MATCHING TERRESTRIAL IMAGES CAPTURED BY A NOMAD SYSTEM TO IMAGES OF A REFERENCE DATABASE FOR POSE ESTIMATION PURPOSE

Arnaud Le Bris and Nicolas Paparoditis

Université Paris-Est, IGN, Laboratoire MATIS
73 Avenue de Paris
94165 SAINT-MANDE Cedex - FRANCE
arnaud.le-bris@ign.fr, nicolas.paparoditis@ign.fr
http://recherche.ign.fr/labos/matis/

Commission III, WG III/3

KEY WORDS: Registration, SIFT, tie points, facades, urban terrestrial image database

ABSTRACT:

Mobile mapping systems have been developed to achieve a fast automated acquisition of huge quantity of georeferenced terrestrial images in urban cities. Stereopolis is such a system making it possible to capture panoramic groups of images. These georeferenced photos are then stored in urban images street scale reference databases.
The issue investigated in this paper is the problem of tie points extraction between new images captured with an approximate georeferencement by a "nomad system" and images from the reference database in order to estimate a precise pose for these new images. Because of several difficulties (diachronism, viewpoint change, scale variation, repetitive patterns) extracting enough correct well distributed tie points is difficult and directly extracted and matched SIFT keypoints from original images are most of the time not sufficient. Nevertheless, results can be improved using ortho-rectified images on the facade plane instead of original images. Rectification parameters (3D rotation) are obtained from the coordinates of vanishing points corresponding to the two main directions of the facade. These points can indeed be extracted from linear features of the facade on the images. However, many point matches remain false and difficult to detect using only their image coordinates. The use of both image coordinates and scale and orientation associated to the matched SIFT keypoints makes it possible to detect outliers and to obtain an approximate similitude model between the two ortho-images. A more accurate model can then be computed from correct tie points.

1 INTRODUCTION

1.1 Context

These last years, mobile mapping systems have been designed and developed to achieve a fast automated acquisition of huge quantity of georeferenced terrestrial images in urban cities. Stereopolis is such a system. It is a vehicle surmounted by a crown of cameras, making it possible to capture panoramic groups of images.These georeferenced photos are then stored in urban images street scale reference databases and can be used for several applications or simply displayed owing to special applications such as for instance (iTowns, last visited on the 1st of March 2010) making it possible to navigate within the image flow (in panoramic geometry) and exploit features extracted from images to provide high semantic level data.
Tools could now be developed to offer the possibility for a user to georeference its own images from the database images. Such tools could then be used for several applications, such as :

- enrich the database : images with better resolution than the ones from the database are captured to enrich the database, especially for monuments.
- update the database : new images are captured punctually to update the database. (for example, photos of the facade of a new shop)
- image based location : an image is taken by a user with a mobile phone + camera + GPS (+ INS), and sent to the tool. It is then matched to the database and georeferenced.

These new images have not necessarily been captured by a precise georeferencing image acquisition system. Therefore, they have to be precisely located to be added to the database (or to an additional database) and their pose have to be coherent with the one of the images of the database. Nevertheless, it can be assumed that an approximate pose is available. Such information can indeed have been obtained either directly from the acquisition system - for example a mobile phone + camera + GPS (+INS) - or by the user (approximate location indicated on a map, name of the street...). As a consequence, the main issue in this paper is not to find matching images in the whole database for the new photos but to extract tie points between them and the images of the database in order to be able to georeference them. Nevertheless, tie points between these two sets of images are not easy to detect and several difficulties are encountered (as explained in section 2).

1.2 Input data

It must here be said that no real data captured by a "nomad system" - such as a mobile + camera + GPS (+INS) - was available and that the methods have been tested on images captured by a standard digital camera (without georeferencing system). The only information about georeferencement is the name and the side of the street.
On the other hand, images of the database are precisely georeferenced. They have indeed been captured by the mobile mapping system Stereopolis.

1.3 SIFT keypoint matching

In this paper, SIFT is used to extract tie points between the new photos and images from the database. SIFT (*Scale Invariant*

Feature Transform) method is described with details in (Lowe, 2004). It is both a multiscale keypoints detector (also known as DoG) and a points matching method.
First, for each image, keypoints are detected and a descriptor called SIFT is computed for each of them :

1. Keypoints are detected in each image. These points have 3 coordinates (image coordinates (x,y) and scale).
2. An orientation is computed and assigned to each detected keypoint. (It corresponds to the main direction of pixel gradients in the neighbourhood of the keypoint).
3. A SIFT descriptor (relative to the orientation of the keypoint) is then calculated for each detected keypoint. This 128 dimensional vector describes the image behaviour in the neighbourhood of the point.

Tie points of an image pair can then be extracted by matching these keypoints : the matching keypoint in image 2 of a keypoint of image 1 is its nearest neighbour among keypoints of image 2 in SIFT descriptors space but this matching is rejected if it is too ambiguous (i.e. if the second nearest neighbour is too near). The matching method is described in details in (Lowe, 2004).

High quality of SIFT descriptors for robust tie points matching has been shown by (Mikolajczyk and Schmid, 2005). SIFT is indeed very robust to scale variations (Morel, 2008), rotations (in image plane), noise and illumination variations. It is also robust to limited viewpoint changes, e.g. affine distortions. A method called ASIFT or Affine-SIFT has been proposed by (Morel, 2009) to extract matching points even if there is an important viewpoint change between the two images.
On the other hand, SIFT is very sensitive to diachronism (when images of a same scene are very different from each other since they have been captured under different conditions). Nevertheless, it can be better to obtain no tie points rather than to obtain a majority of false matches.

For the experiments described in this paper, a modified implementation of sift++ (Vedaldi, 2007) is used to extract SIFT keypoints from images. ANN library - available at (Mount and Arya, last visited on the 22nd of March 2009) and described in (Arya et al., 1998) - is used in the matching process and more precisely to find (approximate) nearest neighbours in SIFT descriptor space.

2 DIFFICULTIES

Several difficulties are encountered when trying to match new images to images from the database and to extract tie points between them. First of all, as these two sets of images have not been captured at the same time, diachronism can be important. Second, they can have been captured under different viewpoints. In addition to these difficulties, facades usually include repetitive objects (such as windows) or details, increasing the difficulty to find correct tie points.

2.1 Diachronism

Diachronism is an important limit in this context. As images to be registered and reference images have not been captured at the same time (neither the same day, nor at the same hour), a same facade can appear very different from one image to an other. Two distinct phenomena must be distinguished.

2.1.1 Illumination and shadows As images have not been captured at the same time, illumination conditions can change a lot. Thus, radiometry and contrast can greatly vary between the two sets of images, giving a different appearance to a same facade.
Variations of shadows (caused by architectural patterns and facade furniture) are an extreme case of this phenomenon : shadows can be very different on the two series of images, and can even be inverted (if one was captured during the morning and the other one during the afternoon), making it difficult to find correct and precise tie points.

2.1.2 Changes Some objects of the scene can have changed between the two data acquisitions. First of all, moving objects such as pedestrians and cars are obviously most of the time different from a set of images to an other. Facades can also be partly masked by other objects such as for instance scaffoldings.
Second, facade furniture can be in different positions : for example some windows and some shutters can be closed on an image and opened on the other one. Besides, reflections on the windows are also often very different from one set of images to an other.
Last, facades can really have changed, particularly if new images are added to the database in order to update it.

2.2 Different viewpoints and scale

Images can have been captured from quite different viewpoints (orientation and position). However, facades usually include repetitive objects (such as windows and balconies) or decorative details. Repetitive structures combined to viewpoint change often lead to matching errors even for images captured at the same time. As long as viewpoint change between the two images to match is not too important, perspective distortion remains limited and matching performs good. On the contrary, when viewpoint change becomes to be more important, this perspective distortion tends to be sufficient to lead to the next situation : a point corresponding to a repetitive pattern in image 1 will not be matched to its true homologous (appearing under a too distorted appearance), but to a same pattern appearing in image 2 with quite the same viewpoint as this point in image 1.
In addition to this, the two sets of images can not have been taken at the same scale, that is to say, neither at the same distance from the facades, nor with the same focal distance. This increases the previous phenomenon.

3 FIND MATCHING IMAGES IN THE DATABASE

Large scale image search and retrieval in huge image database is a very important topic in computer science. Several approaches have been developed and presented in literature.
As explained in section 1.1, it is assumed here that an approximate georeferencement of the query images is available. Such information can have been obtained either directly during data capture (for instance if a light "mobile mapping system" - such as a mobile phone + camera + GPS (+INS) - has been used), or afterwards (for example owing to an interface making it possible for the user to set an approximate position of the images on a map, or even among the panoramic images of the database). Therefore, it is not necessary to search matching images among all available images of the database, as in (Picard et al., 2009), but only in a limited part of the database (such as a street). In this more simple case, existing methods for image retrieval in databases can obviously still be applied to this part of the database.
For the experiments described in this paper, two very simple and quite "naive" methods have been used to find corresponding images in the database. A matching score is computed for each image of the database.

- The query photo is matched to each candidate image from the database (standard SIFT matching method mentioned in section 1.3). For each possible pair of images, the SIFT keypoints of the query image are matched to the SIFT keypoints of the image of the database. The matching score is the number of obtained matched points. (No elimination of outliers is performed, since because of repetitive patterns on facades they can be false tie points but correspond to the same kind of object on the same facade.)
- The k nearest neighbours among SIFT keypoints from the candidate images of the database are found for each keypoint of the query image. The matching score is the number of nearest neighbours found in each database candidate image.

Stereopolis images (i.e. images of the database) have been captured with overlap along the street. As a consequence, query image matches most of the time to several images of the database. Therefore, the matching images correspond to the spatial cluster of images with highest matching scores. (In the present experiments, approximate location is the name of the street. Thus, matching images corresponds to a cluster of images with high matching score along this street.)
To reduce computing time, downsampled images can be used at this step.

4 EXTRACT TIE POINTS : FIRST EXPERIMENTS WITH "NAIVE" METHOD

In these first experiments, the method tested to extract tie points between the query image and its matching image from the database is very simple and naive. It just consists in extracting SIFT keypoints from the images and to match them. Obtained results vary obviously a lot.
The filtering method consists simply in computing a 2 dimensional 8 parameters homography between the two images, taking into account that both of them are photos of facades, i.e. of planar objects.

4.1 Positive example : good results directly with original images

In the next example (figure 1 and 2), no problem is encountered and matching performs very well directly with original images. Matched points are mostly correct and their distribution is quite good. There are enough inliers to detect and filter outliers.

Figure 1: Raw matches (On the left, database image. On the right, query image)

4.2 Bad results with original images, but good results with rectified images

In the next example (see figure 3), results obtained by direct matching between the original images are quite bad and not sufficient to compute a good model. Matched points are mostly false.

Figure 2: Query image (on the left) is registered to the database image (on the right) according to 2D homography (computed from extracted tie points)

It must here be said that better results are obtained with an other

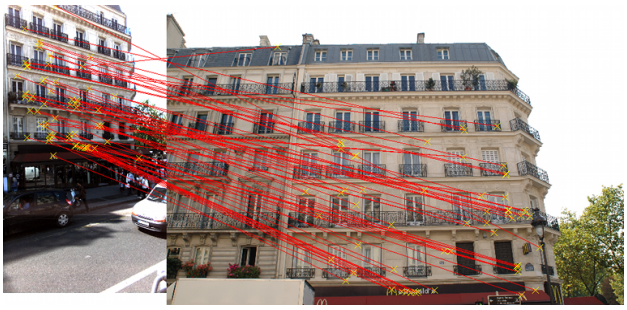

Figure 3: Raw matches (On the left, database. On the right, new image)

image of the database with a nearest viewpoint than this one. Therefore, the problem of viewpoint change mentioned in section 2.2 is very likely encountered here : strong viewpoint changes combined with repetitive structures (windows, balconies, decorative details) lead to some false matches.
As a consequence, two possible solutions could be used to improve results :

- More correct tie points could be detected using a multiple matching method. Such method does not find only the best matching point in image 2 for each point of image 1 and does not reject a match if it is too ambiguous, that is to say if there is an other good candidate. On the contrary, it detects several possible matching points in image 2 for each keypoint of image 1.
  - (Cléry, 2009) detects keypoints of image 2 belonging in SIFT space to the neighbourhood of the keypoint of image 1 (the radius of this neighbourhood is a parametre of the method). The number of correct matches increases with such a method. The number of false matches increases too, but it must be said that the less probable matches can often be rejected immediately during matching. Outliers are then detected during a model estimation owing to a RANSAC process.
  - *A contrario* matching methods as the one presented in (Rabin et al., 2008) could be a good solution but seem to be quite long.

- Perspective distortions caused by viewpoint changes could be corrected :
  - ASIFT (described in (Morel, 2009)) simulates several affine distortions for the two images. It extracts keypoints and tries to match them on each possible pair of transformed images. To be more efficient, the algorithm is first applied to downsampled images in order to identify the best pairs of affine transforms.
  - In this context, both images (query and database) are photos representing facades of buildings, i.e. planar objects with many linear vertical and horizontal features. As a consequence, these images can be simply orthorectified on the facade plane, using vanishing points.

The chosen method is the last one, which consists in working on rectified images on the facade plane. It is a quite fast and simple solution adapted to the present problem. It is described in section 5.1.
As it can be seen on figure 4, better results are obtained with rectified images. A satisfying number of correct tie points is detected. Furthermore, these points are well distributed over the overlapping area.

Figure 4: Raw matches between rectified images (on the left : database image, on the right : query image)

4.3 Bad results even with rectified images

In this example (figure 5), obtained results remain bad even with ortho-rectified images. This can be explained by the presence of repetitive patterns, radiometric changes on facades (and even contrast inversion on some windows).
Nevertheless, very few correct tie points are extracted among many false matches, making it difficult to detect and filter false matches by simply computing a 2D homography transform between the 2 images.

5 PROPOSED METHOD

5.1 Work on rectified images on the facade plane

As it was shown in section 4.2, results can be improved when working with rectified images on the facade plane instead of original images. This is indeed a way to limit perspective distortion caused by viewpoint change between the two images. This proposed method is a quite fast and simple solution well adapted to the present problem.
Since both images (query and database) are photos representing facades of buildings, i.e. planar objects, they can be orthorectified on the facade plane. Besides, facades contain many linear vertical and horizontal features, that can be detected and used to extract vanishing points. The algorithms described in (Kalantari et al., 2008) and (Kalantari, 2009) are used to extract the vanishing points corresponding to the two main directions of the facade from the images.
When the horizontal and vertical vanishing points have been calculated, the (3D) rotation from the image plane to the facade plane can be estimated, making it possible to (ortho-)rectify the image on the facade plane up to a scale factor.
SIFT keypoints are then extracted from these rectified images and matched. It must here be kept in mind that the transform from ortho-rectified image coordinates to original image coordinates is known, making it possible for extracted tie points from rectified images to be used for pose estimation of original images.

Figure 5: Bad results, even with rectified images. (on the left : database image, on the right : query image)

5.2 Detect false matches using information related to SIFT keypoints

At this step, SIFT keypoints have been extracted from rectified images on the facade plane and have been matched. Nevertheless, (as on figure 5) there are sometimes few correct detected tie points among many false matches, making it difficult to detect mistakes by simply detecting outliers when directly estimating a 2D transform between the 2 images.
Work is performed with (ortho-)rectified images on the facade plane. As a consequence, the transform between these ortho-images is a 2D similitude, i.e. a scale factor combined with a 2D rotation and 2D translation in facade plane. Let this transform be

$$S(x;y) = E \cdot \begin{pmatrix} cos(\alpha) & sin(\alpha) \\ -sin(\alpha) & cos(\alpha) \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} T_x \\ T_y \end{pmatrix}$$

The parameters of this transform are estimated in two steps. As explained in section 1.3, an orientation and a scale are associated to each SIFT keypoint. Let $P_i(x_i, y_i, \sigma_i, \theta_i)$ be a SIFT keypoint in image i with $(x_i; y_i)$,σ_i and θ_i standing respectively for image coordinates, scale and orientation angle associated to P_i. Therefore only one correct match (P_1, P_2) of SIFT keypoints from the two images makes it possible to estimate approximate values for all parameters of the 2D similitude.

1. Approximate values for scale factor E and rotation angle α are first estimated not from the image coordinates of the matched points but from other information related to keypoints : $E = \frac{\sigma_1}{\sigma_2}$ and $\alpha = \theta_1 - \theta_2$

2. Second, the 2D translation vector $(T_x; T_y)$ of the similitude can be estimated from the image coordinates of the two matched keypoints, taking into account the scale factor and the rotation estimated at previous step.

5.2.1 Estimate scale and rotation A hypothesis for scale and rotation parameters is estimated for each obtained keypoint match. The histogram of these possible values is calculated (to describe their distribution). The pair of scale and rotation parameters corresponding to the maximum of this histogram is selected.
Knowing these parameters, many false keypoint matches can be detected and filtered if their scale ratio or their difference of orientation is too different from the approximated scale factor and rotation angle previously estimated.

5.2.2 Estimate 2D translation parameters Many false matches have been filtered at previous step. A hypothesis for translation vector is estimated for each remaining keypoint match. The 2 dimensional histogram of the distribution of these calculated values is computed. As at previous step, the translation parameters corresponding to the maximum of the histogram are selected.
As at previous step, remaining false matches can then be detected and filtered.

5.3 Estimate more accurate parameters

At previous steps, approximate parameters of the similitude have been estimated and false matches have been detected and filtered. Therefore, it is now possible to estimate more accurate parameters and to restore correct matches that could have been considered as outliers at previous steps.
Correct tie points are now available and can be used to estimate the pose of the query image.

6 RESULTS

The method presented in section 5 has been tested on several images, giving good results even in the most difficult case shown by figure 6 (same photos as in figure 5).
To evaluate the proposed approach, database images have been registered to the query image according to a 2D 8-parameters homography model. Examples of obtained results are shown by figures 8, 7 and 6. Registered images are superposed with an image per channel : reference image (query image) is red channel whereas two registered database images are green and blue channels.

7 CONCLUSIONS AND FUTURE WORK

Because of combined difficulties such as diachronism, viewpoint change, scale variation and the presence of repetitive structures, extract correct tie points from new images captured by a nomad system and from images from infrastructure database is not an easy task. Obtained results with direct extraction and matching of SIFT keypoints from original images is often not sufficient to provide enough well distributed tie points to detect and filter false matches and compute image orientation. It must here be said that similar results have been obtained with other keypoint extraction methods such as Harris-Laplace, MSER, Harris-Affine, Hessian-Affine.
Nevertheless, as these images are photos of facades of buildings, results can be improved working with rectified images on the facade plane instead of original images. Vanishing points corresponding to the two main directions of the facade can indeed be

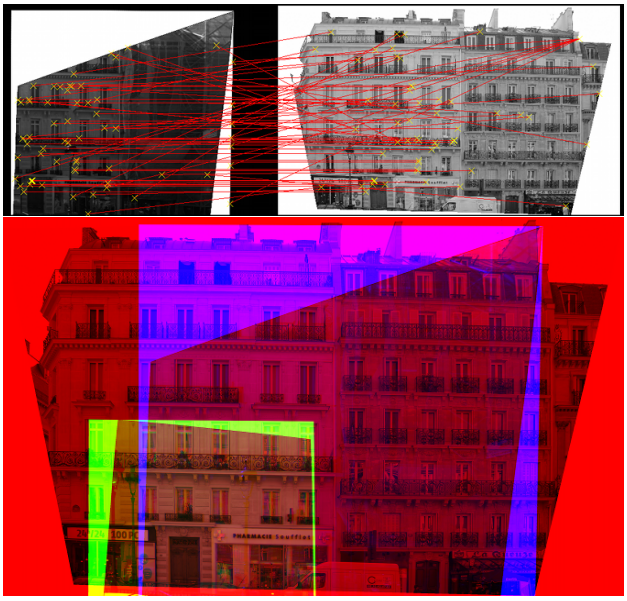

Figure 6: First row, raw matches between (ortho-)rectified images (on the left : database image, on the right : query image). Second row, registered images. (Same photos as figure 5)

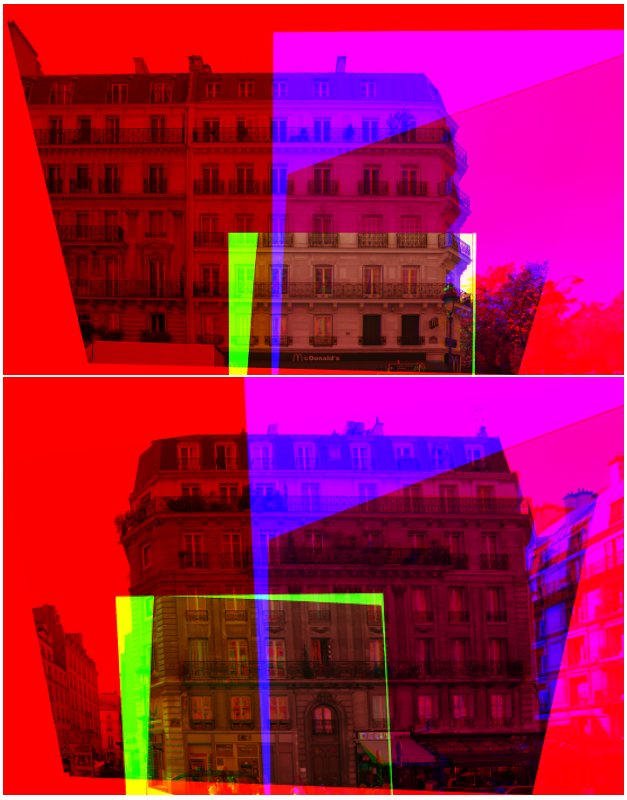

Figure 7: Registered images.

detected from image linear features, bringing orientation information making it possible to ortho-rectify images on the facade plane, up to a scale factor.
However, because of diachronism and repetitive patterns, many obtained point matches remain false and are sometimes difficult to detect using only their image coordinates. Use both image coordinates and other information (scale and orientation) associated to the matched SIFT keypoints makes it possible to detect outliers

Figure 8: Two first rows, raw matches between original images. Two middle rows, filtered matches between (ortho-)rectified images (on the left : database image, on the right : query image). Last row, registered images.

and to obtain an approximate model between the ortho-images. A more accurate model can then be computed from correct tie points and could be used as a predictor to obtain more tie points. At this step, correct tie points are available and can then be used to calculate the pose of the query image. For instance, robust tools described in (Kalantari, 2009) or (Kalantari et al., 2009) make it possible to obtain accurate relative pose from at least 5 points even in planar configurations. Multiple tie points could also be a way to obtain the pose of the query image, since their 3D ground coordinates can be obtained from the poses of database images. Other methods could now be used. For example, working with the whole image texture has advantages. With such methods, as in (Bentrah, 2006), not independent keypoints but textures are matched, making it possible to compute an initial transform which can then be used as a predictor to extract more tie points.

REFERENCES

Arya, S., Mount, D. M., Netanyahu, N. S., Silverman, R. and Wu, A. Y., 1998. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. Journal of the ACM 45(6), pp. 891–923.

Bentrah, O., 2006. Reconstruction de la géométrie d'acquisition de séquences d'images acquises par un véhicule en milieu urbain. PhD thesis, Institut National Polytechnique de Grenoble, France.

Cléry, I., 2009. Test d'algorithmes d'appariement robustes entre des images ambigues et prises dans des conditions différentes. Master's thesis, École Nationale des Sciences Géographiques, France.

iTowns, last visited on the 1st of March 2010. itowns. http://www.itowns.fr/index.html.

Kalantari, M., 2009. Approche directe de l'estimation automatique de l'orientation 3D d'images. PhD thesis, Université de Nantes, France.

Kalantari, M., Jung, F., Guédon, J. and Paparoditis, N., 2009. The five points pose problem : A new and accurate solution adapted to any geometric configuration. In: Lecture Notes in Computer Science (LNCS), Vol. 5414, pp. 215–226.

Kalantari, M., Jung, F., Paparoditis, N. and Guédon, J., 2008. Robust and automatic vanishing points detection with their uncertainties from a single uncalibrated image, by planes extraction on the unit sphere. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences (IAPRS) XXXVII(3A), pp. 203–208. Beijing, China.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), pp. 91–110.

Mikolajczyk, K. and Schmid, C., 2005. A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis ans Machine Intelligence (PAMI) 27(10), pp. 1615–1630.

Morel, J.M. et Yu, G., 2008. On the consistency of the sift method. Technical Report CMLA 2008-26, Centre de Mathmatique et de Leurs Applications (CMLA) UMR 8536, ENS Cachan, Cachan, France.

Morel, J.M. et Yu, G., 2009. Asift: A new framework for fully affine invariant image comparison. SIAM Journal on Imaging Sciences 2(2), pp. 438–469.

Mount, D. M. and Arya, S., last visited on the 22nd of March 2009. Ann: A library for approximate nearest neighbor searching. http://www.cs.umd.edu/ mount/ANN/.

Picard, D., Cord, M. and Valle, E., 2009. Study of sift descriptors for image matching based localization in urban street view context. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences (IAPRS) XXXVIII(3/W4), pp. 193–198. Paris, France.

Rabin, J., Delon, J. and Gousseau, Y., 2008. A contrario matching of sift-like descriptors. In: ICPR, Tampa, Florida, USA, pp. 1–4.

Vedaldi, A., 2007. SIFT++: a lightweight c++ implementation of sift detector and descriptor. http://www.vlfeat.org/ vedaldi/code/siftpp.html.