TRIANGULATION-BASED VIDEO REGISTRATION AND APPLICATIONS

Dimitri Bulatov

Research Institute for Optronics and Pattern Recognition,
Gutleuthausstr. 1, 76275 Ettlingen, Germany
bulatov@fom.fgan.de

KEY WORDS: disparity, geo-referencing, mosaicking, registration, reconstruction, triangulation, video

ABSTRACT:

We present a simple and effectient procedure for registration of video sequences onto high-resolution images for scenes with distinct spatial depth. The high-resolution image could be a synthetic view of a 3D-object representing a piecewise continuous surface. The key issue of the procedure is the inter-sensorial registration which contains 1. computing characteristic points in the video frame and in the image, 2. establishing putative correspondences, 3. obtaining a relative orientation by computing fundamntal matrix, 4. image rectification, 5. computing additional correspondences, and 6. retrieving the triangular maps. We visualize the algorithm and its main applications by two examples and discuss the main advantages and drawbacks of our methods as well as directions of future research.

1 INTRODUCTION AND PRELIMINARIES

Integration of images and videos into geo-referenced 2D- and 3D-models given by high-resolution images or data-bases of different kind is a very important task for many practical applications of computer vision and image processing. Especially in the areas of surveillance, automatic navigation and defense research, a considerable progress was made in development of fast algorithms for geo-referencing. The key issue in many of these problems is to transform every single frame of the video sequence into the coordinate system of the given model, a process called *inter-sensorial registration*. Geometrically, there are three main problems associated with this task: 1. a *2D-2D registration*, where the depth of the scene is negligible and an image-to-image homography is a suitable model for registration, 2. a *3D-3D registration*, where point clouds recovered from image sequences can be used to support the registration, and 3. a *2D-3D registration*, where the depth between sensors is no longer negligible, however due to time factor or a critical geometric configuration, a 3D-reconstruction cannot be performed. Examples for the first kind of registration can be found in (Solbrig et. al., 2008), (Krueger, 2001) (features-based-registration), (Xiao et. at., 2005) (segment-based registration), and (Lin et. al., 2007) (registration by mutual information). In (Solbrig et. al., 2008), several examples of applications, such as motion detection or annotation of objects from data-bases in the video, are demonstrated. For the second kind of problems, we refer to (Martinec et. al., 2006), where the key idea consists of computing 3D-Eucliedean reconstructions from pairs of cameras and merging them into the same coordinate system. Several examples of successful reconstruction of rather complicated objects were presented in this paper. Another possibility of fusing 3D data sets is given by object based methods, such as ICP (Besl and McKay, 1992) and its modifications. The last kind of geometric configurations is a border case since it is not broadly investigated in the literature. Several segmentation-based methods (Xiao et. at., 2005) can perform segmentation of image into planes (layers) without explicitly computing the 3D structure of the scene. Also the efforts of (Schenk and Csathó, 2002) were concentrated on a small number of pictures of rather good quality and not the video sequence with (theoretically) unlimited number of frames. However, integration of video data recorded by cameras of average quality and without internal navigation into the given model in a reasonable time and with a reasonable memory load is particularly challenging.

In this work, we will present our procedure for referencing video data into a high-resolution image. The algorithm will be described in Section 2 and it will be a continuation of work in (Solbrig et. al., 2008). The difference is that we will follow the multi-homography procedure (based on triangular meshes and described in (Bulatov et. al., 2009)) in order to establish a pixel-to-pixel correspondences. An example of annotation of objects in the video will also be considered. The results from two similar data-sets of varying difficulty are presented in Section 3. These results intend to prove the principle of our method, even though they do not include the improvements outlined in Section 4, which also summarizes our work.

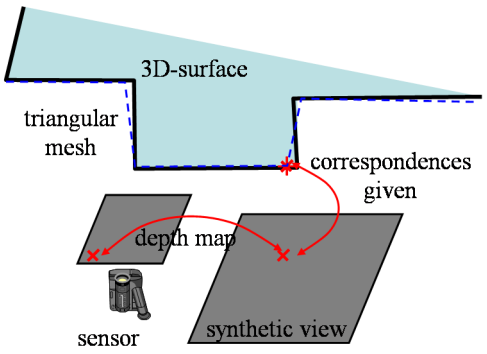

Figure 1: Principle of the geometric configuration. Depth information of every pixel of the synthetic view is assumed to be given, so the key task is to establish relation between pixels of the synthetic view and video frame

Before going into the details of the algorithm, we will describe the data sets we are dealing with and the assumption we make about the data. We assume that we are given a synthetic view of the model. We denote this (high-resolution) image by $\mathcal{J}$. Behind the image data, one can imagine 3D coordinates, as for example in the case of range data or a digital terrain model. Figure 1 visualize the principle of the described configuration. For the scope of this paper, we will exclude this additional information from consideration and concentrate our attention on registration of (almost) every pixel of the video sequence onto the high-resolution image. We do not assume any additional properties of the video: it can be made by a moving sensor or from a zooming/rotating camera (such as no 3D-reconstruction is possible). To initialize

the process of registration, we need the approximate position of the first frame in the video in $\mathcal{J}$; in the case when no GPS data is available, one can manually specify the bounding box of the video frame in the image $\mathcal{J}$.

2 REGISTRATION PROCEDURE AND APPLICATIONS

We denote by F_k the *fundamental matrix* which establishes the *inter-sensorial registration*, i. e. it assigns to a point in the k-th frame of the video sequence ($\mathcal{I}_k$) the epipolar line in the high-resolved image $\mathcal{J}$. For corresponding points $x_k \in \mathcal{I}_k$ and $X \in \mathcal{J}$ in homogeneous coordinates, we have the relation $X^T F_k x_k = 0$. The *homography* H is the 2D image transformation between two frames of the video sequence (*intra-sensorial registration*). For corresponding points $x_k \in \mathcal{I}_k$ and $x_{k+n} \in \mathcal{I}_{k+n}$, the relation $x_{k+n} = H_{k,k+n} x_k$ holds and

$$H_{k,k+n} = H_{k+n-1,k+n} \cdot \ldots \cdot H_{k,k+1}. \quad (1)$$

We will consider mostly homographies between neighboring frames (i.e. for small n) in order to be able to neglect the baseline and, consequently, the effects of depth. A detailed description of intra-sensorial and inter-sensorial registration is given in Sections 2.1 and 2.2, respectively. For more information on multiple view geometry, we refer to (Hartley and Zisserman, 2000).

2.1 Intra-sensorial registration

The intra-sensorial registration succeeds, similar to (Solbrig et. al., 2008) by tracking characteristic points from frame to frame by means of (Lucas and Kanade, 1981). In order to provide the inter-sensorial registration (see Section 2.2) and to save computational time needed for intra-sensorial registration, we use SIFT-points (Lowe, 2004), even though they are not very suitable for tracking. The homography is computed by a modification of RANSAC (Chum and Matas, 2004). Assuming that inter-sensorial registration is given by a fundamental matrix F_k, the next inter-sensorial registration is performed if one of following four conditions holds: 1. not all points of the convex hull of the initially detected points have been successfully tracked, which is a typical case for moving and rotating cameras, 2. the area of the polygon enclosed by that convex hull has become significantly different with respect to its initial value (typical for zooming cameras) 3. the number of inliers yielded by RANSAC is low, and 4. the number of frames performed after the last inter-sensorial registration exceeds a fixed number, typically 4 in our experiments. The coarse estimation of the fundamental matrix between $\mathcal{I}_{k+n}$ and $\mathcal{J}$ is given by $F_{k+n} = F_k \cdot H_{k,k+n}^{-1}$, with $H_{k,k+n}$ as in Equation (1).

2.2 Inter-sensorial registration

The key issue of this section is to register a frame $\mathcal{I}$ of the sequence onto $\mathcal{J}$. We assume that we can establish a number of correspondences of SIFT-points with cost function given by the scalar product of their descriptors. This assumption usually holds if $\mathcal{J}$ and $\mathcal{I}$ do not possess significant radiometric differences, i.e. if $\mathcal{J}$ was taken under approximately the same light and weather conditions. We assume that such an image $\mathcal{J}$ can be chosen from a data-base. Additionally, we can precompute SIFT points in $\mathcal{J}$ in order to accelerate the procedure. From the putative correspondences, we can obtain the relative orientation of two sensors by the 7-point-algorithm (Hartley and Zisserman, 2000) and RANSAC. In the next step, epipolar image rectification of a small patch of $\mathcal{J}_\mathcal{I} \subseteq \mathcal{J}$ and $\mathcal{I}$ is performed. After this step, corresponding pixels in transformed images $\mathcal{J}_\mathcal{I}$ and $\mathcal{I}$ have the same y-coordinate and we can obtain a denser set of correspondences either by computing disparity maps (e. g. by means of (Hirschmüller, 2005), a method which is relatively computationally efficient compared to other global methods) or by establishing additional point correspondences and then approximating disparity maps from triangulations. We follow the method of (Bulatov et. al., 2009) here because we assume that the synthetic view represents a piecewise continuous surface which can be approximated by triangles. The mentioned method has several advantages: it is very fast and less sensitive to differing radiometric information (since two pictures were made by different sensors) than those global and semi-global methods which estimate depths maps pixel per pixel. For this reason, we also do not perform segmentation and computation of similarities of triangles. As indicated in (Bulatov et. al., 2009), the disparity of the point $\mathbf{x} = (x^1, y^1)$ in the triangle with vertices $(x_1^1, y_1^1), (x_2^1, y_2^1), (x_3^1, y_3^1)$ in the first rectified image is given by:

$$d = \begin{pmatrix} x_1^2 & x_2^2 & x_3^2 \end{pmatrix} \begin{pmatrix} x_1^1 & x_2^1 & x_3^1 \\ y_1^1 & y_2^1 & y_3^1 \\ 1 & 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} x^1 \\ y^1 \\ 1 \end{pmatrix} - x^1, \quad (2)$$

where x_1^2, x_2^2, x_3^2 are the x-coordinates of the correspondences in the second image. We denote the first two terms of Equation (2) (independent on $\mathbf{x}$) by $\mathbf{v}$ and store the three entries of $\mathbf{v}$ for every triangle in the given triangulation. It is thus possible to assign to a pixel the number of triangle it belongs to and then its disparity by means of Equation (2).

Remark: We rectify the images $\mathcal{J}_\mathcal{I}$ and $\mathcal{I}$ by calculating two homographies (T_1 and T_2) according to the method of Zhang (Loop and Zhang, 1999). This is possible if and only if the epipoles (center of second camera projected by the first camera and vice versa) are bounded away from the domain of images. But even if the epipoles are within the image domain, our triangulation and multi-homography registration (unfortunately not the algorithm of Hirschmüller in its original implementation!) will succeed. The disadvantages in this case are difficulties in setting bounds for disparity ranges for enriching the point correspondences which we will explicate in the next paragraph as well as inner constraints for the eight entries of every local homography ($HF^T + FH^T = 0$, compare (Hartley and Zisserman, 2000), chapter 13) instead of three entries of $\mathbf{v}$ in Equation (2).

The process of *enriching* (also called guided matching) needed for increasing the number of point correspondences in two rectified images, works similar to (Bulatov et. al., 2009). We specify for a point in the first image a rectangle in which the correspondent point is expected. The width of the rectangle is specified by the disparity ranges of the correspondences obtained in the previous step as well as the bounds provided by edges of the given triangulation. The height of the rectangle is 3, because we only allow a deviation of 1 pixel from the corresponding epipolar lines. The only difference to the mentioned work is the matching cost function. Beside the calculation of normalized cross correlation for points near characteristic edges in the image, we experimented with SIFT points for which we selected putative correspondences in the rectangle mentioned above and then computed scalar products with all putative correspondences. The matches with the highest score were added into the list of correspondences if this score exceeded a fixed threshold (0.75 in our experiments). Both approaches gave results of similar quality. After guided matching, we perform the *Delaunay-triangulation* of points in the video frame and compute the approximation of disparity map according to (2).

In the last paragraph of this subsection we describe the initialization for the next inter-sensorial registration. Let S_k, S_{k+n} be the translation matrices for coordinates of points X of $\mathcal{J}$ into the

patches $\mathcal{J}_{\mathcal{I}_k}, \mathcal{J}_{\mathcal{I}_{k+n}}$. Then, we can map the key-points of $\mathcal{I}_{k+n}$ and $\mathcal{J}_{\mathcal{I}_{k+n}}$ by the homographies $T_2 H_{k,k+n}^{-1}$ and $T_1 S_k S_{k+n}^{-1}$, respectively, into the rectified images of the previous inter-sensorial registration and thus narrow the search window for putative correspondences needed for the fundamental matrix estimation.

2.3 Main applications

Similar to Section 3 of (Solbrig et. al., 2008), we discuss the applications of motion detection as well as object annotation in the triangulation-based geo-referencing of videos. While moving object can be detected from frame to frame of the video sequence, computing their exact position in the high-resolution image is not a trivial task. Geometrically, one can not obtain the depth of the moving object from one single image (since in the next image, it will already change its position), therefore model- and segmentation-based approaches will be needed to solve the problem. Still, the disparity information can be used in order to perform *motion analysis*, e. g. detect false alarms. On the other hand, annotation of objects of the high-resolution photo (for example: windows) in a video frame can be carried out similarly to (Solbrig et. al., 2008). We can assign the depths of these objects with respect to the video frame where inter-sensorial registration was performed for the last time and then use homographies $H_{k,k+n}$ to compute the coordinates of these objects in the current frame. In the case of moving sensor, $H_{k,k+n}$ are not always feasible, therefore segmentation methods or a computation of disparity maps between $\mathcal{I}_k$ and $\mathcal{I}_{k+n}$ could significantly refine the results.

3 RESULTS

We present a video sequence from a rotating and zooming camera recorded from a tower in Ettlingen, Germany. A high-resolution photo was made from approximately, although not exactly, the same position (Figure 2, top). The order of magnitude in the disparity range d_R in the rectified images is computed according to:

$$d_R \sim fcs \cdot \left(\frac{1}{z_{\min}} - \frac{1}{z_{\max}} \right), \tag{3}$$

where s is the scaling factor for rectification homographies (any value close to 1), $f \approx 1000\,\text{pixel}$ is the estimation of the focal length, $c \approx 1\,\text{m}$ is the distance between the cameras, $z_{\min}$ and $z_{\max}$ are furthest and closest point to the camera, respectively. According to the scene depicted in Figure 2, we can assume $z_{\min} = 50\,\text{m}$ and $z_{\max} = \infty$. Then, Equation (3) yields a disparity range of about 20 pixels. This makes the registration with homographies analytically impossible. This fact is also confirmed by the right picture in the middle row of Figure 2. Here we detected points at the frontal wall of the nearest house and matched them by a homography. But since the points – in the background or in the foreground, on the trees – are situated far away from the plane spanned by the frontal wall of the house, the mosaicking is not successful. After matching with our method, there are almost no significant deviations, and the small number of visible deviations is of local origin and is mainly due to the lack of accuracy of the triangular mesh obtained after the last step of guided matching as well as with the fact that transparency levels of objects (especially: trees) were not considered in this work.

The video sequence and the image of Figure 2 were recorded on the same day. The top of Figure 3 presents a photo taken several weeks *after* the video sequence was recorded. Note the clear differences in the vegetation, since the video sequence (on the left of the middle row of Figure 2) was recorded in summer while the high-resolution frame was taken in fall. Despite these radiometric differences, it was possible to register some 25 frames of the video sequence onto the image using our approach. For some interesting objects in the orthophoto (such as windows in the big church, denoted by blue crosses), we successfully obtained their positions in the frames using disparity information, as can be seen in the left picture of the middle row and in the bottom row of Figure 3. In order to visualize the quality of the registration, we compared the squared norms of the gradients:

$$\Delta = \left(\mathcal{J}_x^2 + \mathcal{J}_y^2\right) - \left(\hat{\mathcal{I}}_x^2 + \hat{\mathcal{I}}_y^2\right), \tag{4}$$

where $\cdot_x, \cdot_y$ are image gradients approximated by the Sobel-filter and $\hat{\mathcal{I}}$ is the frame of the sequence transformed by Equation (2) into the coordinate system of $\mathcal{J}$. By representing Δ from Equation (4), we visualize the distances between the corresponding characteristic edges in both images and we notice that they usually do not exceed an (acceptable) value of 5-10 pixels. The mosaic of 25 frames (total number of frames is 100 and every 4th frame was considered for registration) integrated into $\mathcal{J}$ is depicted in the bottom row of Figure 3.

4 CONCLUSIONS AND FUTURE WORK

We presented a fast and simple registration algorithm for referencing videos from scenes with a distinct 3D-structure. The spectrum of geometric configurations which can be handled is thus much wider than in the work (Solbrig et. al., 2008) where only 2D-2D situations were of interest. Our method allows reliable determining 3D coordinate of a pixel in a video frame (if the 3D-coordinates for pixels of the synthetic view are available and the triangulation approximates the surface well), even if no Euclidean reconstruction can be carried out from the video sequence. The results presented in this paper indicate that the triangular meshes created from enriched sets of correspondences provide, in many cases, an acceptable approximation of the surface. The computing time of the mesh does not depend on the disparity range and is less dependent on the image size than other state-of-the-art local and global algorithms for calculation of disparity maps since a lower point density not necessarily means worse results. In the first implementation of the algorithm, the calculation of SIFT-features in the video frame takes the bigger part of the computation time, computation times needed to calculate fundamental matrices, transformations of images, and enriched sets of correspondences are approximately the same, around 1 sec.

The main draw-back of the algorithm consists in the accuracy of and the outliers among the sampling points. The current implementation of the algorithm is usually not able to filter out wrong putative correspondences if they satisfy epipolar equation. Since the dense set of correspondences is spanned from triangles, the complete regions around these points will be given wrong disparities. It has been shown that using redundant information from more than two images (see, for instance, (Stewart and Dyer, 1988)) can significantly improve the performance; therefore we will concentrate our future efforts on integration of fused meshes into our triangulation networks. For example, one could calculate disparity maps from several frames of the video sequence to the image by the procedure described in Section 2 and then take the medians of depth values for overlapping regions. Another interesting aspect consists of taking into account the 3D-information from the synthetic views and calibrated cameras. To achieve this goal, the consideration of robust state-of-the-art methods for surface reconstruction beside image-based methods will be needed. As soon as the relation between a point in a video frame and the 3D-coordinate of the corresponding point in the synthetic view is established, new methods of quality control are made possible. Furthermore, our future work also includes a detailed consideration of applications such as change detection, motion analysis

and annotation of objects in the case of (geo-referenced) scenes with spatial depth.

REFERENCES

Besl, P. and McKay, N., 1992. A Method for Registration of 3-D Shapes, Trans. PAMI, 14, (2), pp. 239–256.

Bulatov, D., Wernerus, P., and Lang, S., 2009. A New Triangulation-based Method for Disparity Estimation in Image Sequences, to appear in Proc. of SCIA, Oslo, Norway

Hartley, R. and Zisserman, A., 2000. Multiple View Geometry in Computer Vision, Cambridge University Press.

Hirschmüller, H., 2005. Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, (2), pp. 807-814, San Diego, USA.

Lucas, B., Kanade, T., 1981. An iterative image registration technique with an application to stereo Vision. In: Proc. 7th International Joint Conference on Artificial Intelligence (IJCAI), pp. 674-679.

Krüger, W., 2001. Robust and Efficient Map-to-image Registration with Line Segments. In: Machine Vision and Applications, 13, (1), pp. 38–50.

Lin, Y., Yu Q., and Medioni, G., 2007. Map-Enhanced UAV Image Sequence Registration. In: IEEE Workshop on Applications of Computer Vision (WACV'07).

Loop, C. and Zhang, Z., 1999. Computing rectifying homographies for stereo vision. Technical Report MSR-TR-99-21, Microsoft Research.

Lowe, D. G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. In: International Journal of Computer Vision. 60, (2), pp. 91–110.

Martinec, D., Pajdla, T., Kostkova J., and Sara R., 2006. 3D reconstruction by gluing pair-wise Euclidean reconstructions, or "How to achieve a good reconstruction from bad images". In: Proc. of the 3D Data Processing, Visualization and Transmission conference (3DPVT), University of North Carolina, Chapel Hill, USA.

Matas, J. and Chum, O., 2004. Randomized Ransac with $T_{d,d}$-test. In: Image and Vision Computing, 22, (10), pp. 837–842.

Schenk, T. and Csathó, B., 2002. Fusion of LIDAR Data and Aerial Imagery for a More Complete Surface Description. In: International Archives of Photogrammetry and Remote Sensing, 3A pp. 310-318

Solbrig, P., Bulatov, D., Meidow, J., Wernerus, P., and Thönnessen, U., 2008. Online annotation of airborne surveillance and reconnaissance videos. In: Proc. 11th International Conference on Information Fusion.

Stewart, C. V. and Dyer, C. R., 1988. The trinocular general support algorithm: a three-camera stereo algorithm for overcoming binocular matching errors. In: Second International Conference on Computer Vision (ICCV), pp. 134-138

Xiao, J., Zhang, Y., and Shah, M., 2005., Adaptive Region-Based Video Registration, IEEE Workshop on Motion, Jan 5-6, Breckenridge, Colorado.

Figure 2: Top row: a high resolution image. Middle row left: a frame from the video sequence, right: Registration with one single homography fails. Bottom row: triangulated meshes from 4 key frames were successfully integrated into the image

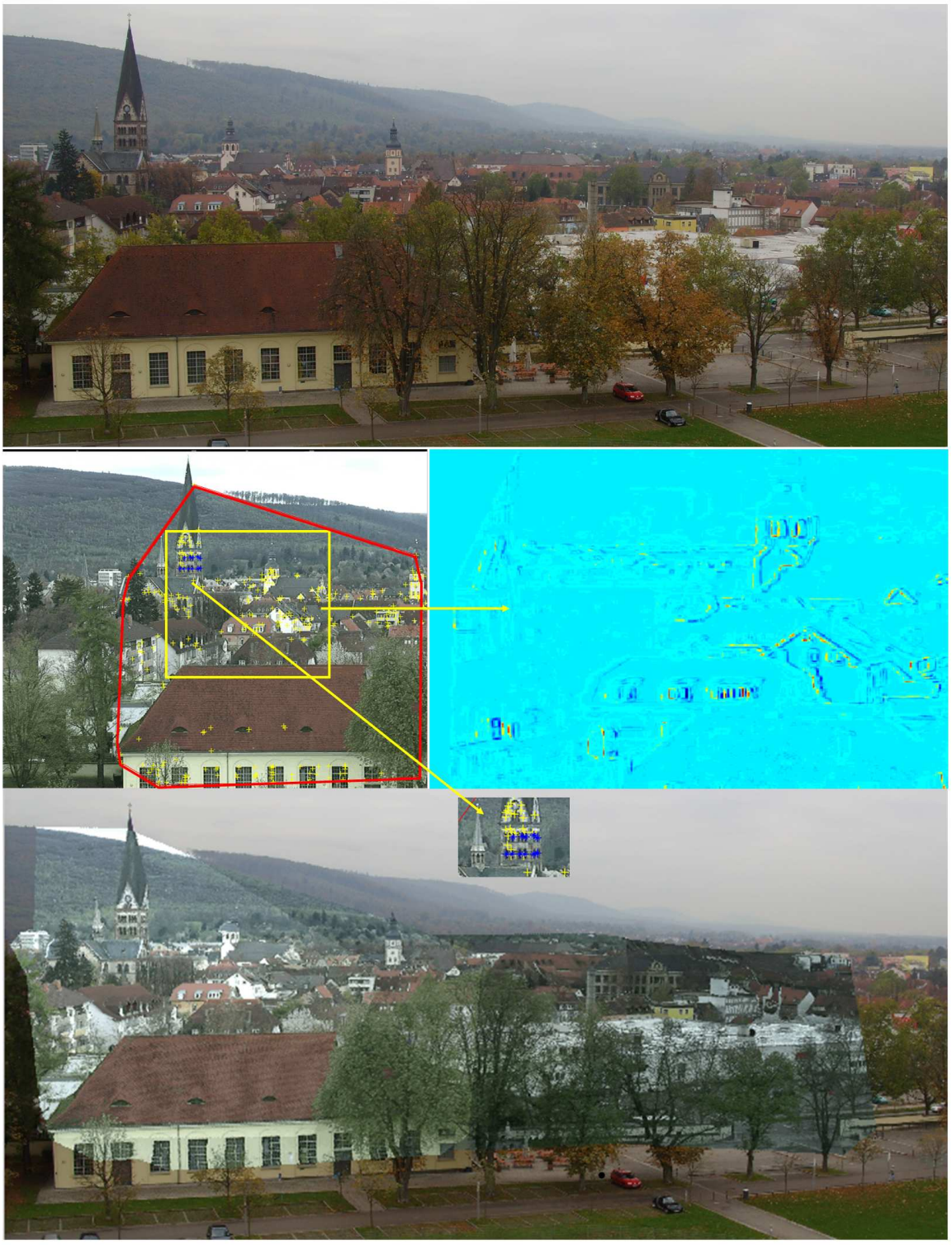

Figure 3: Top row: a high resolution image. Middle row left: a frame from the video sequence with convex hull of enriched correspondences to be integrated into the image. Right: quality control of registration by means of difference of squared norms of gradients for a small patch of frame; the distance between corresponding edges is negligibly low. Bottom row: triangulated meshes from 25 frames were successfully integrated into the image. Annotated objects (windows) are denoted by blue crosses on the left image of the middle row as well as the small fragment in the bottom row