Homography estimation with deep convolutional neural networks by random color transformations

Miguel A. Molina-Cabello[1]
miguelangel@lcc.uma.es
David A. Elizondo[2]
elizondo@dmu.ac.uk
Rafael Marcos Luque-Baena[1]
rmluque@lcc.uma.es
Ezequiel López-Rubio[1]
ezeqlr@lcc.uma.es

[1] Department of Computer Languages and Computer Science
University of Málaga
Málaga, Spain

[2] Department of Computer Technology
De Montfort University
Leicester, United Kingdom

Abstract

Most classic approaches to homography estimation are based on the filtering of outliers by means of the RANSAC method. New proposals include deep convolutional neural networks. Here a new method for homography estimation is presented, which supplies a deep neural homography estimator with color perturbated versions of the original image pair. The obtained outputs are combined in order to obtain a more robust estimation of the underlying homography. Experimental results are shown, which demonstrate the adequate performance of our approach, both in quantitative and qualitative terms.

1 Introduction

One of the fundamental low level tasks to be accomplished in computer vision is that of homography estimation between two images. This task consists in finding a non-singular, linear homography transformation between the points in both images [5, 6]. This is required for many higher level tasks of computer vision such as line matching [15], mosaicing [3], motion detection [18], camera motion estimation [19], tracking from multiple views [7, 8, 9, 10] and action recognition [17].

Classic approaches to homography estimation are mostly based on the Random Sampling and Consensus (RANSAC) technique, where a filtering of pairs of key points coming from both images is carried out, in order to remove erroneous point pairs [1, 2, 14]. Once the bad point pairs have been identified, the estimation of the homography can proceed more accurately. Outlier rejection is therefore the main strategy in which classic approaches are founded [13, 16].

This state of things has changed substantially with the advent of deep learning convolutional neural networks. Deep homography estimation networks have been proposed, which

are able to estimate the homography directly from the input image pair, without explicitly computing sets or pairs of key points. Among these proposals, the model proposed by Nguyen et al. [12] stands as one of the state of the art deep networks for this purpose.

In this work, we aim to enhance the performance of deep homography estimation networks by means of an ensemble approach. This strategy, which has been successfully employed previously [11], involves the presentation of multiple versions of the input to the network, followed by a combination of the outputs that the network yields for the various inputs. The alternative inputs are slightly perturbated versions of the original input. This way, the combination of the produced outputs is expected to be a more robust estimation of the true homography than any of the individual outputs. The kind of perturbations that are considered here are random color transformations, as they generate a significant amount of variability in the set of outputs, while maintaining their quality.

The structure of this paper is as follows. First, our proposed method is described in Section 2. Then the experiments that we have carried out are reported in Section 3. Finally, Section 4 is devoted to conclusions.

2 Methodology

In this section, our homography estimation proposal is detailed. We have observed that the output of the deep homography estimator varies substantially as the input pair of images is subject to color transformations. We propose to employ this effect to the advantage of the estimation process, by combining the results of the estimator as the input pair is transformed by a set of random color transformations.

Let $\mathcal{F}$ be the deep homography estimator, which takes two images of size $M \times N$ pixels as input, and outputs a 8-component vector with the coordinates of the 4 corners of the homography:

$$\mathcal{F} \,:\, [0,255]^{3MN} \times [0,255]^{3MN} \rightarrow ([1,M] \times [1,N])^4 \tag{1}$$

$$\mathcal{F}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{h} \tag{2}$$

where color values in the range $[0,255]$ are assumed. Please note that tristimulus color values are considered here, so that the images have the following structure:

$$\mathbf{X} = (x_{q,r,s})_{q \in \{1,2,3\}, r \in \{1,\ldots,M\}, s \in \{1,\ldots,N\}} \tag{3}$$

Let φ be a color transformation:

$$\varphi \,:\, [0,255]^{3MN} \rightarrow [0,255]^{3MN} \tag{4}$$

Then we propose to estimate the homography as follows:

$$\hat{\mathbf{h}} = \psi(\{\mathbf{h}_i \mid i \in \{1,\ldots,H\}\}) \tag{5}$$

where $\psi \in \{mean, median\}$ is a suitable aggregation function, and H is the number of homographies to be aggregated. Each homography $\mathbf{h}_i$ is obtained by supplying the deep homography estimator with a pair of color transformed images:

$$\mathbf{h}_i = \mathcal{F}(\varphi_{i,1}(\mathbf{X}_1), \varphi_{i,2}(\mathbf{X}_2)) \tag{6}$$

where the color transformations $\varphi_{i,j}$ with $j \in \{1,2\}$ are randomly generated as follows. The color transformations are composed of four stages which are applied in sequence: 1) gamma transformation, 2) brightness transformation, 3) tone transformation, 4) color clipping. Therefore we have:

$$\varphi_{i,j}(\mathbf{X}) = \gamma_{i,j}(\beta_{i,j}(\tau_{i,j}(\kappa_{i,j}(\mathbf{X})))) \tag{7}$$

The gamma transformation is given by:

$$\gamma_{i,j}(\mathbf{X}) = ((x_{q,r,s})^a)_{q\in\{1,2,3\},r\in\{1,\ldots,M\},s\in\{1,\ldots,N\}} \tag{8}$$

where a is a real number drawn from the uniform distribution on the interval $[A_1, A_2]$.

The brightness transformation is as follows:

$$\beta_{i,j}(\mathbf{X}) = (bx_{q,r,s})_{q\in\{1,2,3\},r\in\{1,\ldots,M\},s\in\{1,\ldots,N\}} \tag{9}$$

where b is a real number drawn from the uniform distribution on the interval $[B_1, B_2]$.

Thirdly, the tone transformation is defined as:

$$\tau_{i,j}(\mathbf{X}) = (t_{q,r,s})_{q\in\{1,2,3\},r\in\{1,\ldots,M\},s\in\{1,\ldots,N\}} \tag{10}$$

$$t_{q,r,s} = c_q x_{q,r,s} \tag{11}$$

where c_1, c_2 and $c_3$are three real numbers drawn from the uniform distribution on the interval $[C_1, C_2]$.

Finally, the color clipping is carried out this way:

$$\kappa_{i,j}(\mathbf{X}) = (k_{q,r,s})_{q\in\{1,2,3\},r\in\{1,\ldots,M\},s\in\{1,\ldots,N\}} \tag{12}$$

$$k_{q,r,s} = \min(\max(x_{q,r,s}, 0), 255) \tag{13}$$

The work presented by Nguyen *et al.* [12] used this kind of color shift process.

By combining a large enough number H of homographies obtained from random variations of the original input image pair, the combined estimated homography is expected to be closer to the real homography than most of the combined homographies. Hence the reliability of the homography estimation process is expected to be better than that of a single application of the deep homography estimator.

Figure 1 summarizes the operation of the proposal. First of all, we have a pair of images (I^A and I^B.) as input of the system. Four corners from one of these images (I^A) are also given to be predicted in the other image (the corners are not included in the figure in order to show it in a clearer way). Both images (and the four points) are the inputs of the base homography method, which produces its prediction ($\mathbf{h}_1$). Then, for each one of the additional considered base methods to be used in the consensus method, a color shift process is applied to both images. This process consists of an injection of random color, brightness and gamma shifts. After that, each pair of color shifted images are provided as inputs to the base method, which produces its homography prediction ($\mathbf{h}_2..\mathbf{h}_H$). In the last step, all the produced predictions are considered as inputs of the consensus function, that decides which is the predicted result ($\hat{\mathbf{h}}$). In the figure is also reported the output image formed by the input pair of images, the selected four corners in one of the images (left image, in red, they have been connected

to provide a clearer visualization), the ideal related four corners in the other image (right image, in red) and the predicted related four corners (right image, in yellow). Note that this methodology with a value of $H = 1$ produces the same result than the use of a single deep homography estimator.

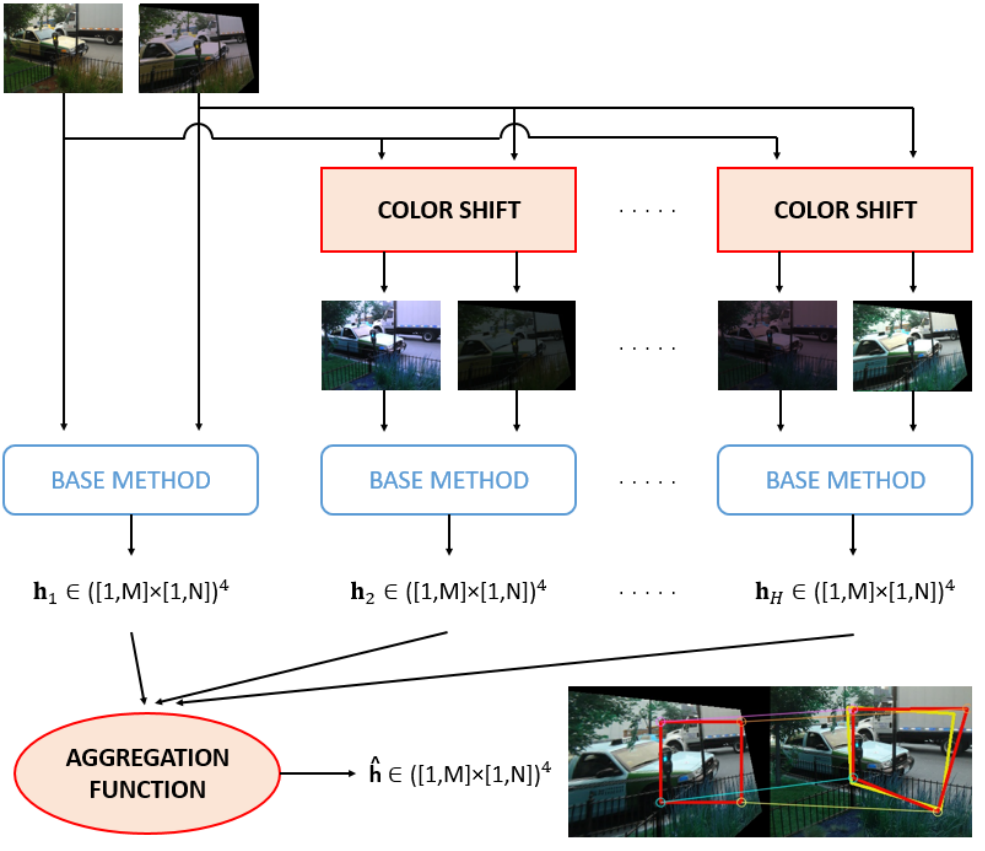


Figure 1: Graphical depiction of the operation of the proposed methodology. Given a pair of images (and four corners from one of these images; the corners are not included in the figure in order to show it in a clearer way) the method predicts the homography between both images by using a number H of homographies and an aggregation function. Note the different color transformation applied to each deep homography estimator.

3 Experimental results

The experiments that have been carried out are described in this section. First of all, Section 3.1 describes the software and the hardware resources employed in the experiments. Then, the image dataset used to test our proposal is depicted in Section 3.2. After that, the parameter selection of our approach is specified in Section 3.3. Finally, the obtained results are reported in Section 3.4.

3.1 Methods

In order to test the performance of the proposed model, we have selected a method from the literature, noted as *base method*. The chosen method is the homography estimation model proposed by Nguyen *et al*. [17], which predicts the homography between two images. It is written in python and it uses the OpenCV and tensorflow libraries. The code of this method can be download from its website [1].

Our proposed method is also implemented in python by employing the same libraries previously commented.

The reported experiments have been carried out on a 64-bit Personal Computer with an eight-core Intel i7 3.60 GHz CPU, 32 GB RAM and a NVIDIA GeForce GTX 1080 Ti.

3.2 Dataset

The images that we have used in our experiments are created by a synthetic data generation process on the COCO dataset similar to the used in [9].

The COCO dataset is a public set with the purpose of a large-scale object detection, segmentation, and captioning[2]. The version that we have used is the 2014 Test which is formed by 40,775 images. It can be downloaded from its website[3].

Our testing dataset is formed by 1000 pair of images. Each pair is composed by two images: an image selected from the COCO dataset (we have chosen the first 1000 images from the test set) and a synthetic image obtained by a randomly transformation of the selected image. This transformation applied consists of a random color, brightness and gamma shifts which are injected. This is done to test the performance under adversal conditions such as large image displacement and illumination varition. In addition, a point perturbation parameter ρ controls the amount of image overlap.

3.3 Parameter selection

We have selected a wide range of values for the parameters of the proposed consensus method. The function that decides how the output of the consensus is produced is analyzed. In this case, as presented in Section 2, we have considered two possible aggregation (or consensus) functions: the mean and the median. These two functions have been chosen because they attempt to describe a set of data by determining the central position within that set with a single value, and their implementation is easy to be done.

Another interesting parameter is the number of base method used in the proposal. The parameters related to the color shift and the point perturbation are the same than the used in [17]. Table 1 shows the selected parameter value configurations.

3.4 Results

A comparison between the selected homography estimator and the proposed one is carried out in this section. First of all, the results of the experiments are depicted from a qualitative point of view.

[1] https://github.com/tynguyen/unsupervisedDeepHomographyRAL2018
[2] http://cocodataset.org/
[3] http://images.cocodataset.org/zips/test2014.zip

Parameter		Values
Number of homographies, H	=	$\{1..10\}$
Aggregation function, ψ	=	{mean, median}
Gamma transformation interval, $[A_1, A_2]$	=	$[0.8, 1.2]$
Brightness transformation interval, $[B_1, B_2]$	=	$[0.5, 2.0]$
Tone transformation interval, $[C_1, C_2]$	=	$[0.8, 1.2]$
Point perturbation, ρ	=	45

Table 1: Considered parameter values for the proposed method, forming the set of experimental configurations.

A result of the operation of the proposed approach is represented in Figure 2. Given an input pair of images (a), the first step is where the proposal calculates the homography between the pair of images by employing a selected homography estimation method (b, first row). In addition, the proposal applies a color shift to both input images and it uses the homography method to estimate the homography between the images (b, second, third and fourth rows). The total amount of considered homographies is H (in this case, $H = 10$). Each one of these 10 images is formed by two subimages: the left subimage shows the points (and the connection between them) to be predicted in the right subimage, which exhibits the predicted homography result by the base method and the ideal result (ground truth). After that, the proposal generates the homography result by using an aggregation function (c), where the homography result predicted by the base method, the Mean and Median Consensus, and the ideal result are shown.

Several results are presented in a visual way in Figure 3. In some cases, the prediction of the base method looks like to be closer to the ground truth (first row). So that, the base method yields a better prediction. Nevertheless, the result produced by the consensus proposals is better than the result provided by the base method in the most of cases. Furthermore, the Mean Consensus seems to works better (second and third row) than the Median Consensus (forth row). According to the result presented in 2, the different predicted homographies are close to the ideal result, but none matches exactly. In this context, the Median provides one of the predicted homographies, while the Mean constructs a new rounded prediction from the predicted homographies.

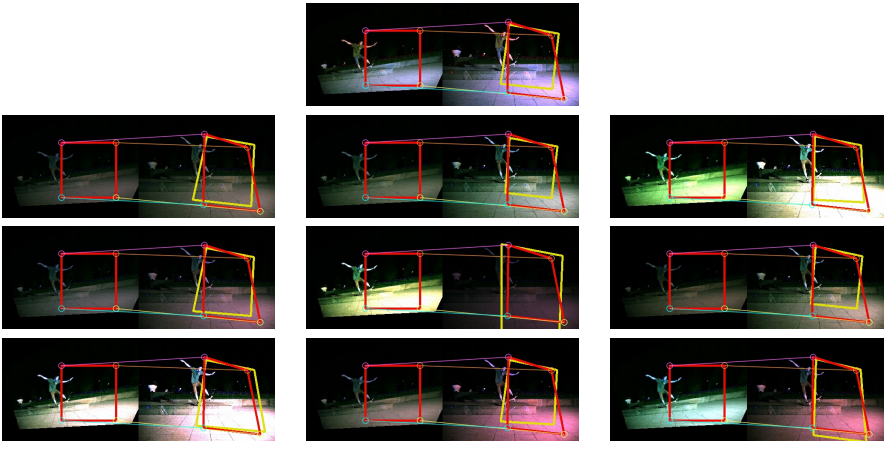
Additionally, a comparison of the performance of the considered methods is described from a quantitative point of view. Given a pair of images and four points which belong to one of them, an homography method predicts the related four points in the other image. So that, a vector $\mathbf{h} \in \mathbb{R}^8$ is predicted (two coordinates per point). Note that each pair of images provides a ground truth mask which is composed by a vector $\mathbf{g} \in \mathbb{R}^8$ that is the ideal result. Thus, a comparison between the result of an homography method and the ground truth can be accomplished in order to obtain a quantitative performance.

Several well known measures have been selected in order to test the performance of the method. The first measure is the error between the predicted and the ideal result of each pair of images. The selected error (E) is those related to the Euclidean norm, where $E \in \mathbb{R}$, $E \geq 0$ and lower is better. Here, the error is defined as:

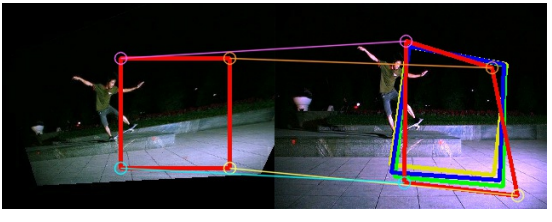
$$E = \|\mathbf{h} - \mathbf{g}\| = \sqrt{(h_1 - g_1)^2 + ... + (h_8 - g_8)^2} \tag{14}$$

where $\mathbf{h}, \mathbf{g} \in \mathbb{R}^8$.

(a) Input pair of images.



(b) Internal operation.



(c) The final result provided by the base method and the mean and median consensus.

Figure 2: Visual operation of the proposal. The input pair of frames are shown in (a), while (b) exhibits how the proposal works internally in order to obtain the homography between the pair of images of (a). In (b), there are H images (in this case, the number of homographies is 10) and each image shows the pair of images, the homography that the method estimates and the ground truth. Note that all pairs of images (second, third and fourth row) present a color shift except the first pair (first row), which is the same pair than (a). In (c), the image represents the result provided by the proposal by combining the internal results through the aggregation function (mean and median consensus). The base method result and the ground truth are also shown in order to compare them in a visual way.

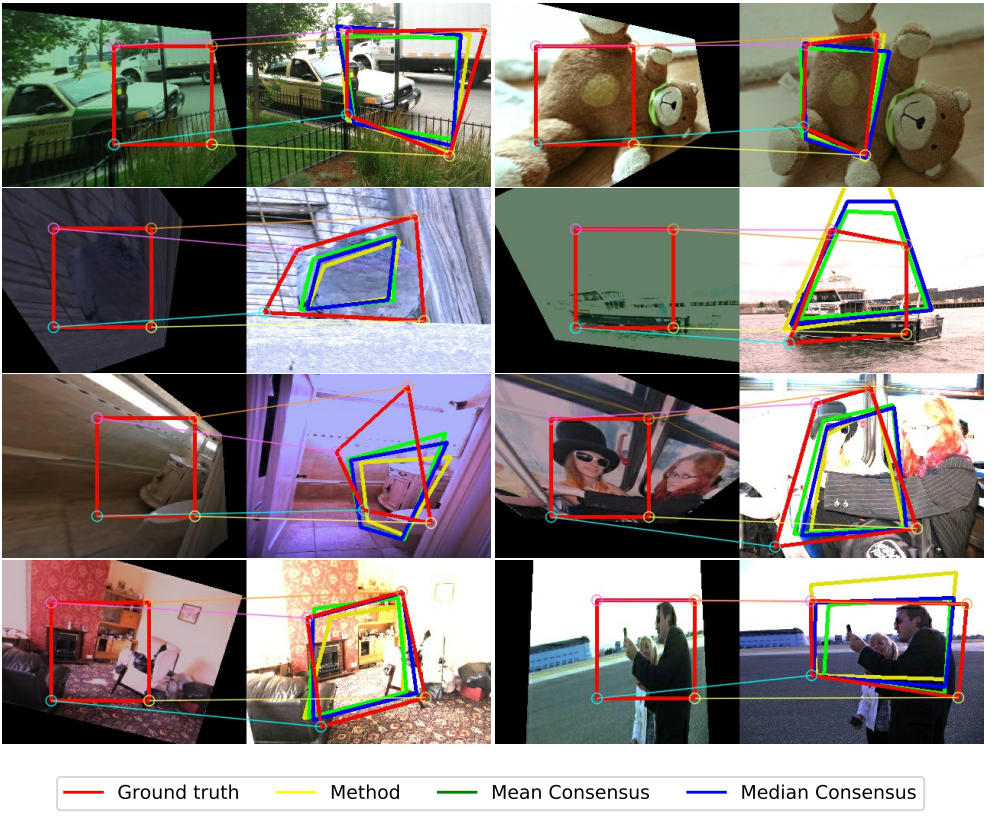


Figure 3: Qualitative results for some benchmark pair of images. From left to right and up to bottom, pair of images generated from the COCO test images: 1, 155, 3180, 5798, 6242, 6504, 8447 and 12320. For each pair of images, the left image exhibits the four selected corners, while the right image shows the ideal result and the predict result provided by the tested methods.

The number of wins of a method against other as a measure was also considered. In this case, the number of wins is defined as follows. Let A, B two homography methods and $E_{A,i}, E_{B,i}$ the error related to their prediction result according to the i-th pair of images, $W_i = 1$ (A wins B) respecting the i-th pair of images if $E_{A,i} > E_{B,i}$. In other case, $W_i = 0$. Thus, the number of wins of the method A against the method B is

$$W = \sum_{i=1}^{K} W_i \tag{15}$$

where K is the total number of pair of images (in this case, $K = 1000$).

The error achieved by each considered proposal is shown in Table 3.4. As it can be observed, the Mean Consensus yields the best performance. It must be highlighted that the higher the value of H used in the consensus, the more the Mean Consensus improves its performance. This does not happen in the case of the Median Consensus, where the error slightly swings. Moreover, not only the mean error is lower, the standard deviation of the

H	Mean Consensus	Median Consensus
1 (base method)	63.139 ± 28.955	63.139 ± 28.955
2	62.944 ± 26.374	62.944 ± 26.374
3	61.793 ± 24.794	64.696 ± 27.122
4	60.850 ± 23.927	62.631 ± 26.088
5	60.164 ± 23.188	62.906 ± 26.429
6	59.922 ± 22.778	62.024 ± 25.920
7	59.892 ± 22.594	62.956 ± 26.327
8	59.635 ± 22.069	62.383 ± 25.615
9	59.673 ± 22.061	62.828 ± 25.708
10	59.580 ± 21.885	62.300 ± 25.314

Table 2: Error results yielded by the tested methods for the test dataset. First column specifies the parameter H, while the remaining columns report the mean error and the standard deviation obtained by the Mean Consensus and the Median Consensus, respectively.

error is also lower. Therefore, the consensus method is more robust than the use of a single base method.

The number of wins of the different considered consensus approaches against the base method is also studied. This information is reported in Figure 4. The left image depicts the number of wins of the base method versus the Mean Consensus, while the right image reports the number of wins of the base method versus the Median Consensus. With the same value of H, the Mean Consensus wins more times than the Median Consensus. In addition, the Mean Consensus improves its performance by increasing H; however, in the Median Consensus, the number of wins remains practically the same after a certain value of H.

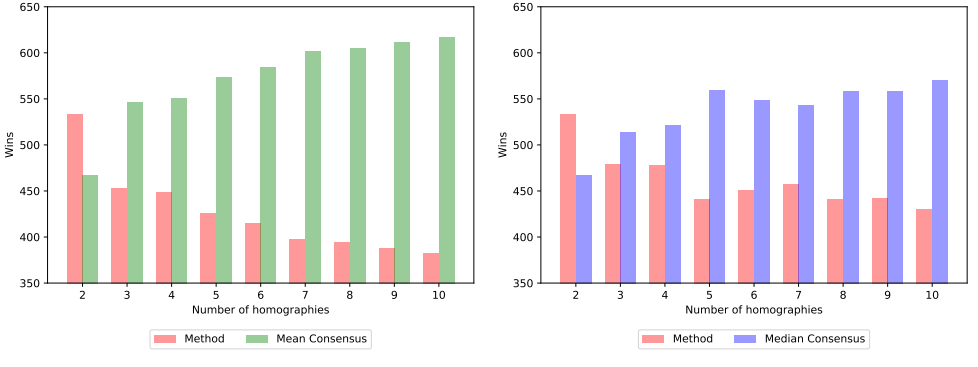


Figure 4: Number of wins of the consensus proposals against the base method. Images correspond to Mean Consensus and Median Consensus, respectively. Each image reports the number of wins according to the value of the parameter H.

4 Conclusions

A new homography estimation method has been proposed, which combines the outputs obtained from a deep neural network homography estimator for perturbated versions of the

original input pair of images. The outputs are combined by a suitable aggregation function, namely the mean or the median. The kind of perturbations which have been considered are color transformations of the input images. The obtained results demonstrate that the quality of the estimated homography consistently increases as the number of combined outputs grows. This demonstrates the stability and suitability of the proposed approach.

Acknowledgments

This work is partially supported by the Ministry of Economy and Competitiveness of Spain under grants TIN2016-75097-P and PPIT.UMA.B1.2017. It is also partially supported by the Ministry of Science, Innovation and Universities of Spain [grant number RTI2018-094645-B-I00], project name Automated detection with low cost hardware of unusual activities in video sequences. It is also partially supported by the Autonomous Government of Andalusia (Spain) under project P12-TIC-657, project name Self-organizing systems and robust estimators for video surveillance. It is also partially supported by the Autonomous Government of Andalusia (Spain) under project UMA18-FEDERJA-084, project name Anomalous behaviour agent detection by deep learning in low cost video surveillance intelligent systems. All of them include funds from the European Regional Development Fund (ERDF). The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Málaga. They also gratefully acknowledge the support of NVIDIA Corporation with the donation of two Titan X GPUs used for this research. The authors acknowledge the funding from the Universidad de Málaga.

References

[1] S. Choi, T. Kim, and W. Yu. Performance evaluation of RANSAC family. In *British Machine Vision Conference, BMVC 2009 - Proceedings*, 2009.

[2] O. Chum, J. Matas, and J. Kittler. Locally optimized RANSAC. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2781:236–243, 2003.

[3] D. Corrigan, K. Sooknanan, J. Doyle, C. Lordan, and A. Kokaram. A low-complexity mosaicing algorithm for stock assessment of seabed-burrowing species. *IEEE Journal of Oceanic Engineering*, 2018.

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016.

[5] Y. Huang, C. Wang, and C. Li. Translucent image recoloring through homography estimation. *Computer Graphics Forum*, 37(7):421–432, 2018.

[6] R. Juarez-Salazar and V.H. Diaz-Ramirez. Homography estimation by two PClines Hough transforms and a square-radial checkerboard pattern. *Applied Optics*, 57(12): 3316–3322, 2018.

[7] F. Kahl and R. Hartley. Multiple - view geometry under the linf-norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9):1603–1617, 2008.

[8] F. Kahl, S. Agarwal, M.K. Chandraker, D. Kriegman, and S. Belongie. Practical global optimization for multiview geometry. *International Journal of Computer Vision*, 79(3): 271–284, 2008.

[9] S. Liu, J. Chen, C.-H. Chang, and Y. Ai. A new accurate and fast homography computation algorithm for sports and traffic video analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2993–3006, 2018.

[10] W.-L. Lu, J.-A. Ting, J.J. Little, and K.P. Murphy. Learning to track and identify players from broadcast sports videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1704–1716, 2013.

[11] Miguel A Molina-Cabello, Rafael Marcos Luque-Baena, Ezequiel López-Rubio, and Karl Thurnhofer-Hemsi. Vehicle type detection by ensembles of convolutional neural networks operating on super resolved images. *Integrated Computer-Aided Engineering*, 25(4):321–333, 2018.

[12] Ty Nguyen, Steven W Chen, Shreyas S Shivakumar, Camillo Jose Taylor, and Vijay Kumar. Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robotics and Automation Letters*, 3(3):2346–2353, 2018.

[13] N. Qi, S. Zhang, L. Cao, X. Yang, C. Li, and C. He. Fast and robust homography estimation method with algebraic outlier rejection. *IET Image Processing*, 12(4):552–562, 2018.

[14] H.K. Sangappa and K.R. Ramakrishnan. A probabilistic analysis of a common RANSAC heuristic. *Machine Vision and Applications*, 30(1):71–89, 2019.

[15] Y. Shen, Y. Dai, and Z. Zhu. Efficient line matching with homography. *Measurement Science and Technology*, 29(3), 2018.

[16] H. Wang, T.-J. Chin, and D. Suter. Simultaneously fitting and segmenting multiple-structure data with outliers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1177–1192, 2012.

[17] H. Wang, D. Oneata, J. Verbeek, and C. Schmid. A robust and efficient video representation for action recognition. *International Journal of Computer Vision*, 119(3): 219–238, 2016.

[18] L. Zelnik-Manor and M. Irani. Multiview constraints on homographies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):214–223, 2002.

[19] H. Zhu, X. Wen, F. Zhang, X. Wang, and G. Wang. Homography estimation based on order-preserving constraint and similarity measurement. *IEEE Access*, 6:28680–28690, 2018.