Model-image registration of a building's facade based on dense semantic segmentation

Antoine Fond, Marie-Odile Berger, Gilles Simon

Model-image registration of a building's facade based on dense semantic segmentation

Antoine Fond[1], Marie-Odile Berger[2] and Gilles Simon[2]

Abstract— **This article presents an efficient approach for accurate registration of a building facade model "dressed" with dense semantic information. Localization sensors such as the GPS as well as vision-based methods are able to provide a camera pose in an efficient and stable way, but at the expense of low accuracy. We propose here to rely on semantic maps to improve the accuracy of a rough camera pose. Simultaneously we aim to iteratively improve the quality of the semantic map through the registration. Registration and semantic segmentation are jointly refined in an Expectation-Maximization framework. We especially introduce a Bayesian model that uses prior semantic segmentation as well as geometric structure of the facade reference modeled by Generalized Gaussian Mixtures. We show the advantages of our method in terms of robustness to clutter and change of illumination on urban images from various databases.**

Keywords: Image Registration, Semantic Segmentation, Camera Pose Estimation.

I. Introduction

Urban localization plays a major role in many applications including navigation aid [1], labeling of local touristic landmarks [2], [3], and robot localization [4]. The outdoor accuracy of mobile phone GPS is only 12.5 meters [5] and can be easily worse in urban areas where the street is flanked by buildings on both sides. Vision-based solutions are prone to be more accurate. However, a recent benchmark [6] shows that the long-term visual localization problem is far from solved.

Most of the image-based solutions rely on correspondences between features in the image and features from a city model build using Structure from Motion (SfM) algorithm [7], [8], [9], [10]. Though often successful and quite accurate, these methods can fail if the newly acquired image is taken in very different conditions compared to the SfM data. The main reason is that both hand-crafted [11] and learned [12] features are non-invariant e.g. to image blur, day-night and large viewpoint changes [13]. Furthermore, these features are based on local information, which leads to multiple hypotheses in the presence of similar or repeated patterns. In large-scale environments, disambiguation between all hypotheses can be difficult if not impossible, and computationally expensive.

By contrast, global (image-level) descriptors and particularly ConvNet features exhibit strong robustness against appearance changes induced by the time of day, seasons, or weather conditions [14] and are weakly impacted by the presence of similar or repeated patterns. However, those features only allow to approximate the pose of a query image by the one related to the closest image in a database, which is obviously not accurate unless a very dense view sampling is used [6].

An intermediate approach was proposed in [15]. In this approach, semi-global ConvNet features are used to match facade proposals generated in a query image with reference images of facades. However, the detected boundaries of a facade in the query image rarely fit exactly the reference boundaries, which usually yields coarse pose estimation.

Finally, some convolutional neural networks have been designed to regress the 6-DOF camera pose from a single RGB image in an end-to-end manner. The most emblematic of these is PoseNet [16]. However, the main drawback of the PoseNet approach is that it is inaccurate (see e.g. [17]) unless an excessively large and well sampled training set is used.

In short, it seems that solving urban localization is faced with the choice between accurate but unstable versus stable but inaccurate methods. Several suggestions have been formulated in [6] to increase the robustness of (local) feature-based methods, such as designing novel features, e.g., based on scene semantics [18], or using multiple images for pose estimation. However, whatever improvements made to feature-based methods, they will be confronted with the local nature of the features, which, in particular, does not allow the case of repeated patterns to be handled correctly. For instance, augmenting the features with pixel-wise semantic labels such as "windows", "facade" etc. would not help to distinguish between two features in the center of similar windows on the same or a different facade. Moreover, the complexity of the SfM models (millions of points for Dubrovnik [19] i.e. several Go of memory) can make the method unsuitable to run on mobile devices.

On the other hand, with city-scale data available (Google Street View/Maps), interests grew in 3D textured model as an alternative to SfM models. Their richer geometric information (i.e. facade planes) in combination with vanishing points [15] or coarse sensor pose prior [20], [21], [22] enables more reliable model-image registration in strong viewpoint changes. Furthermore, the texture information can be condensed into low dimension descriptors [15], semantic parts [21], edges only [22] or even not considered at all [20] leading to much lighter models.

Relying on a semantic segmentation to register a reference image of a facade in a target image [20], [21] has

[1]Antoine Fond is with Synthesia Technologies, London and with Université de Lorraine & Inria

[2]Marie-Odile Berger and Gilles Simon are with Inria and Université de Lorraine, France marie-odile.berger@inria.fr, gilles.simon@loria.fr

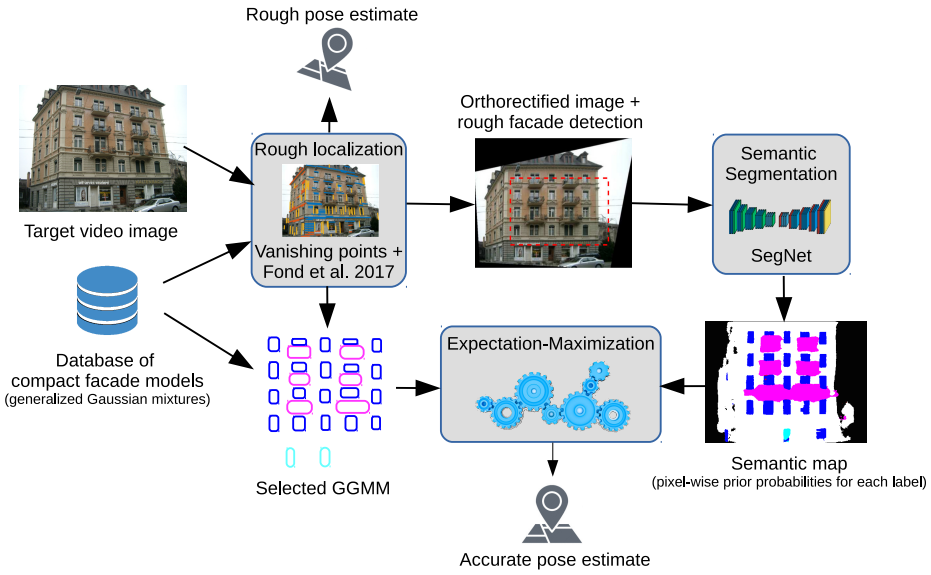

Fig. 1. Overview of the method. Starting from an initial estimate of the pose, a Generalized Gaussian Mixture Model of the identified facade is registered by Expectation-Maximization into the semantic map of the rectified image. All–not only the highest– pixel-wise classification scores are considered over iterations so that both pose and segmentation are refined during this process.

several advantages. First, there is no need for a complex similarity metric as semantic segmentation already manages appearance changes between the two images [23]. Second, the registration focuses on meaningful components on both images reducing possible local minima. In the methods proposed by [20] and [21], the highest classification score for each pixel is used for 3D-2D registration. Unfortunately, segmentation can include many misclassified pixels. For instance, a door can be easily misclassified as a window at the output of the neural network with a slightly higher score than the one assigned to the correct class.

Our method for model-image registration is summarized in Fig. 1 and is composed of three main blocks. In a first initializing step, a rough pose estimation is computed thanks to vanishing point detection, image rectification and rough facade detection. In a second step, a semantic segmentation is realized in the target rectified image. In a third step, registration is performed through an expectation-maximization (EM) algorithm based on the semantic labeling of the target and the reference image. This algorithm is the major contribution of our approach. The main idea is to consider all (and not only the highest) pixel-wise classification scores as segmentation priors in a Bayesian framework. The EM algorithm is used to iteratively compute the pose that best assigns the image pixels to their related structural elements (window, door, ...) extracted from a ground-truth, orthorectified semantic map. The final assignments can be seen as a posterior segmentation of the target image. Thus, if semantic segmentation makes it possible to guide the registration, registration in turn makes it possible to guide the semantic segmentation (Fig. 2). This idea is similar, in spirit, to older works estimating simultaneoulsy the pose and point correspondences in an iterative process [24], [25], [26]. Of course, as we assume the ground-truth semantic segmentation of the reference to be known, the main goal of our method is registration. The online improvement of the target segmentation through the joint approach is rather a means to achieve that goal.

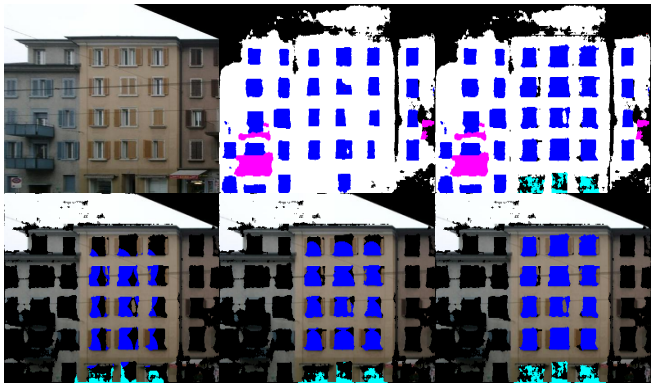

Fig. 2. Evolution of the semantic segmentation during the EM on the first 3 iterations. Top: (from left to right) the target image I with the orange building as reference, the prior semantic segmentation $P(l_j|i,I)$, and the posterior semantic segmentation after registration. Bottom: The doors on the ground-floor wrongly classified in the prior are progressively correctly classified as well as they are guiding the registration.

Furthermore, operating in a Bayesian framework makes it possible to use a very compact model, namely the parameters of a generalized Gaussian mixture model (GGMM), with one generalized Gaussian per structural element of the facade. The weights of the GGMM being allowed to vary during the iterations of the EM algorithm, our method is robust to facade occlusions. Last but not least, because a GGMM has an infinite support, our method is also robust to poorly initialized localization.

Section II provides a state of the art of 3D-2D registration methods based on a camera pose prior. The initialization steps of our method are described in section III.

The proposed bayesian model and expectation-maximization algorithm are presented in section IV. Finally, extensive experimental results are provided in section V.

II. Related work

In addition to extensive benchmarks presented in [6], a recent survey on visual-based localization can be found in [27]. This section focuses on 3D-2D registration methods based on a prior camera pose and a 2.5D or 3D model of a building or a facade, "dressed" with texture and/or semantic information. The pose prior is usually obtained through localization sensors (e.g. the GPS and a magnetic compass), though any robust, even not accurate visual-based method such as those presented above may be used. In these methods, the pose is iteratively refined starting from the prior, so that the projection of the model gets accurately aligned with its image counterpart. The different approaches differ essentially with respect to which information is used to measure this alignment.

A. Feature-based methods

Feature-based methods rely on textured 3D models of buildings [22], [28] or reference images of facades in frontal view [29]. In the case of a building, the 3D model is rendered using the prior pose. In the case of a facade, the target video image is orthorectified based on the prior pose. In both cases we obtain two close images between which feature matching is facilitated. Features can be edgels [22] or points [29], [28] and the matching procedure is based on 1D [22] or 2D [29], [28] cross-correlation, or comparison between descriptors [28]. Unfortunately, edge detection is sensitive to illumination, shadows and occlusions, as well as scene-to-camera distances. On the other hand, [28] has shown that fast-to-compute corner features such as Harris [30] or FAST [31] are weakly repeatable between a rendered frame and a video frame, due to depth blur changes. Unless an adaptive depth blur is applied to the synthetic image, which requires a tedious calibration of the camera response curve, feature points such as SIFT [11] or SURF [32], both based on approximations of Laplacian of Gaussian, prove more repeatable. However, as already mentioned, SIFT features have several drawbacks, which is confirmed by our experiments (see section V).

B. Template-based methods

Template based-methods usually aim at tracking a planar patch between two image I_0 and I_i. More exactly, homographies $\mathbf{H}_0^i$ are computed so that $I_0 \circ \mathbf{H}_0^i$ is "similar" to I_i inside the patch. The similarity measure is very important with these methods.

The first measure used was the sum of the squares of the differences between pixel gray levels (L_2-norm) [33]. Optimization is done very quickly by gradient descent using the Gauss-Newton algorithm. While the transformation was initially limited to a simple 2D translation, the geometric models were then enriched to cover affine transformations [34] and homographies [35]. The computational efficiency of the minimization has been improved in [36] by a second-order approximation without calculating the Hessian. The similarity measure based on the L_2-norm remains sensitive to changes in illumination and occlusions. In [37], the reference image is decomposed into a pyramid of sub-images that are registered independently according to the L_2-norm. The global solution of the registration is searched recursively in the parameter space in such a way that it maximizes the number of sub-registrations. If decomposition makes it possible to effectively treat occlusions, it can be sensitive to frequent repetitions on facades.

Kim et al. [38] use a M-estimator for a more robust similarity measure. Mutual information between images is also a measure of similarity that is less sensitive to changes in illumination and occlusions [39] and has long been used for multimodal registration in medical imaging [40]. While these measures significantly increase the complexity of optimization, progress has since been made that allows for effective resolution [41]. Nevertheless, the convergence of all these methods depends strongly on the accuracy of the initialization.

In the case of translational shift only, the global solution can quickly be found by phase shift in the frequency domain. This method can be generalized to similarities [42] and homographies [43]. However, areas of the current image that do not correspond to the reference image can disturb the Fourier transform and cause the method to fail. This happens regularly in urban images where a building can be observed at very different scales.

C. Semantic-based methods

Semantic segmentation has been used in at least two previous works. In [20], an initial pose provided by a GPS is refined by fitting a coarse 2.5D (building's footprints and height) city model to the image. The rotation is computed from vanishing points and translational hypothesis are generated by matching vertical ridges of the model with vertical lines detected in the image. A semantic segmentation using a SVM classifier on local image descriptors allows pixels that belong to *facades* and pixels from the background to be distinguished. The log-likelihood between that probability of classification and the projected facades of the model is then maximized over all the pose hypotheses. Though this method is interesting, the accuracy of the registration relies on the pixel-wise segmentation, which is noisy and does not separate adjacent facades. Moreover, structural elements on the facades (windows, doors, etc.) are not detected by the classifier (they are simply classified as *facade*), though these elements would be useful to get a more accurate registration.

Chu et al. [21] exploit this structural information to better estimate the camera location as well as some geometric parameters of the building's model (height of each floor, vertical positions of windows and doors, etc.). As in [20], the method assumes the camera pose to be initialized by GPS and requires geo-referenced footprint of buildings as a base for creating the 3D models. The problem is formulated as inference in a Markov random field, which encourages

the projection of the 3D model to match the image edges, semantics (based on SegNet [44]) and location of doors and windows (based on Edgeboxes [45] and AlexNet [46]) and to differ from the background in all GoogleStreetView images around the building. Nevertheless the complexity of the inference that uses a discretized parameters search space and multiple views are disadvantages for real time applications to urban localization.

III. Initialization

Initialization of the Expectation-Maximization procedure is based on four steps (see Fig. 3): (i) vanishing points as well as the camera intrinsic parameters are computed from the image content. (ii) the image is rectified so that the facades of the buildings appear as if they where fronto-parallel to the camera (several rectified images can be obtained), (iii) facades in the rectified images are detected (approximate bounding boxes are obtained) and recognized among facades of the model, (iv) semantic segmentation and registration are initialized from the bounding boxes of the recognized facades. In the following, this initialization step will be referred to as $t = t_0$. We now detail each of its subtasks.

A. Autocalibration and plane rectification

Steps (i) and (ii) of the initialization process are performed using the method described in [47]. Horizontal vanishing points of the image are detected by exploiting accumulations of oriented segments around the horizon line. The principal point is assumed to be at the center of the image and the focal length is computed from a detected pair of orthogonal vanishing points. For each detected vanishing point a homography is computed, that transforms all vertical planes in the direction of the vanishing point to a fronto-parallel view of the planes.

B. Facade detection and recognition

Facades are detected and recognized in the rectified images using the method presented in [15]. This method relies on image cues that measure facade characteristics such as shape, color, contours, semantic structure (windows and balconies are detected using SegNet [44]) and symmetry. These cues are combined to generate a few facade candidates quickly. The candidates are then classified into "facade" and "non facade" through a neural network using SPP descriptors [48]. The remaining facades are matched with the facade database using a metric learned through a siamese neural network [49] taking the SPP descriptors as inputs.

C. Registration and segmentation initialization

In this method we aim to jointly solve the registration of the recognized reference to the detected facade in the target image and the segmentation of the latter into semantic parts. As the image has been previously rectified using calibrated camera intrinsics the only remaining parameters to register the reference image onto the target image are one scale parameter s (the aspect ratio is preserved) and two translational parameters (t_x, t_y). Facade recognition enables

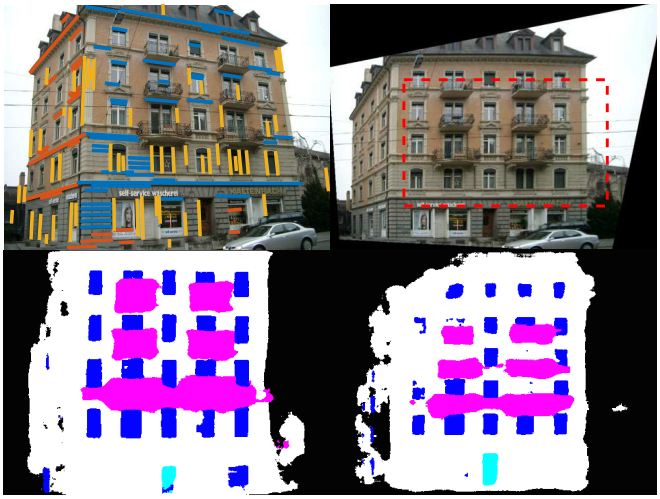

Fig. 3. Initialization steps. (i) Vanishing points are detected using the method described in [47] (top-left, vanishing points are shown by their supporting line segments, with one color per vanishing point). (ii) The target image is rectified so that the facades of the buildings appear as if they where fronto-parallel to the camera (top-right, here the yellow and blue vanishing points were used). (iii) A facade is detected (red dashed rectangle at top-right) and recognized by using the method presented in [15] (the selected reference facade is shown in Fig. 6, right). (iv) The semantic segmentation is computed inside the enlarged detected region (bottom-left). For sake of comparison, the semantic segmentation obtained in the whole rectified image is shown at bottom-right.

to select the correct facade reference to be registered in a larger facades database. Moreover thanks to facade detection we can estimate a first initialization of the registration parameters by solving the least-square problem that maps the four transformed corners of the reference to the four corners of the detection.

As the facade detection [15] step relies on semantic segmentation, it also provides a first initialization of the latter. However the SegNet [44] inference is sensitive to scale (Fig. 3, bottom row). To improve the initial segmentation we zoom into the target image and we perform another inference. Rather than restricting I to the detected facade which would mean to put too much confidence on the detection, we restrict I to an enlarged detection scaled by a constant value of 40 %.

IV. Joint registration and semantic segmentation

A. Bayesian model

We wish to register the recognized image reference I_{ref} onto the target image I in which the facade has been detected through a transformation T and simultaneously improve the quality of the semantic segmentation in the target image.

We denote $L = \{l_j\}_{1 \leq j \leq K}$ the different labels from the semantic segmentation that are characteristic of a facade architecture such as "window", "door" and "balcony". The target image is considered as sets of 2D labeled points. Let $X = \{X_i\}_{1 \leq i \leq N}$ be a set of N data points $X_i = (x_i, y_i)$ from the target image I. These points are the coordinates of the pixels i from the target image I that have a fair probability

of being one of the labels $P(l_j|i,I) \geq 0.01$ (Fig. 4). This probability $P(l_j|i,I)$ is the score of the last layer of the CNN for semantic segmentation at X_i.

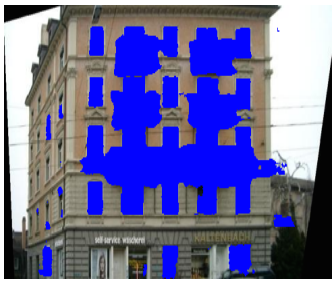

Fig. 4. Data points X from the target image I. Only the points which are likely ($P(l_j|i,I) \geq 0.01$) to be a characteristic facade architecture components are considered.

The ground-truth semantic segmentation corresponding to the image reference I_{ref} is assumed known (Fig. 6, left) and is modeled in a compact way as follows. For each label l_j, we extract the connected components of the reference segmentation, and a generalized Gaussian $\mathcal{N}_p$ with shape parameter p is fitted to each of them. Generalized Gaussians [50] are well suited for facade architectural components as the L_p norm $\|M\|_{p,\Sigma}^p = \frac{m_x^p}{\Sigma_{xx}} + \frac{m_y^p}{\Sigma_{yy}}$ unit ball is roughly rectangular with a high value of p (Fig. 5). To properly model the rectangular shape of facade components and keep the computation tractable we choose $p = 4$ (Fig. 6, right).

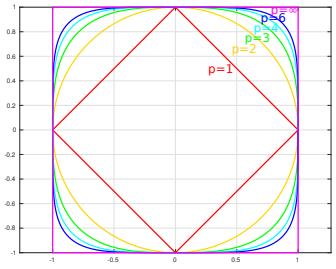

Fig. 5. L_p norm unit ball for different values of p.

As the image is rectified and the shape of the connected components is typically rectangular, the axes of the generalized Gaussians are aligned with the image axes. Let m_j be the number of the connected component labelled with l_j. The centers of the generalized Gaussians $\left(\mu_{k_j}\right)_{1\leq k_j \leq m_j}$ are initialized to the mean of the pixels coordinates of the connected components and the covariances $\Sigma_{k_j} = \text{diag}\left(\sigma_x^{p/2}, \sigma_y^{p/2}\right)$ are initialized from their vertical and horizontal variances (respectively σ_x and σ_y). They are then refined by minimizing the error between the connected component and the true generalized Gaussian form using Gauss-Newton.

Finally, the ground-truth semantic segmentation is modeled by a mixture of generalized Gaussian distributions for each label l_j: $\left(\pi_{k_j}, \mu_{k_j}, \Sigma_{k_j}\right)_{1\leq k_j \leq m_j}$. The mixture priors $\left(\pi_{k_j}\right)_{1\leq j\leq K, 1\leq k_j \leq m_j}$ are initialized such as π_{k_j} is the ratio of the number of points from the connected component k_j over the total number of points from the image reference. Then they are normalized so that $\sum_{j,k_j} \pi_{k_j} = 1$.

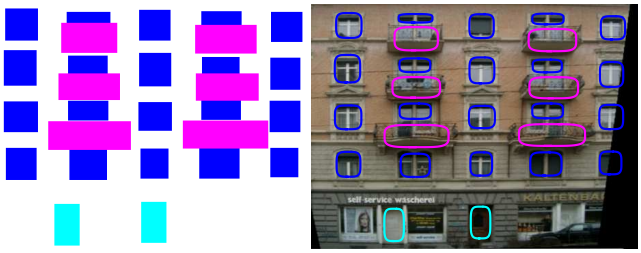

Fig. 6. Ground-truth of the semantic segmentation from the reference image I_{ref} (left) and the associated generalized Gaussian mixtures (right).

The goal is to estimate the geometric transformation $T(\Theta)$ of parameters $\Theta = (t_x, t_y, s)$ that registers these generalized Gaussians to the set of observed data points X from the target image I. In addition, the assignment of a data point X_i to a transformed generalized Gaussian can be seen as a posterior segmentation. Assuming that the observed data X are independent and taking the logarithm, the *a posteriori* distribution is maximized to find Θ :

$$\begin{aligned}\Theta^\star &= \arg\max_\Theta \ln P(X|\Theta,I)P(\Theta) \\ &= \arg\max_\Theta \sum_{i=1}^{N} \ln P(X_i|\Theta,I) + \ln P(\Theta),\end{aligned} \tag{1}$$

with N the number of data points. Using the law of total probability, $P(X_i|\Theta,I)$ can be expressed as follows:

$$\begin{aligned}P(X_i|\Theta,I) = &\sum_{j=1}^{K} P(X_i|l_j,\Theta,I)P(l_j|i,\Theta,I) \\ &+ P(X_i|o,\Theta,I)P(o|i,\Theta,I),\end{aligned} \tag{2}$$

where K is the number of labels and:

- $P(X_i|l_j,\Theta,I)$ is the likelihood of an observation X_i given its assignment to label j through the transformation T, and is modeled by a mixture of transformed generalized Gaussians:

$$\begin{aligned}P(X_i|l_j,\Theta,I) &= \sum_{k_j=1}^{m_j} \pi_{k_j} \mathcal{N}_p\left(X_i|T\mu_{k_j}, s^p\Sigma_{k_j}\right) \\ &= \sum_{k_j=1}^{m_j} \pi_{k_j} \frac{\exp\left(-\left\|X_i - T\mu_{k_j}\right\|_{p,s^p\Sigma_{k_j}}^p\right)}{4/p^2\Gamma(1/p)^2|s^p\Sigma_{k_j}|},\end{aligned} \tag{3}$$

- $P(l_j|i,\Theta,I)$ is the segmentation prior probability. Thanks to the scale reestimation and the invariance of CNN to small translations, the semantic segmentation inference is pretty stable. Thus we can assume that $P(l_j|i,\Theta,I) = P(l_j|i,\Theta^{(t_0)},I)$;
- the likelihood of an observation given its assignment to the outlier class $P(X_i|o,\Theta,I) = P(X_i|o,I)$ is modeled as

a uniform distribution $P(X_i|o,I) = \frac{1}{HW}$, with H,W the dimensions of the target image;
- $\nu = P(o|i,\Theta,I)$ is the outliers rate.

To be more robust to clutter we let the mixture weights free to vary during the inference but, as a tradeoff, we assume a prior distribution over them. We can actually add the mixture weights to $\Theta = \left(t_x, t_y, s, \{\pi_{k_j}\}_{1\leq j\leq K, 1\leq k_j\leq m_j}, \alpha\right)$ without changing Eq. 1. We don't assume any prior for the transformation parameters (t_x, t_y, s) but we choose a Dirichlet distribution as a prior for the mixture weights π_{k_j}:

$$P(\Theta) = \mathscr{D}ir\left(\pi_{k_j}|\alpha_{k_j}\right)_{1\leq j\leq K, 1\leq k_j\leq m_j} \propto \prod_{j,k_j} \pi_{k_j}^{\alpha_{k_j}-1} \tag{4}$$

Gauvain et al. [51] show that Dirichlet distribution is a practical prior candidate for mixture distributions that enables closed-form optimizations in the EM framework. $\left(\alpha_{k_j}\right)_{1\leq j\leq K, 1\leq k_j\leq m_j}$ are set to the same values as the initialized mixture priors $\alpha_{k_j} = \pi_{k_j}^{(t_0)}$.

B. Expectation-Maximization

This Maximum *A Posteriori* (MAP) problem can be solved in the framework of expectation-maximization. We define the latent variables $Z = \left\{z_{i,j,k_j} \in \{0,1\}, z_{i,o} \in \{0,1\}\right\}_{1\leq i\leq N, 1\leq j\leq K, 1\leq k_j\leq m_j}$ such that $z_{i,j,k_j} = 1$ means that X_i is assigned to a generalized Gaussian $\left(T\mu_{k_j}, s^p\Sigma_{k_j}\right)$ from the label l_j and $z_{i,o} = 1$ means that X_i is assigned to the outlier extra class o. The Expectation-Maximization algorithm seeks to find the solution iteratively by alternating between calculating the expected complete-data log-likelihood $Q(\Theta|\Theta^{(t)})$ with respect to Z given X and the current parameters $\Theta^{(t)}$ and finding the parameters Θ that maximizes this quantity :

$$\begin{aligned} Q\left(\Theta|\Theta^{(t)}\right) &= \mathbb{E}_{Z|X,\Theta^{(t)}} \ln P(X,Z|\Theta) \\ &= \sum_Z P(Z|X,\Theta^{(t)}) \ln P(X,Z|\Theta) \\ &= \sum_{i,j}\sum_{k_j} \beta_{i,j,k_j}\left(\ln \pi_{k_j} + \ln P(l_j|i,I)\right) \\ &+ \sum_{i,j}\sum_{k_j} \beta_{i,j,k_j} \ln \mathscr{N}_p\left(X_i|T\mu_{k_j}, s^p\Sigma_{k_j}\right) \\ &+ \sum_i \gamma_i \ln \frac{\nu}{HW} \end{aligned} \tag{5}$$

with $\beta_{i,j,k_j} = \mathbb{E}\left(z_{i,j,k_j}|X,\Theta^{(t)}\right)$ and $\gamma_i = \mathbb{E}\left(z_{i,o}|X,\Theta^{(t)}\right)$

Thus the Expectation-Maximization framework iterates between the two steps :

- **E-Step:** compute β_{i,j,k_j} and γ_i
- **M-Step:** $\Theta^{(t+1)} = \arg\max_\Theta Q\left(\Theta|\Theta^{(t)}\right) + \ln P(\Theta)$

The **E-Step** can be seen as the computation of an assignment probability of each data point X_i to a generalized Gaussian $\left(T\mu_{k_j}, s^p\Sigma_{k_j}\right)$ from the label l_j knowing the current parameters $\Theta^{(t)} = \left(\{\pi_{k_j}\}^{(t)}_{1\leq j\leq K, 1\leq k_j\leq m_j}, \alpha^{(t)}, t_x^{(t)}, t_y^{(t)}, s^{(t)}\right)$. Using Bayes rule and by denoting $\lambda = \frac{\nu}{HW}$, we can write :

$$\begin{aligned} \beta_{i,j,k_j} &= \mathbb{E}\left(z_{i,j,k_j}|X,\Theta^{(t)}\right) \\ &= \frac{\pi_{k_j}\mathscr{N}_p\left(X_i|T\mu_{k_j}, s^p\Sigma_{k_j}\right)P(l_j|i,I)}{\sum_{j',k'_{j'}} \pi_{k'_{j'}}\mathscr{N}_p\left(X_i|T\mu_{k'_{j'}}, s^p\Sigma_{k'_{j'}}\right)P(l_{j'}|i,I) + \lambda} \end{aligned} \tag{6}$$

$$\begin{aligned} \gamma_i &= \mathbb{E}\left(z_{i,o}|X,\Theta^{(t)}\right) \\ &= \frac{\lambda}{\sum_{j',k'_{j'}} \pi_{k'_{j'}}\mathscr{N}_p\left(X_i|T\mu_{k'_{j'}}, s^p\Sigma_{k'_{j'}}\right)P(l_{j'}|i,I) + \lambda} \end{aligned} \tag{7}$$

In the **M-Step** we aim to maximize $R = Q\left(\Theta|\Theta^{(t)}\right) + \ln P(\Theta)$ knowing the assignments $\beta_{i,j,k}$ and γ_i. By replacing the expressions of the distribution from equations 3 and 4 and by ignoring the constant terms, R can be re-written as $\tilde{R}$:

$$\begin{aligned} \tilde{R} = &-\sum_{i,j,k_j} \frac{\beta_{i,j,k_j}}{2}\left(\ln|s^p\Sigma_{j,k_j}| + \left\|X_i - T\mu_{k_j}\right\|^p_{p,s^p\Sigma_{j,k_j}}\right) \\ &+ \sum_{i,j,k_j} \beta_{i,j,k_j}\ln\pi_{k_j} + \sum_i \gamma_i \ln\lambda + \sum_{j,k_j}\left(\alpha_{k_j} - 1\right)\ln\pi_{k_j} \end{aligned} \tag{8}$$

The form of $\tilde{R}$ permits independent maximization of each of the following parameters sets (t_x, t_y, s), $\{\pi_{k_j}\}_{1\leq j\leq K, 1\leq k_j\leq m_j}$ and α. From the partial derivatives $\frac{\partial\tilde{R}}{\partial t_x} = \frac{\partial\tilde{R}}{\partial t_y} = \frac{\partial\tilde{R}}{\partial s} = 0$ we can derive a polynomial system which cannot be solved in closed-form for $p = 4$. Our solving strategy is similar to the one we used in the initialization of the mixture from the reference. First, we solve the polynomial system in closed-form with $p = 2$, setting the partial derivatives of $\tilde{R}$ to zero. This leads to solving a polynomial system of one quadratic equation in s and two linear equations in t_x and t_y. The closed-form solution is the following:

$$\begin{cases} s = \frac{-2a_2a_7a_8 + a_3a_5a_8 + a_4a_6a_7 \pm \sqrt{\Delta}}{4a_9a_7a_8} \\ t_x = \frac{-a_3 - 2a_5 s}{2a_7} \\ t_y = \frac{-a_4 - 2a_6 s}{2a_8} \end{cases}$$

with

$$\Delta = -16a_1a_7^2a_8^2a_9 + 4a_2^2a_7^2a_8^2 - 4a_2a_3a_5a_7a_8^2 - 4a_2a_4a_6a_7^2a_8 + a_3^2a_5^2a_8^2 + 4a_3^2a_7a_8^2a_9 + 2a_3a_4a_5a_6a_7a_8 + a_4^2a_6^2a_7^2 + 4a_4^2a_7^2a_8a_9$$

$$a_1 = -\sum_{i,j,k_j} \beta_{i,j,k_j} \left(\frac{x_i^2}{\sigma_{k_j,x}} + \frac{y_i^2}{\sigma_{k_j,y}} \right)$$

$$a_2 = \sum_{i,j,k_j} \beta_{i,j,k_j} \left(\frac{x_i \mu_{k_j,x}}{\sigma_{k_j,x}} + \frac{y_i \mu_{k_j,y}}{\sigma_{k_j,y}} \right)$$

$$a_3 = 2\sum_{i,j,k_j} \beta_{i,j,k_j} \frac{x_i}{\sigma_{k_j,x}}$$

$$a_4 = 2\sum_{i,j,k_j} \beta_{i,j,k_j} \frac{y_i}{\sigma_{k_j,y}}$$

$$a_5 = -\sum_{i,j,k_j} \beta_{i,j,k_j} \frac{\mu_{k_j,x}}{\sigma_{k_j,x}}$$

$$a_6 = -\sum_{i,j,k_j} \beta_{i,j,k_j} \frac{\mu_{k_j,y}}{\sigma_{k_j,y}}$$

$$a_7 = -\sum_{i,j,k_j} \frac{\beta_{i,j,k_j}}{\sigma_{k_j,x}}$$

$$a_8 = -\sum_{i,j,k_j} \frac{\beta_{i,j,k_j}}{\sigma_{k_j,y}}$$

$$a_9 = 2\sum_{i,j,k_j} \beta_{i,j,k_j}$$

Then, we refine the result by minimizing $J = \frac{\partial \tilde{R}}{\partial t_x}^2 + \frac{\partial \tilde{R}}{\partial t_y}^2 + \frac{\partial \tilde{R}}{\partial s}^2$ for $p = 4$ using gradient descent. As J is polynomial both the gradient and the hessian can be computed using their polynomial expression in the Gauss-Newton algorithm. The convergence is reached after a few iterations and we can update the transformation parameters $\left(t_x^{(t+1)}, t_y^{(t+1)}, s^{(t+1)}\right)$. The update for the mixtures weights π_{k_j} and the outliers rate α follows the formula from [51]:

$$\pi_{k_j}^{(t+1)} = \frac{\sum_i \beta_{i,j,k_j} + \alpha_{k_j} - 1}{\sum_{i,k'_{j'}} \beta_{i,j,k_{j'}} + \sum_{k'_{j'}} \left(\alpha_{k'_{j'}} - 1\right)} \tag{9}$$

$$\alpha^{(t+1)} = \frac{\sum_i \gamma_i}{\sum_{i,k'_{j'}} \beta_{i,j,k_{j'}} + \sum_{k'_{j'}} \left(\alpha_{k'_{j'}} - 1\right)} \tag{10}$$

V. Results

A. Implementation and efficiency

Unlike most EM approaches, in our method the generalized Gaussian parameters are fixed except for the mixture prior weights. Indeed here the generalized Gaussians model the semantic components of the reference facade. This compact representation of a facade enables our method to be efficient. The number of generalized Gaussians is in the order of the number of windows (typically between 2 and 30). If we assume that the image is full of adjacent facades and the empty space between windows is as large as the window itself we can approximate the number of data points $N \approx 0.25HW$. In our testing data, this approximation is valid with an average $\hat{N} = 31000$. Actually registration does not request the points to be sampled at each pixel. In our implementation we use a multi-resolution scheme with 2 levels. The EM algorithm is executed on a downsampled version of the set of points X until convergence $\left\|\Theta^{(t+1)}(1:2) - \Theta^{(t)}(1:2)\right\| \leq \varepsilon_t$ and $\left\|\Theta^{(t+1)}(3) - \Theta^{(t)}(3)\right\| \leq \varepsilon_s$ and then executed again on the full set X from the last estimated $\Theta^{(t)}$.

The complexity for one iteration t of the EM algorithm is $O(NK \max_j m_j)$ and parallelization is easy for the E-Step as β_{i,j,k_j} computations are independent. This efficient complexity is also a consequence of the partial solvability of the M-Step in closed-form with negligible Gauss-Newton inner-iterations. The code of our implementation is in Matlab with the EM in C. The average computation time for one iteration t is 0.023 second on an I7-3520M CPU. The number of steps for the EM to converge strongly depends on the initialization. In our testing data, only 6 iterations are needed to converge for the downsampled level and 2 more for the upper level (Fig. 7). Our M-Step optimization scheme is also faster and more accurate on this problem than homotopy continuation methods. Thus the average computation time of the whole EM is 0.121 second.

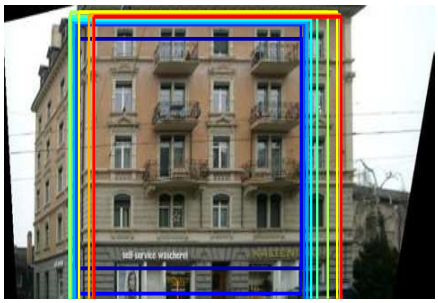

Fig. 7. The registered reference boundaries of the image reference for each iteration of the algorithm are drawn in color according to the jet colormap. From dark blue for the initial iteration to red for the final one.

To avoid the problem of the EM converging to a local maximum, we use several initializations in practice. We apply our method not only to the detected facade but also to the top-20 facade proposals [15] that overlap the detected facade. The final solution is the one with the highest R values.

B. Validation with ground-truth semantic references

We test our method on 3 different datasets. The first one is VarCity 3D [1]. It consists of 401 street-view images of buildings along the same street. Images are also semantically labelled and a SfM reconstruction of the scene is available as well as the camera parameters. The image viewpoints are roughly fronto-parallel and facades cover most of the image (see e.g. Fig. 9). Therefore the change of scale from the reference is minor but the translation value can be high with large image parts not visible.

The second one is the first 100 buildings from Zurich Buildings Database (ZuBuD) with 5 different viewpoints per building. Among those scenes we keep only the ones that have been correctly reconstructed by SfM [2]. The diversity

[1] https://varcity.ethz.ch/3dchallenge

[2] http://ccwu.me/vsfm

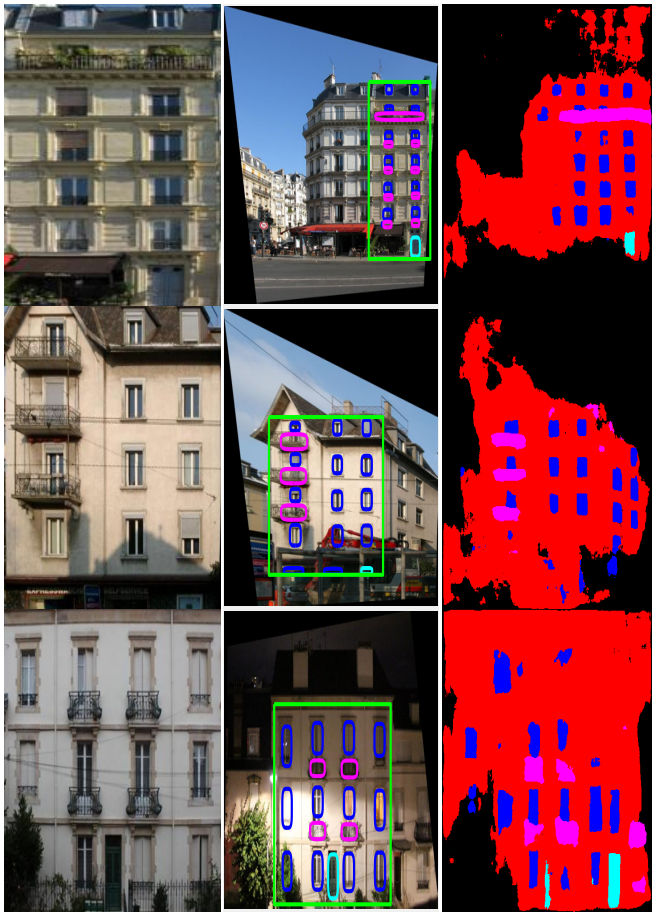

Fig. 8. Examples of registration results on VarCity 3D (first line), ZuBuD (second line) and NancyLights (third line). Left: reference facade; middle: registration result; right: *A posteriori* segmentation.

of viewpoints in this dataset enables a wider range of scale as well as occlusions.

The last dataset NancyLights[3]aims to show the robustness of the proposed method to change in illumination. It consists of 2 time-lapses of the same facade taken from the same viewpoint at sunrise and sunset for a total of 56 images.

For each building in all 3 databases we select the facade reference from the most fronto-parallel viewpoint where the facade is fully visible with the least occlusions possible. The reference is manually segmented into the 3 semantic labels "window", "door" and "balcony" (Fig. 6). The ground truth boundaries of the reference are transferred to all the images where this facade is visible using the geometric information from the SfM model.Examples of registration results are shown in Fig. 8. More examples can be found in the supplementary material.

We compare our method to both template-based and feature-based registration between the rectified target image and the reference image. In the first category we are competing against raw detection [15], L_2 norm minimization between images by gradient descent [36], Mutual Information maximization [52], [53], and phase correlation [42]. For the optimization methods the same initializations as for our method are chosen. For the feature-based method we extract SIFT descriptors in the rectified image with fixed orientation. 2 pairs of matched SIFT descriptors using Lowe's criteria [54] are used to generate transformation samples in a RANSAC framework. The comparison is done in the image itself computing the cumulative normalized histogram of the error in translation and scale. For ZuBuD and VarCity 3D the SfM models enable us to also show the error on the camera pose translation deducted from the registration (Table I and Fig. 9).

[3]This dataset is freely available on our team website https://magrit.loria.fr/dataset.html

	SIFT	PhCorr	LstSqr	MutInf	Ours
VarCity 3D	0.04	0.02	0.37	0.35	0.03
ZuBuD	0.22	0.67	0.33	0.44	0.12

TABLE I

MEDIAN ERRORS FOR THE 3D CAMERA TRANSLATION (RELATIVE TO THE FACADE DISTANCE)

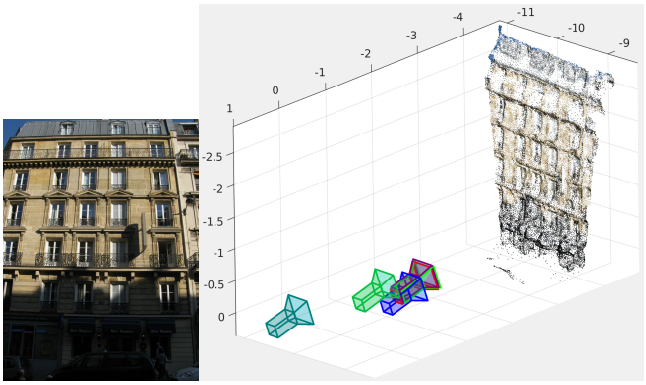

Fig. 9. An example image from VarCity 3D (left) and the related SfM model (right). The ground-truth pose is shown in light green. Poses obtained by using SIFT+RANSAC, phase correlation, least squares, mutual information and our method are shown in blue, red, cyan, dark green and purple, respectively. Our result is superimposed on the ground truth.

The good results on VarCity 3D (Fig. 10) show that our method can handle large translations thanks to the infinite generalized Gaussian support. Even when this phenomenon concurs with very repetitive patterns, the multiple initializations that exploit those repetitions and symmetries as well as the MAP regularization globally provide a correct registration. On the contrary these conditions are a major weakness for template-based methods that get easily stuck in a local minima (Fig. 11, top). Still, in our method, the lack of discriminative architectural components like doors can cause the same shift in registration aligning the wrong floor or windows when SIFT can handle it using other features.

On the other hand, our approach gives the best results on this dataset and benefits from a decent initial detection (Fig. 10, black, middle). Occlusions are another consequence of the diversity in viewpoints. Updating the mixture weights during the EM enables our method to be robust to them (Fig. 11, bottom) as well as hidden parts (Fig. 11, middle)

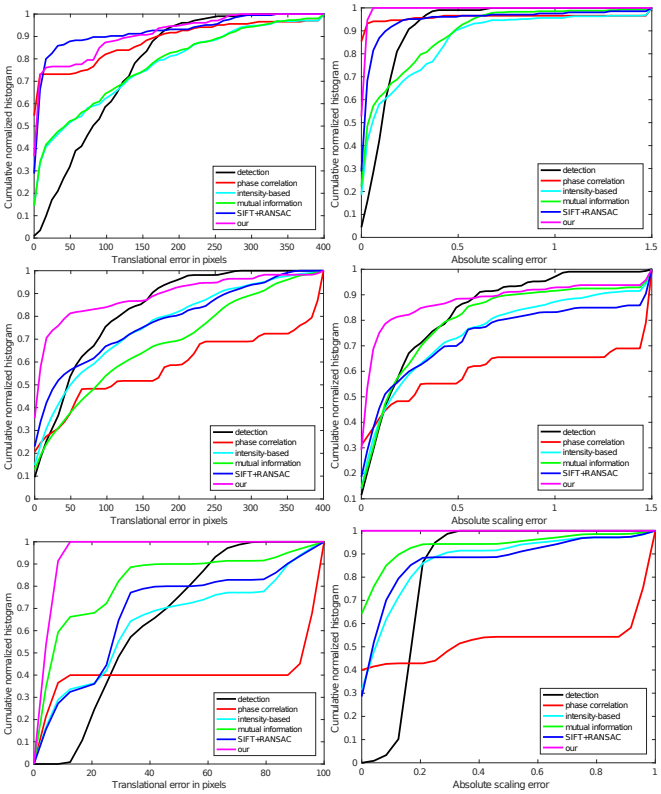

Fig. 10. Registration errors in Varcity (top), ZuBuD (middle) and NancyLights (bottom)

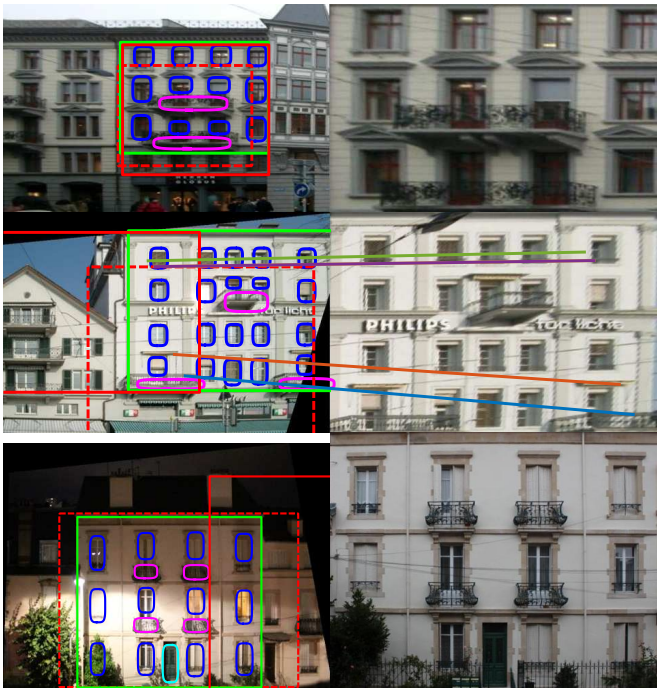

Fig. 11. Examples where other methods (red) fail to estimate the registration whereas our method (green) succeeds. The initial (dashed line) and final (plain line) registered reference boundaries overlay the target image. Top: intensity-based registration fall into local minima. Middle: SIFT+RANSAC registration fails due to facade symmetry. Bottom: strong change of illumination makes phase-correlation registration to fail.

as π_{k_j} values can decrease if a component is not visible. Acting as a regularizer, the Dirichlet prior on mixture weights avoids complete ignorance of data by keeping the mixture weights close to their original values α_{k_j} as shown in Fig. 12. These results were obtained with $p = 2$ but this behavior is independent of the value of p.

The visual appearance of facades can change a lot : windows can change according to sun reflexions and to the presence of closed shutters, balconies orientation are dependent on viewpoints. If it can be noticed on ZuBuD it is clearer for the last database where the robustness to illumination changes is evaluated (Fig. 10, bottom). Relying on semantic segmentation enables our method to focus on the geometric structure of the facade whereas the changes in appearance are encoded in the network. The illumination invariance of the network handles extreme changes in lighting that make other methods fail (Fig. 11).

Though the semantic segmentation prior $P(l_j|i,I)$ is not updated during the EM, label assignments can change from one iteration to another as shown in Fig. 2 where the prior and the posterior semantic segmentation are shown: some points not classified as windows in the prior are correctly classified in the posterior segmentation. The doors on the ground-floor wrongly classified in the prior segmentation are progressively correctly classified as well as they are guiding the registration process. Globally, if misclassification is common for visually similar labels like "door" and "window", the prior probability of the expected label can be increased by the generalized Gaussian influence during registration.

C. Method analysis and discussion

Our approach is well suited for images with sparse structures as facades but cannot be generalized to all kind of images because of spatial distributions chosen to model them (generalized Gaussians and uniform distribution for outliers). Moreover, in cases where data points are close to a uniform distribution densely sampled (e.g. a facade densely covered with windows), the method tends to label all points as outliers or as belonging to one Gaussian if the initialization is not close enough (see e.g. Fig. 13).

Using ground-truth semantic references can be seen as a limitation as this kind of information is not easily available for augmented reality or robotics applications. However, ground-truth segmentation may be carried out by precise and efficient automatic methods, although costly in terms of computing time. We are thinking of methods based on shape grammars [55]. The structural organization of facades, which originates from architectural rules, make them good candidates for this kind of model. The fact that these methods are expensive in computing time is not an issue since these operations can be done offline, and must be done only once. Another possible way to automate segmentation in the reference images would be to simply use our segmentation network. An example is shown in Fig. 14. Registration from a manually labeled reference image (left) is compared with registration from an automatically segmented one (right). Au-

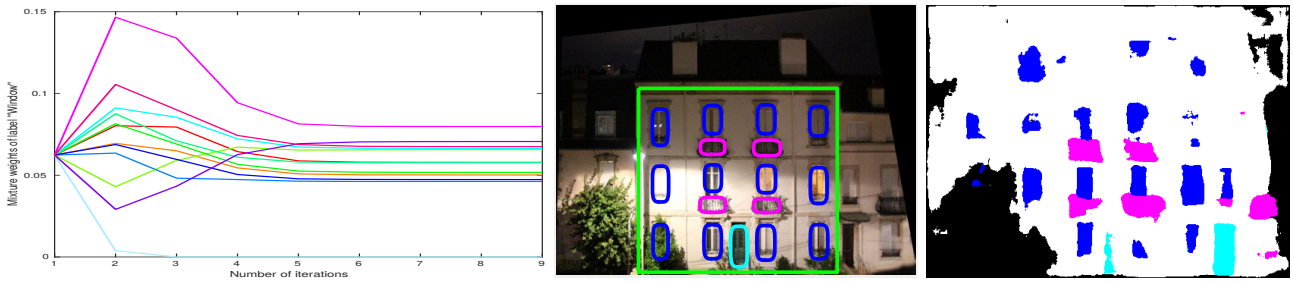

Fig. 12. Evolution of the mixture weights of label "window" over the iterations (left) until the registration converges in the target image (middle). The weight of the occluded bottom-left window drops (light blue). The Dirichlet distribution keeps the other weights close to their initial values despite the bad prior segmentation (right).

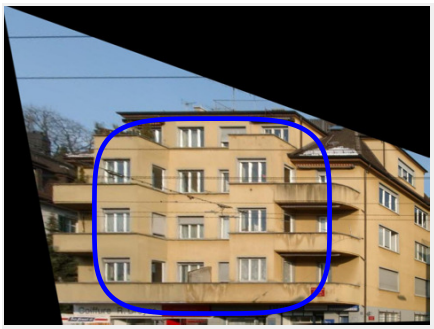

Fig. 13. The unique presence of the "window" semantic label in conjunction with non respect of the planar geometry hypothesis led our method to end in a local minimum where all points are assigned to a single large Gaussian.

tomatically calculated labels are much more noisy than those set manually, but the noise is partially neutralized during the registration process thanks to the overall coherence of the labels, leading to a less precise but not outlier result. Moreover, automatic segmentations may, again, be offline post-processed by introducing regularizing information based on architectural rules [56], [57].

Our method depends on the accuracy of the facade rectification resulting from the detection of Manhattan vanishing points. Figure 15 sheds light on the sensitivity of segmentation and registration to the accuracy of this stage. A Gaussian noise of standard deviation varying from 0.2 to 1.2 was applied to the ends of the line segments used for the vanishing point calculation (ten times for each noise level), which results in more imprecise rectifications than what we normally obtain. Segmentation and registration are performed on the warped images and the registration results are transferred back-again into the unwarped image. This experiment thus contributes to evaluate how rectification and segmentation errors affect the whole process. Figure 15(top) shows the resulting variability of the facade boundaries, which is increasingly greater as the noise increases. The graphics plot the mean orientation errors and the mean position relative errors versus the noise level. These errors are lower than 2 and, respectively, 3% as long as the faade deformation induced by Gaussian noise remains reasonable. Beyond that (Fig. 15(bottom-left), it is no longer a question of deformations obtained due to inaccuracies in the line segment detection, but rather to a notable failure of the vanishing point extraction procedure, against which nothing can be done anyway. It can be noticed that the semantic segmentation itself is very little affected by these geometric deformations, even when they are relatively large as in the example in Fig. 15.

Finally, a mean to improve the accuracy of the framework and to reduce the sensitivity of the semantic segmentation to scale could be to perform semantic segmentation at each step of the EM algorithm. Indeed, the semantic segmentation network was mainly trained with close-up facades, which explains that the segmentation is all the better the closer we are to the boundaries of the facade. In order to allow for convergence, segmentation is done on an enlarged window (40%) around the predicted position of the facade, in order to be as close as possible to the facade but to be sure that the targeted facade is inside the window. Performing semantic segmentation at each step of the EM in conjunction with a progressive decrease of the enlargement factor could thus improve the quality of segmentation as the transformation becomes close to the actual one. We do not adopt this strategy in our experiments in order to meet real time.

VI. Conclusion

We have presented a Bayesian model to solve jointly facade registration and semantic segmentation. The method is efficient and handles registration issues like occlusions, repetitions and changes in illumination.

Registration is currently done from one facade in the presented work. However, if a full model of a building is available we could also extend this work to perform co-registration of the visible facades.

Also in our tests, the initialization was close enough to the solution to assume that the semantic segmentation inference was stable enough and does not need to be re-estimated online. In future work, this assumption could be relaxed in the model to improve accuracy and dependence on initialization.

Declaration of competing interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

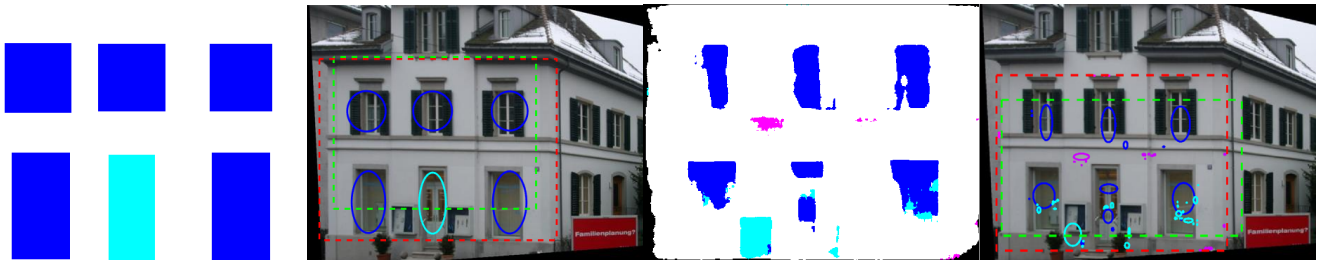

Fig. 14. On the left, registration is performed from a handmade ground truth, on the right, from a semantic segmentation inferred by the network.

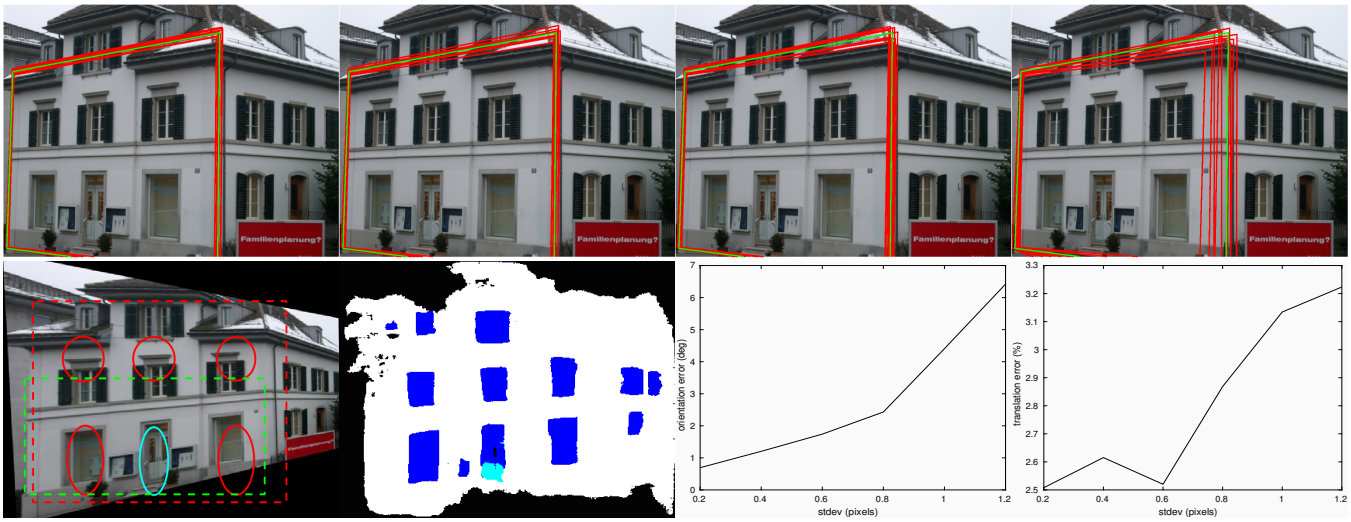

Fig. 15. Registration versus vanishing points accuracy. Top: Gaussian noise of standard deviation 0.2, 0.4, 0.8 and 1.2 pixels (resp. from left to right) is applied to the ends of the line segments used to compute the vanishing points. Registration results are shown for ten trials per noise level. The reference semantic segmentation is shown in Fig. 14 left. The initial box is the same for all trials and is shown in dashed green in the bottom left image, obtained with 1.2 pixel noise. It can be seen (next image) that the targeted semantic segmentation is little affected by the geometric deformation, which is quite large in this example. Bottom right : orientation and relative position errors are plotted in function of the noise level.

Acknowledgments: The work of Antoine Fond was funded by the *French Ministre de lEnseignement suprieur de la Recherche et de lInnovation* and by Inria.

REFERENCES

[1] J. Krolewski and P. Gawrysiak, "The mobile personal augmented reality navigation system," in *Man-Machine Interactions 2*, T. Czachórski, S. Kozielski, and U. Stańczyk, Eds., Berlin, Heidelberg, 2011, pp. 105–113.

[2] D. M. Chen, G. Baatz, K. Koser, S. S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk, "City-scale landmark identification on mobile devices," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011, pp. 737–744.

[3] K. Xu, A. D. Cheok, K. W. Chia, and S. J. D. Prince, "Visual registration for geographical labeling in wearable computing," in *Proceedings. Sixth International Symposium on Wearable Computers,*, 2002, pp. 109–116.

[4] A. Wendel, A. Irschara, and H. Bischof, "Natural landmark-based monocular localization for mavs," *IEEE International Conference on Robotics and Automation*, pp. 5792–5799, 2011.

[5] P. A. Zandbergen and S. J. Barbeau, "Positional accuracy of assisted gps data from high-sensitivity gps-enabled mobile phones," *Journal of Navigation*, vol. 64, no. 3, pp. 381–399, 2011.

[6] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla, "Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, United States, June 2018, pp. 18–23.

[7] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2599–2606.

[8] Y. Li, N. Snavely, D. P. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3d point clouds," in *Large-Scale Visual Geo-Localization*. Springer, 2016, pp. 147–163.

[9] T. Sattler, B. Leibe, and L. Kobbelt, "Improving image-based localization by active correspondence search," in *European conference on computer vision*. Springer, 2012, pp. 752–765.

[10] ——, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1744–1756, 2017.

[11] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, Los Alamitos, CA, 1999, pp. 1150–1157.

[12] K. M. Yi, E. Trulls Fortuny, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," *European Conf. on Computer Vision*, vol. 9910, pp. 17. 467–483, 2016.

[13] J. L. Schönberger, H. Hardmeier, T. Sattler, and M. Pollefeys, "Comparative evaluation of hand-crafted and learned local features," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 6959–6968.

[14] N. Sünderhauf, F. Dayoub, S. Shirazi, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," *International Conference on Intelligent Robots and Systems (IROS)*, pp. 4297–4304, 2015.

[15] A. Fond, M.-O. Berger, and G. Simon, "Facade Proposals for Urban Augmented Reality," in *IEEE International Symposium on Mixed and Augmented Reality*, Nantes, France, Oct. 2017, pp. 32–41.

[16] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional net-

information," *International Journal on Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.
[40] J. P. Pluim, J. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE transactions on medical imaging*, vol. 22, no. 8, pp. 986–1004, 2003.
[41] A. Dame and E. Marchand, "Accurate real-time tracking using mutual information," in *IEEE International Symposium on Mixed and Augmented Reality*, 2010, pp. 47–56.
[42] B. S. Reddy and B. N. Chatterji, "An fft-based technique for translation, rotation, and scale-invariant image registration," *IEEE Transactions on Image Processing*, vol. 5, no. 8, pp. 1266–1271, 1996.
[43] S. Zokai and G. Wolberg, "Image registration using log-polar mappings for recovery of large-scale similarity and projective transformations," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1422–1434, 2005.
[44] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
[45] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conf. on Computer Vision*, Zurich, Switzerland, Sept. 2014, pp. 391–405.
[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1097–1105.
[47] G. Simon, A. Fond, and M.-O. Berger, "A Simple and Effective Method to Detect Orthogonal Vanishing Points in Uncalibrated Images of Man-Made Environments," in *Eurographics*, Lisbon, 2016, pp. 33–36.
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conf. on Computer Vision*, 2014, pp. 346–361.
[49] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, USA, 2005, p. 539546.
[50] S. Yu, A. Zhang, and H. Li, "A review of estimating the shape parameter of generalized gaussian distribution," *Journal of Computational Information Systems*, vol. 8, pp. 9055–9064, 11 2012.
[51] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.
[52] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellen, and W. Eubank, "Nonrigid multimodality image registration," *Medical imaging*, vol. 4322, no. 1, pp. 1609–1620, 2001.
[53] R. Smriti, D. Stredney, P. Schmalbrock, and B. Clymer, "Image registration using rigid registration and maximization of mutual information," in *MMVR13. The 13th Annual Medicine Meets Virtual Reality Conference, Long Beach, CA*, 2005, p. 74.
[54] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[55] O. Teboul, I. Kokkinos, L. Simon, P. Koutsourakis, and N. Paragios, "Parsing facades with shape grammars and reinforcement learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1744–1756, 2013.
[56] M. Kozinski, R. Gadde, S. Zagoruyko, G. Obozinski, and R. Marlet, "A mrf shape prior for facade parsing with occlusions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2820–2828.
[57] A. Cohen, A. G. Schwing, and M. Pollefeys, "Efficient structured parsing of facades using dynamic programming," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3206–3213.

Antoine Fond is a Computer Vision Research Scientist at Synthesia in London. He was a member of the Inria-project Magrit between 2014 to 2018. His research interests include computer vision and deep learning. He received a MS in Robotics from Ecole Centrale de Nantes in 2014 and a PhD in Computer Science from University of Lorraine in 2018

work for real-time 6-dof camera relocalization," in *IEEE International Conference on Computer Vision*. IEEE Computer Society, 2015, pp. 2938–2946.
[17] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother, "Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2016, pp. 3364–3372.
[18] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic visual localization," *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 6896–6906, 2018.
[19] Y. Li, N. Snavely, and D. P. Huttenlocher, "Location recognition using prioritized feature matching," in *European conference on computer vision*. Springer, 2010, pp. 791–804.
[20] C. Arth, C. Pirchheim, J. Ventura, D. Schmalstieg, and V. Lepetit, "Instant outdoor localization and SLAM initialization from 2.5d maps," *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 11, pp. 1309–1318, 2015.
[21] H. Chu, S. Wang, R. Urtasun, and S. Fidler, "Housecraft: Building houses from rental ads and street views," in *European Conf. on Computer Vision*, 2016, pp. 500–516.
[22] G. Reitmayr and T. Drummond, "Going out: Robust model-based tracking for outdoor augmented reality," in *IEEE International Symposium on Mixed and Augmented Reality*, 2006, pp. 109–118.
[23] F. Castaldo, A. R. Zamir, R. Angst, F. Palmieri, and S. Savarese, "Semantic cross-view matching," in *2015 IEEE International Conference on Computer Vision Workshop, ICCV Workshops 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 1044–1052.
[24] P. David, D. DeMenthon, R. Duraiswami, and H. Samet, "Softposit: Simultaneous pose and correspondence determination," in *European Conf. on Computer Vision*, 2002, pp. 698–714.
[25] F. Moreno-Noguer, V. Lepetit, and P. Fua, "Pose priors for simultaneously solving alignment and correspondence," in *European Conf. on Computer Vision*, 2008, pp. 405–418.
[26] E. Serradell, M. Özuysal, V. Lepetit, P. Fua, and F. Moreno-Noguer, "Combining geometric and appearance priors for robust homography estimation," in *European Conf. on Computer Vision*, 2010, pp. 58–72.
[27] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet, "A survey on Visual-Based Localization: On the benefit of heterogeneous data," *Pattern Recognition*, vol. 74, pp. 90 – 109, Feb. 2018.
[28] G. Simon, "Tracking-by-Synthesis Using Point Features and Pyramidal Blurring," in *IEEE International Symposium on Mixed and Augmented Reality*, Basel, Switzerland, Oct. 2011, pp. 85–92.
[29] D. Robertson and R. Cipolla, "An image-based system for urban navigation," in *British Machine Vision Conference*, 2004, pp. 819–828.
[30] C. Harris and M. Stephens, "A combined corner and edge detector," in *In Proc. of Fourth Alvey Vision Conference*, 1988, pp. 147–151.
[31] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European Conf. on Computer Vision*. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 430–443.
[32] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *European Conf. on Computer Vision*, pp. 404–417, 2006.
[33] B. D. Lucas, T. Kanade, *et al.*, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 2. Vancouver, BC, Canada, 1981, pp. 647–679.
[34] G. D. Hager and P. N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 1025–1039, 1998.
[35] S. Baker and I. Matthews, "Equivalence and efficiency of image alignment algorithms," in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2001.
[36] S. Benhimane and E. Malis, "Real-time image-based tracking of planes using efficient second-order minimization," in *Proceedings of the International Conference on Intelligent Robots and Systems*, 2004, pp. 943–948.
[37] F. Jurie and M. Dhome, "Real time robust template matching," in *British Machine Vision Conference*, 2002, pp. 1–10.
[38] J. Kim and J. A. Fessler, "Intensity-based image registration using robust correlation coefficients," *IEEE transactions on medical imaging*, vol. 23, no. 11, pp. 1430–1444, 2004.
[39] P. Viola and W. M. Wells III, "Alignment by maximization of mutual

Gilles Simon is an associate professor at the University of Lorraine, France, working with the Inria project-team Magrit which is part of the LORIA laboratory, a combined Research Unit (UMR 7503) common to CNRS, INPL, Inria and University of Lorraine. His research interests include computer vision and augmented reality. He received a PhD in Computer Science in 1999 from the University of Nancy 1. In 2000, he worked as a research assistant in the Visual Geometry Group of the University of Oxford, UK.

Marie-Odile Berger is a senior researcher at Inria Nancy Grand Est. She is the head of the Inria project-team Magrit. Her research interests include augmented reality, computer vision, and medical imaging. She received a BS in mathematics in 1986 and a PhD in computer science in 1991 from the National Polytechnic Institute of Lorraine.

APPENDIX A. SUPPLEMENTARY MATERIAL

Supplementary material is provided as a separate file.