AUTOMATIC MRF-BASED REGISTRATION OF HIGH RESOLUTION SATELLITE VIDEO DATA

C. Platias, M. Vakalopoulou, K. Karantzalos

Remote Sensing Laboratory, National Technical University of Athens,
Zographou campus, 15780, Athens, Greece
platiasx@gmail.com; *mariavak@central.ntua.gr*; *karank@central.ntua.gr*

Commission I, WG V

KEY WORDS: Video Sequence, Co-registration, Descriptors, Deformable Registration, STAR, FREAK

ABSTRACT:

In this paper we propose a deformable registration framework for high resolution satellite video data able to automatically and accurately co-register satellite video frames and/or register them to a reference map/image. The proposed approach performs non-rigid registration, formulates a Markov Random Fields (MRF) model, while efficient linear programming is employed for reaching the lowest potential of the cost function. The developed approach has been applied and validated on satellite video sequences from Skybox Imaging and compared with a rigid, descriptor-based registration method. Regarding the computational performance, both the MRF-based and the descriptor-based methods were quite efficient, with the first one converging in some minutes and the second in some seconds. Regarding the registration accuracy the proposed MRF-based method significantly outperformed the descriptor-based one in all the performing experiments.

1. INTRODUCTION

Currently the remote sensing community is expecting during the following years a paradigm swift from spare multi-temporal to every-day monitoring of the entire planet through mainly micro-satellites at a spatial resolution of a few meters or centimeters (in the raster world), but also from other cutting-edge technology including hyperspectral sensors and UAVs. Moreover, apart from the standard imaging products video streaming from earth observation satellites significantly expands the variety of applications that can be addressed.

In particular, high resolution satellite video sequences [Murthy et al., 2014,d'Angelo et al., 2014,Kopsiaftis and Karantzalos, 2015] have become available and enrich the existing geospatial data and products. Skybox Imaging[1] and Urthecast[2] are already providing high resolution video datasets with a spatial/temporal resolution of approximately 1 meter and 30 frames per second. However, due to the continues movement of the satellite platform the acquired frames are not registered between each other. Moreover, in order to combine and fuse information from other geospatial data and imagery for any application or analysis their registration to a local/national geo-reference system is required. Therefore, the automated co-registration of video frames and/or their registration to a reference image/map is still an open matter.

The problem of image registration has been heavily studied and numerous approaches have been proposed [Zitova and Flusser, 2003, Sotiras et al., 2013]. The methods fall into two main categories depending on the employed model *i.e.,* rigid-based and non-rigid (deformable-) based ones. The first category consists of descriptor-based methods, which automatically detect and match points in the pair of images and then define a global transformation to register them. A variety of descriptors, such as SIFT [Lowe, 2004], ASIFT [Morel and Yu, 2009], SURF [Bay et al., 2008], DAISY [Tola et al., 2010], FREAK [Alahi et al., 2012], *etc* have been employed for a plethora of applications like face recognition, object identification, motion tracking and satellite imagery. Under such a framework one million of satellite RGB images have been registered by Planet Labs[3] in just one day [Price, 2015]. The second category contains non-linear registration methods. A similarity function is used to calculate the similarity of each pixel (from the first image) to a neighbourhood of pixels in the other image and find the best displacement which recovers the geometry. This kind of methods have been widely used in computer vision and medical imaging [Sotiras et al., 2013], while recently validated for very high resolution satellite data [Karantzalos et al., 2014] delivering high accuracy rates for both optical and multi-modal data.

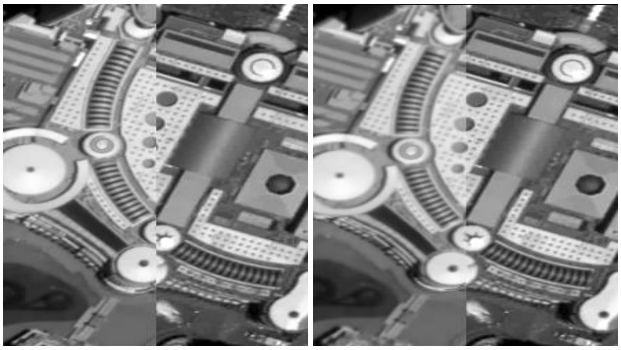


Figure 1: The developed methodology manages to co-register the acquired video frames. Unregistered frames (left), registered frames after the application of the developed method (right). Data are from Skybox Imaging (Terra Bella).

In this paper, a MRF-based registration framework is proposed for the co-registration of satellite video frames and/or their registration to a reference map/image (Figure 1). In particular, the developed method calculates a deformation map, while certain similarity functions (*e.g.,* normalised cross correlation, mutual infor-

[1] https://terrabella.google.com/
[2] https://www.urthecast.com/
[3] http://planet.com/

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-III-1-121-2016

mation, sum of absolute difference, *etc.*) were employed for calculating the displacement of every pixel. An energy formulation through an MRF model was defined and its minimization was performed using linear programming. The methodology was applied and validated based on Skybox Imaging data and certain corresponding reference images (Table 1). Experimental results were compared with the ones obtained from a descriptor-based technique [Price, 2015] which is based on a rigid registration framework using the STAR [Agrawal et al., 2008] and FREAK [Alahi et al., 2012] algorithms for establishing and matching correspondences. These correspondences were used for defining the homography transformation parameters and register the pair of images. Both methods have been quantitative and qualitative evaluated based on manually collected ground control points (GCPs).

2. METHODOLOGY

2.1 Image Registration

Lets denote in a pair of images I_t: $\Omega \mapsto \mathbb{R}^2$ as the reference/target image and I_s: $\Omega \mapsto \mathbb{R}^2$ as the source image that should be registered. The goal of registration is to define a transformation $T : \Omega \mapsto \mathbb{R}^2$ which will project the source to the target in the image pair.

$$I_t(x) = I_s(x) \circ T(x) \tag{1}$$

For the rigid registration, the displacement of each pixel in the image is calculated using the same transformation parameters. On the other hand, for the non-rigid registration the displacement of every pixel is calculated independently using only certain constraints for local smoothness defined by the model. Regarding the co-registration of satellite video frames, in our experiments the reference image corresponds to the first frame of the video sequence.

2.2 Rigid, descriptor-based registration

The most commonly used approach is based on a rigid registration [Le Moigne et al., 2011, Vakalopoulou and Karantzalos, 2014, Price, 2015] and calculates a global transformation for image pairs. The framework has four main components: i) the keypoint detector, which detects and holds the information about the position of every keypoint in each image, ii) the keypoint descriptor, which contains the characteristics of the keypoints, in order to be able to compare them, iii) the matcher, which matches the different keypoints in the source and target images and finally, iv) the image transformation method, which calculates the parameters of the transformation, based on the calculated correspondences.

For the evaluation of the proposed MRF-based approach the rigid registration method employed, here, is based on the recently proposed approach in [Price, 2015] including: a keypoint detector, the Star Detector (STAR), based on Center Surround Extremas (Censure) [Agrawal et al., 2008], a keypoint descriptor, the Fast Retina Keypoint algorithm (FREAK) [Alahi et al., 2012] and as matcher the brute force matcher (BFMatcher). Last but not least, the transformation used to register the source image to the target/reference was the homography one.

Generally speaking, the STAR algorithm detects numerous keypoints in each frame. Since the consecutive frames do not change a lot, many correspondences between the two frames were created. In order to reduce the outliers, the RANSAC [Fischler and Bolles, 1981] algorithm was used with a reprojection threshold of one pixel. Additionally, the false correspondences were removed, using a filter that allowed only matches below a specified threshold to participate to the transformation. The threshold was set to a fraction of the maximum distance between the matches. In all our experiments only those matches with a distance less than or equal to 65 percent of the maximum distance participated in the formulation of the transformation.

The homography parameters are defined after the minimization of the following error (Equation 2).

$$min[\sum_i (x_i^{'} - \frac{h_{11}x_i + h_{12}y_i + h_{13}}{h_{31}x_i + h_{32}y_i + h_{33}})^2 + (y_i^{'} - \frac{h_{21}x_i + h_{22}y_i + h_{23}}{h_{31}x_i + h_{32}y_i + h_{33}})^2] \tag{2}$$

where $h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32}, h_{33}$ are the homography parameters, x_i, y_i are the coordinates of the keypoint i in the reference image and $x_i^{'}, y_i^{'}$ the coordinates of the keypoint i in the source image.

2.3 The proposed MRF-based satellite video registration framework

The proposed approach is based on a deformable registration using different similarity metrics. A MRF model was defined and the solution is minimizing the following energy function (Equation 3) [Glocker et al., 2011]. The label space for the model contains all the possible displacements $(d^1, \ldots, d^n)$, such as: $l_p = [d^1, \ldots, d^n]$. A graph was superimposed on the target frame, and each node was connected to a neighbourhood of pixels using an interpolation function $\eta(.)$. The total energy was formulated as below:

$$E_{reg} = \sum_{p \in G} V_p + \lambda \cdot \sum_{(p,q) \in N} V_{pq} \tag{3}$$

where p, q are nodes in the graph G and N the neighbourhood of p in the other image, V_p is the unary term, V_{pq} is the pairwise term and λ is the weight which defines the use of the pairwise term in the energy minimization.

The unary and the pairwise terms are formulated as follows:

$$V_p = \int_\Omega \hat{\eta}(||x - p||)\rho(I_s(x), I_t(x + d^{l_p}))dx \tag{4}$$

where $\rho()$ is the similarity function used (normalised cross correlation, mutual information, *etc*). The interpolation function $\hat{\eta}$ which connects with a weight propositional to the distance the pixels with the nodes of the grid and reverse. A typical example of a projection function would be cubic B-splines which is the one employed here.

$$V_{pq} = ||d^{l_p} - d^{l_q}|| \tag{5}$$

where V_{pq} penalises neighbour nodes with different displacement labels depending on the difference of their displacement.

Dataset	Acquisition Date	Spatial Resolution	Frame Rate	Duration	Frames
Burj Khalifa	9 /4/2014	1m	30 f/s	30s	900
Las Vegas	25/3/2014	1m	30 f/s	60s	1800
Las Vegas-night	11/2014	1m	30 f/s	30s	887

Table 1: The satellite video datasets that were employed for the validation of the developed registration framework.

3. IMPLEMENTATION

The formulation follows a multiscale approach concerning both the image and the graph, meaning that the energy was calculated at different levels of the grid and the image. Concerning the grid levels a sparse grid was implemented and as the levels of the grid augmented, the grid became more and more dense. At each level a number of iterations was performed in order to calculate the minimum energy. In different grid levels the source image was transformed and updated, so in the next level it was closer to the target one. This way the label space for the displacements was also changing in each grid level, being closer to the optimal. Finally, for different image levels a subsampling of the image was performed for less computational complexity.

For the *Burj Khalifa*[4] Skybox video dataset the set of parameters was defined as follows. The node distance was set to 10 pixels, the grid levels to 3 and the image levels to 2 with 5 iterations at each level. The label space at each grid level changed to 0.8 times of the previous one. Normalized Cross Correlation (NCC) was used as the similarity function, which, according to the literature, performed better than other functions [Karantzalos et al., 2014] for the registration of remote sensing data. Finally, the lambda parameter was set to 40, the sampling steps to 25 and cubic b-splines was used as the interpolation function. All the parameters were tuned after grid search.

Using the above set of parameters, a co-registration between smaller groups was initially performed and then all groups were registered to the first frame. In particular, three groups with a lower number of frames and thus smaller displacements were formed *i.e.,* every 300 frames. The registration of each group was performed using as target image the 1^{st}, 300^{th} and 600^{th} frame, respectively. Then all were registered to the first one.

For the two *Las Vegas* Skybox video datasets, the configuration consisted as in the previous case of: a node distance of 10 pixels, 3 grid levels and 2 image levels. Moreover, the number of iterations was set to 15, the sampling steps to 65, lambda was set to 15 and the label space to 0.67 times the previous one for each grid level. The similarity function and the interpolation method was the same as for the *Burj Khalifa* sequence. Again the registration was performed firstly in groups and in particular, for the *Las Vegas*[5] dataset the grouping was every 300 frames and for the *Las Vegas-night*[6] video dataset every 150 frames.

4. EXPERIMENTAL RESULTS AND EVALUATION

The proposed MRF-based methodology was evaluated both qualitatively and quantitatively. For the quantitative evaluation a number of manually collected GCPs were selected. It is important to note, that for the descriptor-based approach a set of fixed parameters did not perform well for all the video frames, since even the smallest shift between the frames affected the keypoint detection and respectively the registration accuracy. For this reason, the tuning of the parameters was performed for each pair of frames using grid search. This was the main drawback of the descriptor-based framework since even though the multithreaded implementation in OpenCV [Culjak et al., 2012] requires two to three seconds per image pair, the manual tuning of the parameters required significantly more.

The experimental results included satellite video sequences of *Burj Khalifa*, *Las Vegas* and *Las Vegas Night* (Table 1) from Skybox Imaging. The main challenges for the registration of the video datasets were mainly the relative tall buildings, their shadows and any other moving object (*e.g.,* airplanes). In particular, the different angles of the sun and the satellite acquisition affect the geometry of terrain objects and their corresponding shadows.

For the quantitative evaluation the results after the implementation of both registration methods are presented in Table 2. In all cases the proposed MRF-based approach outperformed the descriptor-based one and managed to register all the different frames with a mean displacement error of less than 1.5 pixels. These errors correspond to the overall registration error from all frames since they were calculated between the first and last frame of the video dataset. The resutled higher registration errors from the descriptor-based approach along with the fact that these errors were not equally distributed in image plane indicated a significant lower performance than the proposed MRF-based approach.

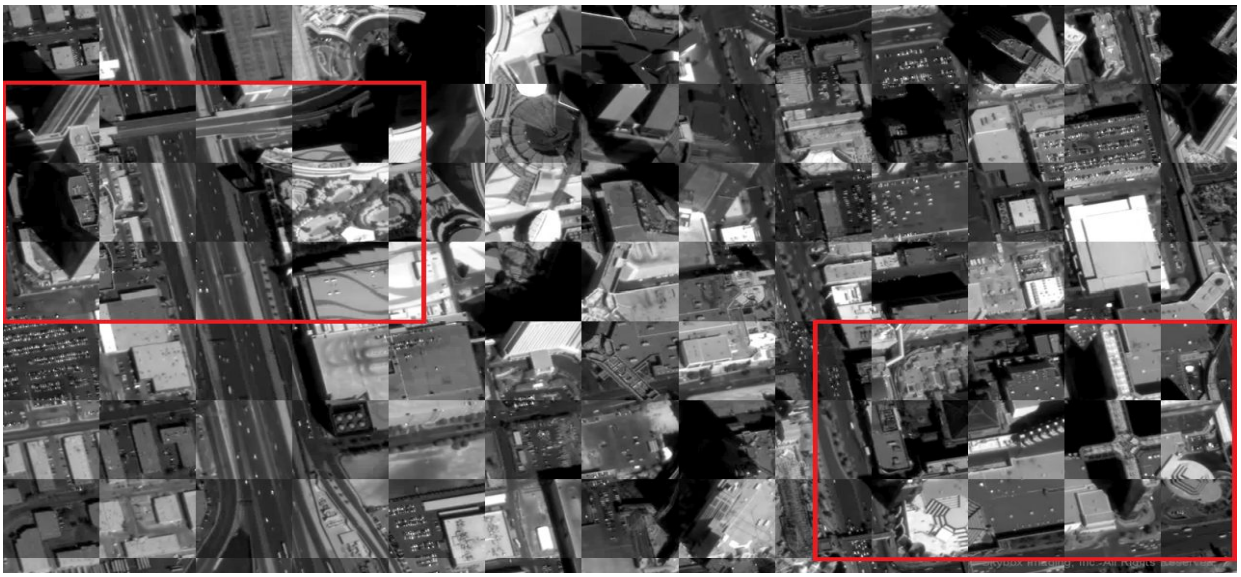
Moreover, the registration of the *Burj Khalifa* dataset to a Google Earth's image mosaic was performed using the proposed MRF-based approach. Quantitative results are quite promising with mean displacement errors less than 1.6 pixels (Table 3).

For the qualitative evaluation different checkerboard visualisations are presented in Figures 2, 4, 3, 5, along with certain zoom-in at selected sub-regions. Each checkerboard visualisation is a blend of the first and last frame of the unregistered and registered datasets. After a closer look on the marked with a red color areas one can observed that the unregistered data possessed large initial displacements. In particular, in Figure 2 one can observe quite large displacements between the different frames, with significant spatial discontinuities in roads, bridges and buildings (*e.g.,* inside the red circles). The MRF-based registration recovered the geometry and managed to register accurately the video frames.

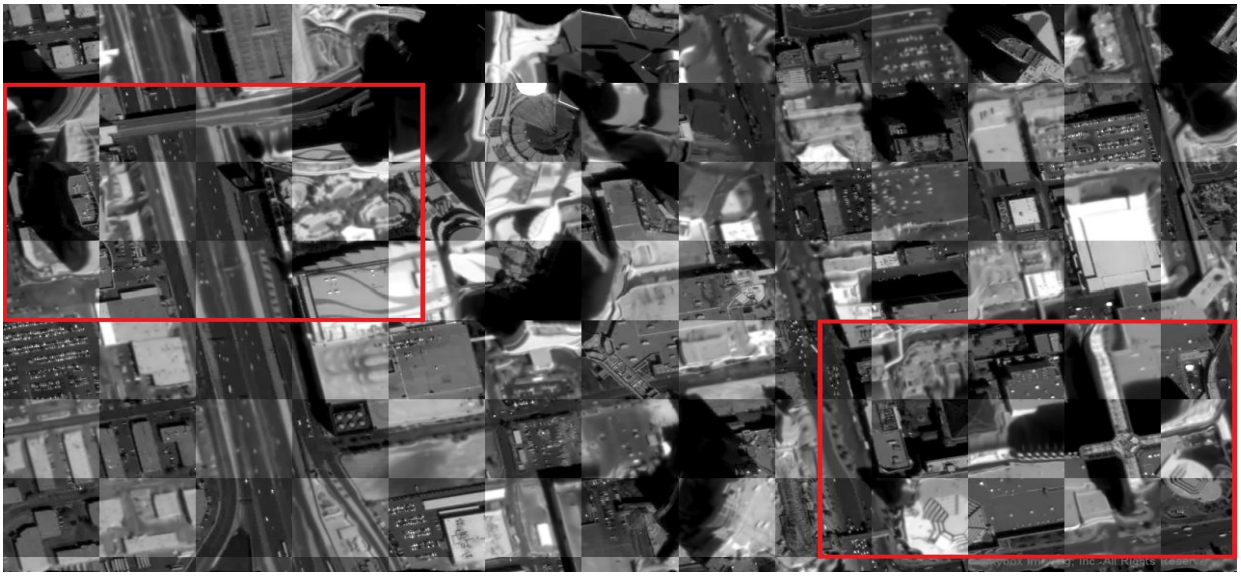
		Mean Displacement Errors *(in pixels)*		
Dataset	**Method**	**DX**	**DY**	**DS**
Burj Khalifa	Unregistered data	0.7	3.2	3.3
	Descriptor-based	0.4	3.8	3.9
	MRF-based	**0.4**	**0.8**	**0.8**
Las Vegas	Unregistered data	1.2	6.4	6.5
	Descriptor-based	1.0	8.8	8.9
	MRF-based	**0.7**	**1.3**	**1.5**
Las Vegas Night	Unregistered data	1.7	13.7	13.8
	Descriptor-based	3.0	13.3	13.7
	MRF-based	**0.8**	**0.8**	**1.1**

Table 2: Quantitative evaluation results after the application of the proposed MRF-based registration method. In all cases the developed approach managed to registered the satellite video frames with a mean displacement error of less than 1.5 pixels.

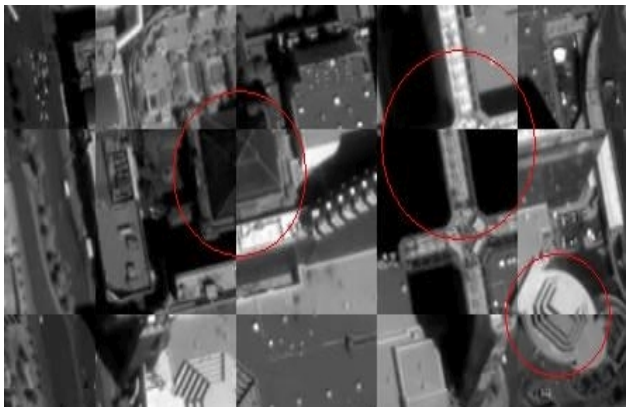
[4] https://www.youtube.com/watch?v=aW1-ZWencvA
[5] https://www.youtube.com/watch?v=lKNAY5ELUZY
[6] https://www.youtube.com/watch?v=uw7CSkJKJYw



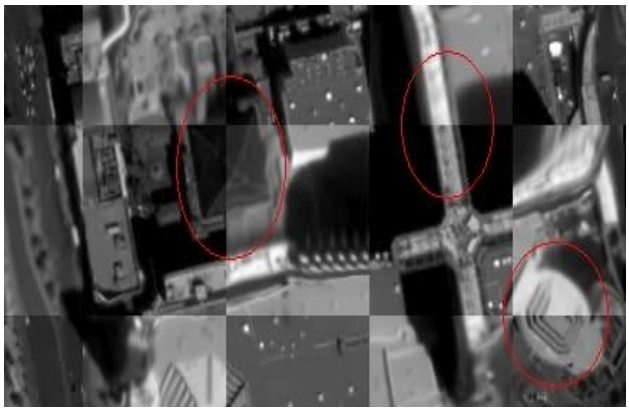
(a) Blending frames from the unregistered *Las Vegas* video dataset



(b) Blending frames from the registered *Las Vegas* video dataset

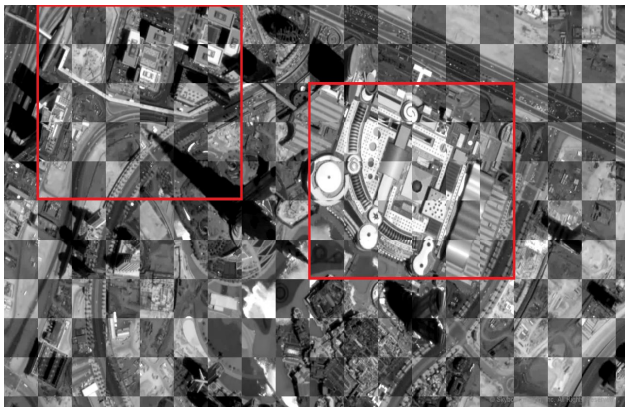


(c) Unregistered (zoom-in area)

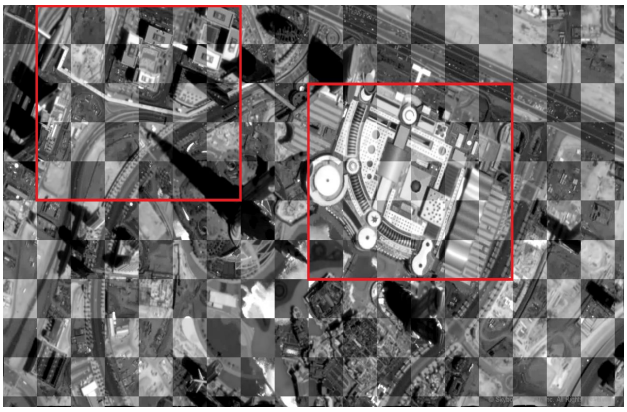


(d) Registered (zoom-in area)

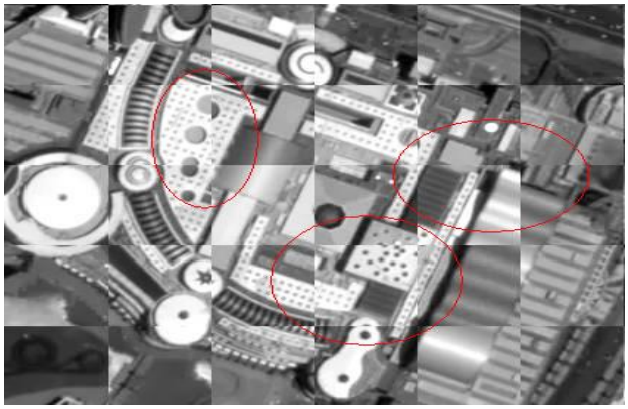
Figure 2: Chessboard visualizations from the *Las Vegas* Skybox dataset. Frames from the unregistered dataset (a) and frames after the registration process (b) are shown in the first two rows. Zoom-in areas are shown in the third row for the unregistered (c) and registered (d) frames.



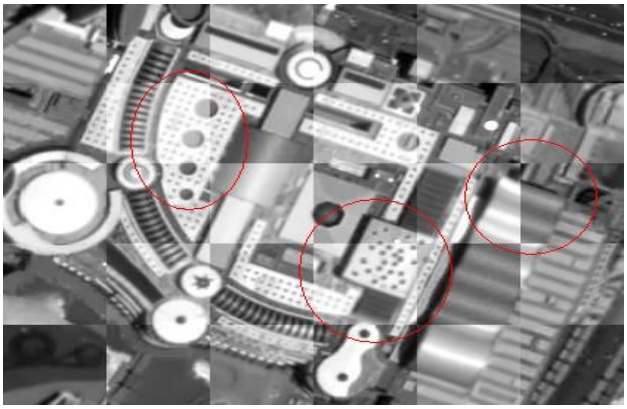
(a) Frames from the unregistered *Burj Khalifa* video dataset



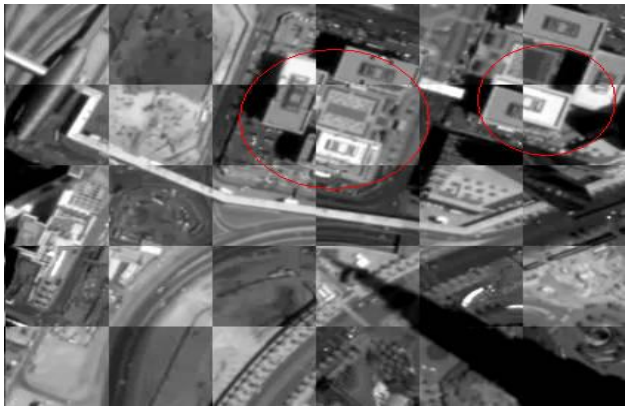
(b) Frames from the registered *Burj Khalifa* video dataset



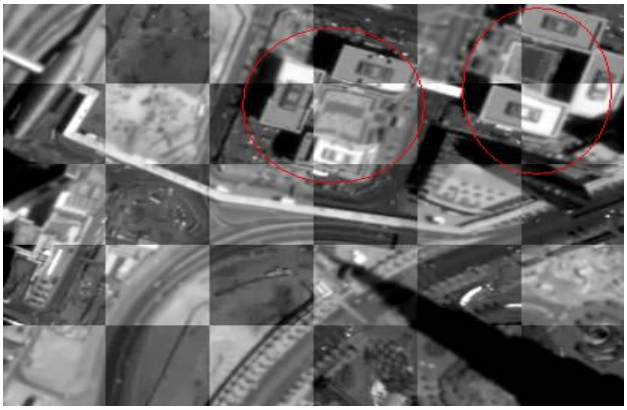
(c) Frames from the unregistered data (zoom-in area #1)



(d) Frames from the registered data (zoom-in area #1)



(e) Frames from the unregistered data (zoom-in area #2)



(f) Frames from the registered data (zoom-in area #2)

Figure 3: Chessboard visualization from the *Burj Khalifa* video dataset. Unregistered (left) and registered (right) data before and after the application of the proposed methodology.

As expected the image regions with the most mis-registration errors were those with significant relief displacements with very tall man-made objects, buildings and skyscrapers.

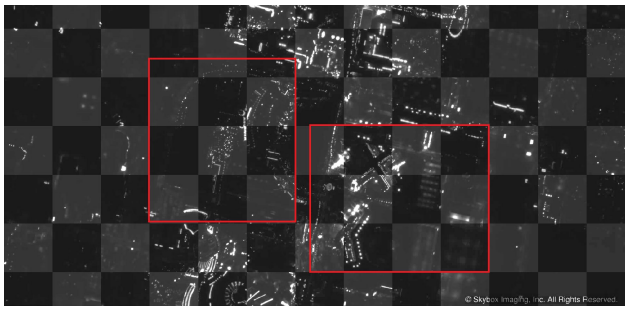
With a chessboard visualisation, results for two other datasets are presented in Figure 3 and 4. Once again after a closer look one can observe the robustness of the proposed approach towards recovering scene's (frame's) geometry. Moreover, Figures 2 and 4 depict the same region in different acquisition times. Even though in Figure 4, the satellite video dataset was acquired during the night, the proposed MRF-based method performed significantly well, resulting into an overall mean displacement error of less than one pixel in both axis (Table 2).

In order to qualitatively compare the results of the proposed MRF-based approach with the descriptor-based one, results on the same

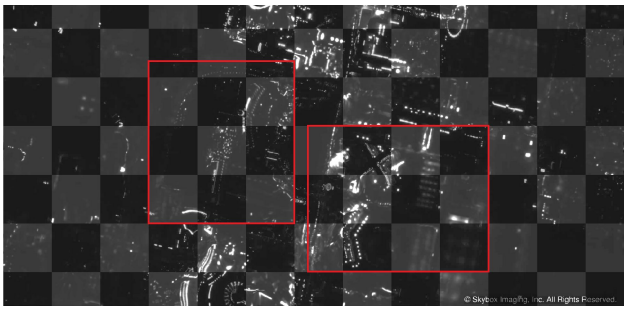
Registration of *Burj Khalifa* to Google Earth
Mean Displacement Errors

	DX	DY	DS
Unregistered	33.2	25.2	41.7
Registered	1.1	1.1	1.6

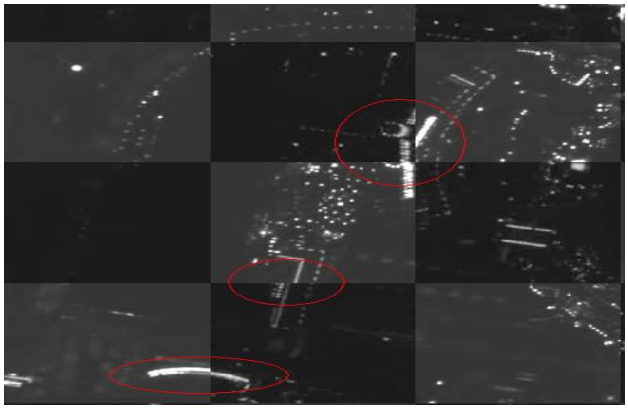
Table 3: Quantitative evaluation results after the registration of the *Burj Khalifa* satellite video dataset to an image mosaic acquired from Google Earth.



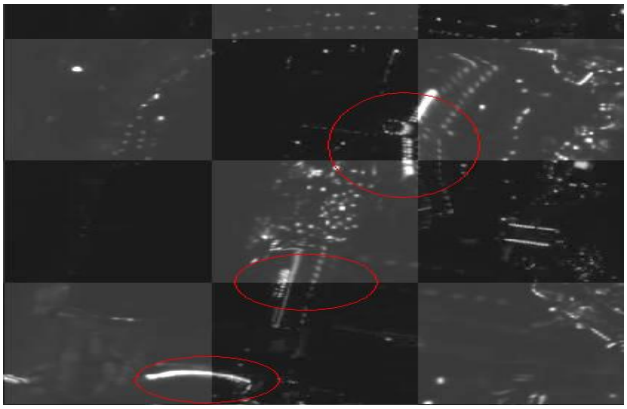
(a) Frames from the unregistered *Las Vegas-night* video dataset



(b) Frames from the registered *Las Vegas-night* video dataset

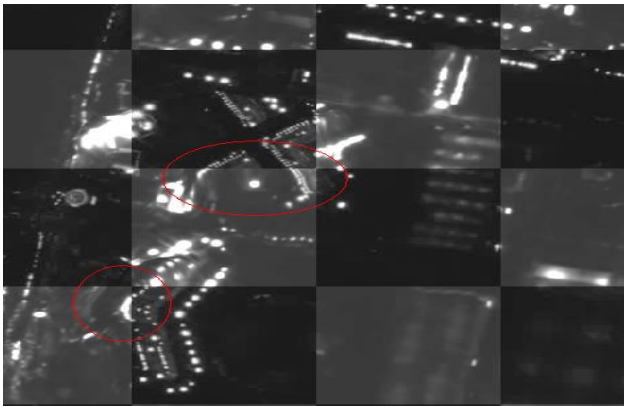


(c) Frames from the unregistered data (zoom-in area #1)



(d) Frames from the registered data (zoom-in area #1)

(e) Frames from the unregistered data (zoom-in area #2)



(f) Frames from the registered data (zoom-in area #2)

Figure 4: Chessboard visualization from the *Las Vegas-night* video dataset. Unregistered (left) and registered (right) data before and after the application of the proposed methodology.

datasets are presented in Figure 5 after the application of the descriptor-based method. Although a large number of correspondences have been established the rigid nature of the transformation could not recover scene's geometry adequately.

5. CONCLUSION

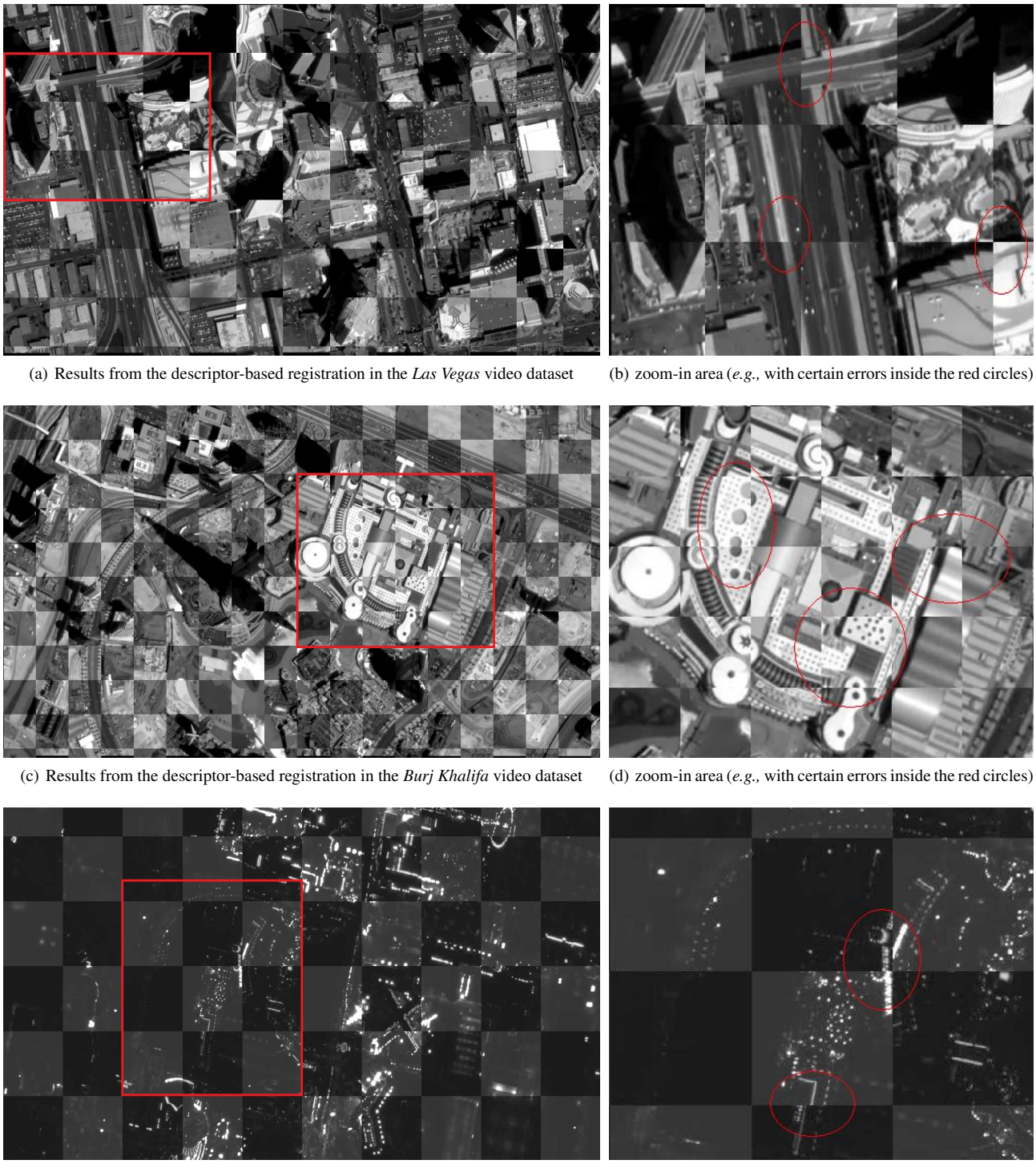
In this paper an MRF-based registration approach was developed for the accurate co-registration of satellite video frames as well as the registration of the video dataset to reference map/image. The method was applied and validated based on satellite video data from Skybox Imaging and compared with a standard descriptor-based registration framework. Experimental results indicate the great potentials of the proposed approach which managed to recover the geometry in all cases with registration errors of less than 1.5 pixels at both x and y axis.

REFERENCES

Agrawal, M., Konolige, K. and Blas, M., 2008. Censure: Center surround extremas for realtime feature detection and matching. In: D. Forsyth, P. Torr and A. Zisserman (eds), Computer Vision ECCV 2008, Lecture Notes in Computer Science, Vol. 5305, Springer Berlin Heidelberg, pp. 102–115.

Alahi, A., Ortiz, R. and Vandergheynst, P., 2012. Freak: Fast retina keypoint. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 510–517.

Bay, H., Ess, A., Tuytelaars, T. and Gool, L. V., 2008. Speeded-up robust features (surf). Computer Vision and Image Understanding 110(3), pp. 346 – 359. Similarity Matching in Computer Vision and Multimedia.



(a) Results from the descriptor-based registration in the *Las Vegas* video dataset

(b) zoom-in area (*e.g.,* with certain errors inside the red circles)

(c) Results from the descriptor-based registration in the *Burj Khalifa* video dataset

(d) zoom-in area (*e.g.,* with certain errors inside the red circles)

(e) Results from the descriptor-based registration in the *Las Vegas-night* video dataset

(f) zoom-in area (*e.g.,* with certain errors inside the red circles)

Figure 5: Registration results after the application of the *descriptor-based* approach.

Culjak, I., Abram, D., Pribanic, T., Dzapo, H. and Cifrek, M., 2012. A brief introduction to opencv. In: MIPRO, 2012 Proceedings of the 35th International Convention, pp. 1725–1730.

d'Angelo, P., Kuschk, G. and Reinartz, P., 2014. Evaluation of skybox video and still image products. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-1, pp. 95–99.

Fischler, M. A. and Bolles, R. C., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24(6), pp. 381–395.

Glocker, B., Sotiras, A., Komodakis, N. and Paragios, N., 2011. Deformable Medical Image Registration: Setting the State of the Art with Discrete Methods. Annual Review of Biomedical Engineering 13, pp. 219–244.

Karantzalos, K., Sotiras, A. and Paragios, N., 2014. Efficient and

automated multi-modal satellite data registration through mrfs and linear programming. IEEE Computer Vision and Pattern Recognition Workshops.

Kopsiaftis, G. and Karantzalos, K., 2015. Vehicle detection and traffic density monitoring from very high resolution satellite video data. In: Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International, pp. 1881–1884.

Le Moigne, J., Netanyahu, N. S. and Eastman, R. D., 2011. Image Registration for Remote Sensing. Cambridge University Press.

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

Morel, J.-M. and Yu, G., 2009. Asift: A new framework for fully affine invariant image comparison. SIAM J. Img. Sci. 2(2), pp. 438–469.

Murthy, K., Shearn, M., Smiley, B. D., Chau, A. H., Levine, J. and Robinson, D., 2014. Skysat-1: very high-resolution imagery from a small satellite. In: Proc. SPIE, Vol. 9241, pp. 92411E–92411E–12.

Price, S., 2015. Rectifying the planet. In: FOSS4G Free and Open Source Software for Geospatial.

Sotiras, A., Davatzikos, C. and Paragios, N., 2013. Deformable medical image registration: A survey. Medical Imaging, IEEE Transactions on 32(7), pp. 1153–1190.

Tola, E., Lepetit, V. and Fua, P., 2010. DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(5), pp. 815–830.

Vakalopoulou, M. and Karantzalos, K., 2014. Automatic descriptor-based co-registration of frame hyperspectral data. Remote Sensing 6(4), pp. 3409–3426.

Zitova, B. and Flusser, J., 2003. Image registration methods: a survey. Image and Vision Computing 21(11), pp. 977 – 1000.