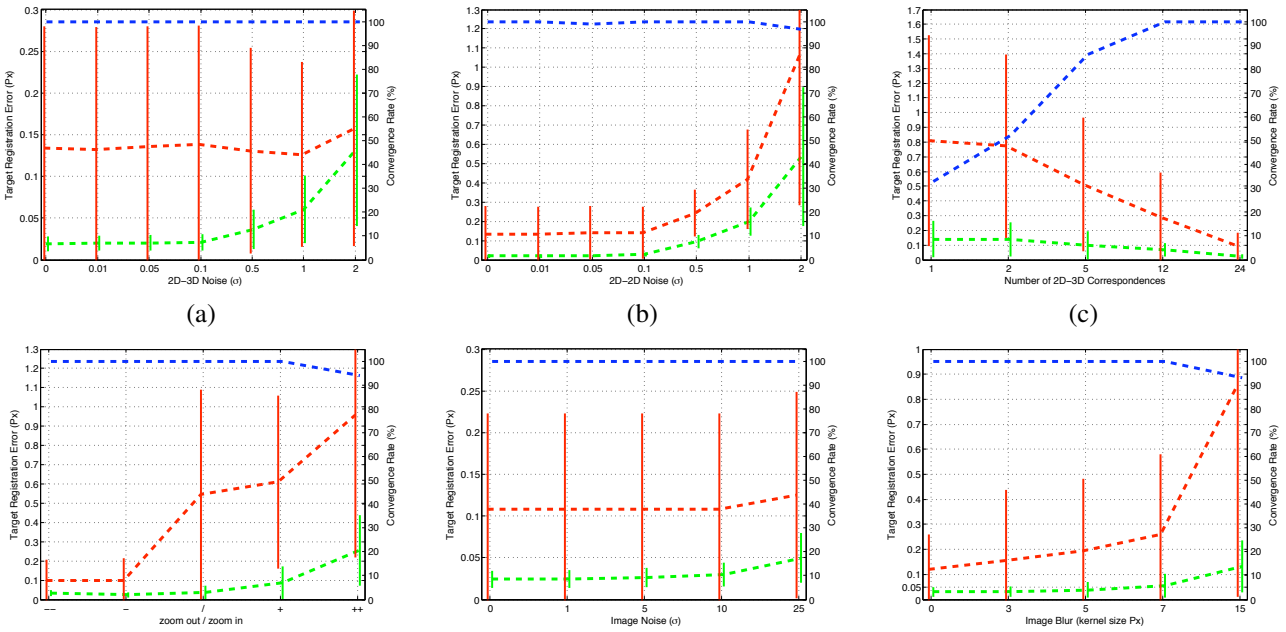

Figure 3. Experiments on synthetic data: the red line represents the mean target registration error (RE) after the template search, the green the RE after the nonlinear refinement, the blue the convergence rate and vertical bars standard deviations. The graphs show the results of experiments with (a) noise in the 2D-3D correspondences; (b) noise in the 2D-2D correspondences; (c) increasing number of 2D-3D correspondences; (d) change in scale (from strong zoom-out on the left to strong zoom-in) (e) image noise; (f) image blur.

ror (RE). Whether the algorithm converged will be decided based on this RE. It is declared that the method converged if the RE is bellow the noise level plus precision limit: 0.05 pixel (this value is explained in 6.1.1). All the mean RE plots are based RE that converged. Six experiments were conducted which will now be explained.

6.1.1 Error in the Keyframe Registration

In order to verify the stability of the method towards misregistration of the keyframe, we simulated slightly wrong alignment between the 3D points and the 2D points. This is performed by adding Gaussian noise to the 2D points and then compute the pose of the keyframe. This induced a small error in the alignment between the model and the image, which includes an error in the orientation of the normal. We ran these experiments with seven different levels of noise on one hundred images. The results are summarized in figure 3(a).

The algorithm always converged to good solution with respect to the input noise. The refinement step always improves the result of the template matching. Furthermore, the resulting RE was small. The experiment also revealed the numerical limitation of the use of this intensity.A target registration error bellow 0.05 was rarely achieved. We speculate that this originates from the discrete method used to create the images and the loss of information after warping.

6.1.2 Error in the Relative Pose

In a second experiment, we tested the stability of the method against an error in the relative pose estimate because it is more than likely that the feature points used to obtain the relative pose are localized with some error. In order to simulate this error, a Gaussian noise was added to the Harris corner (2D points) in both keyframe and target. This had a direct impact on the quality of the initial relative pose (even with the use of the Gold Standard algorithm [10]). We ran these experiments with seven different noise levels on one hundred images. The obtained convergence rates and RE of the approach is summarized in figure 3(b).

The method rarely diverges (up to 4%). We assume that the algorithm only diverged when the error in the relative pose was so large that the epipolar lines resulting from the image points (of the 2D-3D correspondences) miss their true corresponding points by far. Again we see that the refinement step drastically improves the result obtained by the template matching. Furthermore, in comparison to the previous experiment 6.1.1 the results obtained with the template matching diverge from the refinement step results. This can be explained by the fact that we use additional information (photometric) during the nonlinear optimization which corrects the wrong measurement included.

6.1.3 Stability with respect to Number of 2D-3D Correspondences

We studied the impact of the number of 2D-3D correspondences used for the method. We randomly select a subset of the 2D-3D correspondences. The threshold for the convergence was set to 1 pixel error because the lower bound threshold was selected using 24 2D-3D correspondences. Such a precision however cannot be expected for sparse correspondences. We used one hundreds poses, and the results are presented in 3(c).
The first comment, that can be drawn from this experiment, is that the number of 2D-3D correspondences has a direct impact in the convergence rate. We suppose that the randomly selected 2D-3D correspondences were not always well visible (e.g. perspectively too distorted to be recognized in the target image). Such perspective distortion rarely happens with real images because the relative pose is often not computable in this case. The second comment is that the precision of the method is satisfactory even when only one correspondence is available.

6.1.4 Stability to Scale Change

We then tested the effect of the scale changes between the keyframe and the target. We sampled poses with five different zoom factors. Again, we used a threshold of 1 pixel error. The result are presented in 3(d).
The differences in error is the consequence of the varying scale. When the object is closer to the camera the RE increase (automatically) even if the underlying error in pose is the same. Often the focal is increased to compensate for this phenomena; we decided to not apply this idea in order to let the experiment describe the underlying problem. This experiment shows that the proposed method is stable even when the scale factor varies drastically.

6.1.5 Noise and Image Blur

To verify the performance of the approach in actual usage, one needs to know its stability with respect to noise and blur. When images are acquired using a camera, noise is always present because of diffuse illumination, and blur can occur when the camera is handled by a human and not fix on a tripod. We perform two experiments. First, we add an independent Gaussian noise to the intensity of the keyframe and the target. Second we blur the keyframe and the target. The blur is performed using an increasing kernel size. Both experiments were performed on more than hundreds of images. The convergence threshold for both experiment was 0.5 pixel. The results for the noise is visible in figure 3(e) and for the blur in figure 3(f).
The noise experiments demonstrate again than the refinement step is crucial to obtain a precise full pose. Furthermore, we can see that the standard deviation of the resulting error is small which proves that we always reach a stable optimum even with massive disturbances. This is because we simultaneously minimize the photometric and geometric cost over all the observations. The method does a good job of handling the loss of information due to the blurring effect. The obtained RE after the refinement, at maximum, doubles from the non-blurred image (smaller that 0.2 pixel).

6.2. Industrial Application

We implemented the presented method within our Industrial Augmented Reality Software. The goal of the software is to provide a better documentation of the CAD model of power-plants and to help performing a discrepancy check (i.e. verify correctness of the built item compared to the planned CAD model). Giving access to images of the built plant offers insights about undocumented features (for example electrical wires), misplaced or modified components. This visualization can be achieved in any CAD software having texture capabilities to display the images, but such software will then requires well designed methods to register images to the coordinate system of the CAD model. This was one of the main motivation to develop an automatic method to register images using a single keyframe.

The keyframes were registered using anchor-plates [8]. Anchor-plates are metallic plates embedded in concrete structures (walls, ceiling and floor) of rectangular shape used to fix different components. The corners of these anchor-plates and their corresponding image points are used as 2D-3D correspondences.The relative pose is estimated using SIFT points [11], RANSAC [5] and the Gold Standard algorithm in order to obtain a relative pose. In order to register an image the user has to select a keyframe.The method has been successfully used on powerplant's images over the past two years. Some results are visible in figure 4.

6.3. Discussion

Experiments demonstrate that the proposed method is stable during zooming, and is robust to noise as well as blur. The main reason for such accurate performances is the use of a well initialized hybrid nonlinear refinement that handles different types of data and therefore can deal with the erroneous measurements in intensities and/or in 2D coordinates.
An interesting direction for future work could be investigated is the variation of the scale parameter s as a function of the baseline. The example in figure 2 shows a linear variation of the scale when the matching point moves along the epipolar line, but this is not always the case. One could try to use the monotonicity of the variation along the epipolar line to perform the search of the scale more efficiently. Note that the nonlinear 6 DoF estimation could be applied to bundle adjustment when 2D-3D correspondences are available.

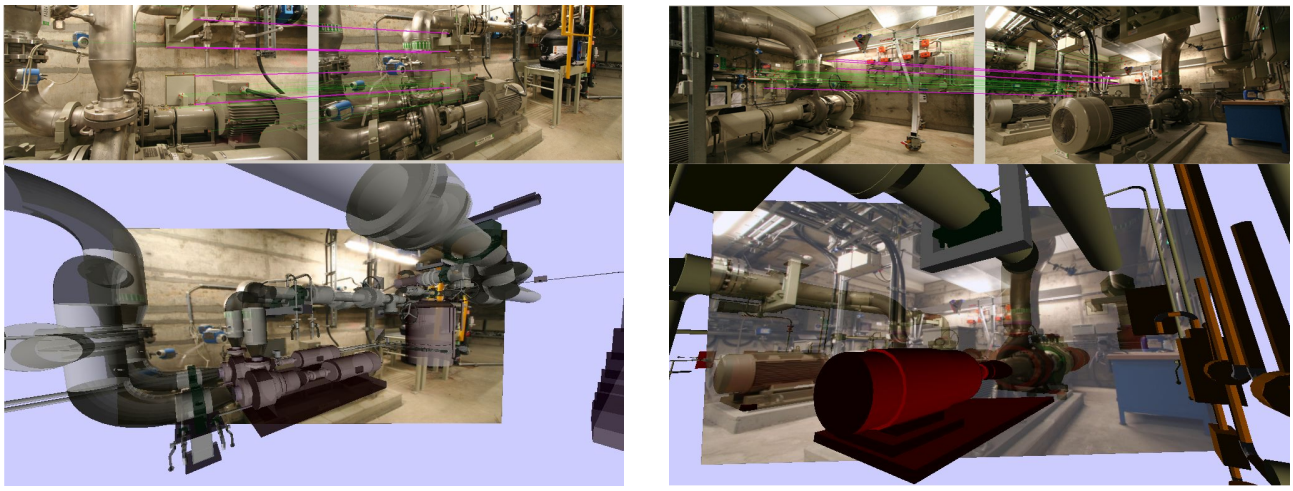

Figure 4. Industrial Application: (top) The matching and propagation results: propagated 2D-3D correspondences in pink (left the keyframe, right the target), unmatched 3D points in yellow; matched features in green; (bottom) The resulting augmentation.

7. Conclusion

In this paper, we present an automatic method to extend a relative pose to a full pose. The relative pose is sufficient for many Computer Vision applications. However, in Augmented Reality the full pose is needed to correctly superimpose the virtual object onto the real view of the world. In such applications, relative pose is of limited use.

The method introduces a homographic warp that is parametrized by the translation length. It uses an hybrid 6 DoF pose estimation which has no gauge freedom. The hybrid pose estimation minimizes at the same time intensity differences and the reprojection error. We have demonstrated through extensive synthetic experiments the robustness and precision of the proposed method and shown its applicability in the context of an industrial photo-based augmented reality application. Not requiring multiple pre-registered images or multiple 2D-3D correspondences greatly broaden the application of keyframe for Photo-based Augmented Reality.

References

[1] M. Appel and N. Navab. Registration of Technical Drawings and Calibrated Images for Industrial Augmented Reality. *Machine Vision and Applications*, 13(3):111–118, 2002.

[2] R. Azuma. A survey of Augmented Reality. *Presence*, 6(4):355–385, 1997.

[3] K. Chia, A. Cheok, and S. Prince. Online 6 DoF Augmented Reality Registration from Natural Features. In *ISMAR*, 2002.

[4] T. Drummond and R. Cipolla. Real-time tracking of complex structures with on-line camera calibration. In *BMVC*, 1999.

[5] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Com. of ACM*, 24(6), 1981.

[6] P. Georgel, S. Benhimane, and N. Navab. A Unified Approach Combining Photometric and Geometric Information for Pose Estimation. In *BMVC*, 2008.

[7] P. Georgel, P. Schroeder, S. Benhimane, M. Appel, and N. Navab. How to Augment the Second Image? Recovery of the translation scale in image to image registration. In *ISMAR*, 2008.

[8] P. Georgel, P. Schroeder, S. Benhimane, S. Hinterstoisser, M. Appel, and N. Navab. An Industrial Augmented Reality Solution For Discrepancy Check. In *ISMAR*, 2007.

[9] C. Harris and M. Stephens. A combined corner and edge detection. In *Proc. of Alvey Vision Conference*, 1988.

[10] R. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision (2nd ed.). *Cambridge press*, 2003.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[12] D. D. Morris, K. Kanatani, and T. Kanade. Gauge fixing for accurate 3d estimation. *CVPR*, 2001.

[13] K. Pentenrieder, C. Bade, F. Doil, and P. Meier. Augmented Reality-based Factory Planning - an Application Tailored to Industrial Needs. In *ISMAR*, 2007.

[14] J. Platonov, H. Heibel, P. Meier, and B. Grollmann. A mobile markerless AR system for maintenance and repair. In *ISMAR*, 2006.

[15] G. Schall, E. Mendez, and D. Schmalstieg. Virtual Redlining for Civil Engineering in Real Environments. In *ISMAR*, 2008.

[16] B. Schwerdtfeger, , and G. Klinker. Supporting Order Picking with Augmented Reality. In *ISMAR*, 2008.

[17] T. Sielhorst, M. Feuerstein, and N. Navab. Advanced Medical Displays: A Literature Review of Augmented Reality. *J. Display Technology*, 4:451–467, 2008.

[18] D. Stricker and N. Navab. Calibration Propagation for Image Augmentation. In *IWAR*, 1999.

[19] L. Vacchetti, V. Lepetit, and P. Fua. Stable Real-time 3d Tracking using Online and Offline Information. *IEEE Trans. PAMI*, 26(10):1385–1391, 2004.