Aggregation of Convolutional Neural Network Estimations of Homographies by Color Transformations of the Inputs

MIGUEL A. MOLINA-CABELLO\textsuperscript{1,2}, DAVID A. ELIZONDO\textsuperscript{3}, (Senior Member, IEEE), RAFAEL MARCOS LUQUE-BENA\textsuperscript{1,2}, AND EZEQUIEL LÓPEZ-RUBIO\textsuperscript{1,2}

\textsuperscript{1}Department of Computer Languages and Computer Science, University of Málaga, 29071 Malaga, Spain
\textsuperscript{2}Biomedic Research Institute of Málaga (IBIMA), 29010 Malaga, Spain
\textsuperscript{3}Department of Computer Technology, De Montfort University, Leicester LE1 9BH, U.K.

Corresponding author: Miguel A. Molina-Cabello (miguelangel@lcc.uma.es)

This work was supported in part by the Ministry of Economy and Competitiveness of Spain under Grant TIN2016-75097-P and Grant PPIT.UMA.B1.2017, in part by the Ministry of Science, Innovation and Universities of Spain, through the Project Automated Detection with Low Cost Hardware of Unusual Activities in Video Sequences, under Grant RTI2018-094645-B-I00, in part by the Autonomous Government of Andalusia, Spain, through the Project Detection of Anomalous Behavior Agents by the Deep Learning in Low Cost Video Surveillance Intelligent Systems, under Grant UMA18-FEDERJA-084, and in part by the European Regional Development Fund (ERDF).

ABSTRACT The standard approach to the estimation of homographies consists in the application of the RANSAC algorithm to a set of tentative matches. More recent strategies based on deep learning, namely convolutional architectures, have become available. In this work, a new algorithm for the estimation of homographies is developed. It is rooted in a convolutional neural network for homography estimation, which is provided with a range of versions of the input pair of pictures. Such versions are generated by perturbation of the color levels of the input images. Each generated pair of images yields a distinct estimation of the homography, and then the estimations are combined together to obtain a final, more robust estimation. Experiments have been designed and carried out to test the validity of our approach, including qualitative and quantitative performance measures. In particular, it is demonstrated that our approach consistently outperforms the baseline approach consisting of using the output of the homography estimation deep network for the original input pair of images.

INDEX TERMS Deep convolutional neural networks, homography estimation, color transformations.

I. INTRODUCTION

One of the basic low level processes to be performed in computer vision consists of estimating a homography between two pictures. This task involves finding the best non-unique linear transformation between the key points in both images, in order to reduce the correspondence error as much as possible [1], [2]. This is essential for plenty of high-level computer vision challenges such as mosaicing [3], line matching [4], foreground detection [5], camera movement [6], action recognition [7] and multi-view tracking [8]–[11].

Feature-based methods have been the most used to date for homography estimation [12], [13]. Most techniques mainly focus on locating points of interest in each image using detectors based on computer vision such as the Harris detector [14] or the SIFT method [15]. Subsequently, those non-valid correspondence pairs are filtered using the Random Sampling and Consensus (RANSAC) method, providing the best possible homography estimation [16]–[18]. These methods perform better than traditional or direct methods [19]. Still, they may be ineffective when the number of points of interest is insufficient (mainly due to poorly textured and homogeneous images) or when mismatches occur due to substantial changes in lighting or significant differences in the viewpoints in both images. Due to its complexity, other techniques for extracting points of interest have been developed [20] but, although they are more efficient, their performance is slightly lower.

There are different approaches, either from a supervised to an unsupervised point of view. In [21], DeTone et al. designed a deep convolutional network with supervised learning for homography estimation from two images,
called HomographyNet. In addition to the images, they use four points defined in both images as input for the training phase, instead of the standard $3 \times 3$ homography matrix [22]. Other approaches directly estimate the parameters of the homography matrix using deep convolutional models [23]. A modification of the DeTone et al. model is proposed in [24], by including hierarchical networks, in which the images are managed by independent deep networks and, like the classic iterative models, the estimation of the homography is progressively improved in several stages. In the same line, another similar model is proposed in [25], although lighter in complexity and number of parameters, whose novelty consists of the inclusion of a homography refinement module to improve the final homography matrix. This refinement is performed by minimizing the masked pixel-level photometric discrepancy between the warped image and the destination image using a gradient-descent algorithm. Compressed convolutional networks are also considered to robust homography estimation [26]. On the other hand, in [27], Siamese networks are applied together with spatial pyramid pooling modules for supervised homography estimation. To avoid or reduce overfitting, they propose invertibility constraints, taking into account that there are a couple of homographies from each pair of images that are inverse matrices. With regard to unsupervised approaches, it is worth highlighting the Nguyen et al. paper [28], where the previous supervised proposals are improved using an unsupervised deep model in which it is not necessary to provide the ground truth labels (i.e., the correspondence points for each pair of images in the training phase). This model is the primary reference for our proposal.

Therefore, our objective is to outperform the performance of deep homography estimation networks by applying an ensemble approach. This strategy, which has been successfully used in previous works [29], consists of presenting altered versions of the input to the network and combine all the outputs which the network produces, by applying a kind of aggregation function. These modified inputs are slightly disturbed versions of the original one. In this way, it is expected that the combination of the outputs produced by the deep model provides a more robust estimate of the homography than each individual one. Random color transformations are the type of variation of the input data considered in this study, because of the fact that they incorporate a significant amount of variability in the result set, while the image quality is maintained.

The paper is structured as follows. Firstly, the methodology of this approach is described in Section II. The experimental results are detailed in Section III, whereas conclusions are exposed in Section IV.

II. METHODOLOGY

This section is devoted to the presentation of our method for the estimation of homographies. The rationale behind our proposal is as follows. Deep homography estimators output significantly different estimations for color transformed versions of the same input pair. Since the true homography does not vary under such color transformations, we propose to take advantage of this effect by aggregating the outputs associated with several random color transformations in order to yield a single estimated homography.

Let us note $\mathcal{F}$ the deep learning based homography estimator. The estimator accepts two images ($I^A$ and $I^B$) with size $M \times N$ pixels as the input, while it returns a vector of 8 components as output, which represents the two coordinates of the four corners of the homography:

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

Here it must be noted that tristimulus color values are considered, which lie in the interval $[0, 255]$.

Since a tristimulus color space is assumed, it follows that the images have this arrangement:

$$x = (x_q, r, s)_{q \in \{1, 2, 3\}, r \in \{1, \ldots, M\}, s \in \{1, \ldots, N\}}$$

where $q$ is the color channel index, $r$ is the pixel row coordinate, and $s$ is the pixel column coordinate.

Now, let us note $\varphi$ a color transformation:

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

As explained before, our proposal is to obtain an aggregated estimation by combining the homography estimations associated to several color transformations:

$$\hat{h} = \psi (S)$$

where $\psi$ is an appropriate aggregation function for homographies, $S$ is the set of homographies to be aggregated, and $H$ is the number of elements of $S$. An individual homography $h_i$ is obtained as the output of the neural homography estimator with a pair of randomly transformed images:

$$h_i = \mathcal{F} (\varphi_{i,1} (X_1), \varphi_{i,2} (X_2))$$

where $\varphi_{i,j}$ stand for the color transformations, and $u \in \{1, 2\}$ are the pixel coordinate indices.

The procedure to randomly generate color transformations $\varphi_{i,u}$ is detailed next. Here we propose to consider color transformations which entail four phases to be executed sequentially. First, a gamma transformation $\gamma_{1,u}$ is applied. Then, a brightness transformation $\beta_{1,u}$ is done. Thirdly, a tone transformation $\tau_{1,u}$ is carried out. Finally, color clipping $\kappa_{1,u}$ is performed. Hence the overall color transformation is given by:

$$\varphi_{i,u} (X) = \kappa_{i,u} (\tau_{i,u} (\beta_{i,u} (\gamma_{i,u} (X))))$$

The gamma transformation is defined as follows:

$$\gamma_{1,u} (X) = \left( (x_q, r, s)^a \right)_{q \in \{1, 2, 3\}, r \in \{1, \ldots, M\}, s \in \{1, \ldots, N\}}$$

where $a$ is a random number coming from the uniform probability distribution on the interval of real numbers $[A_1, A_2]$. 
Secondly, the brightness transformation is given by:

\[ \beta_{i,u}(X) = (b x_{q,r,s})_{q \in \{1,2,3\}, r \in \{1,\ldots,M\}, s \in \{1,\ldots,N\}} \] (10)

where \( b \) is a random number coming from the uniform probability distribution on the interval of real numbers \([B_1, B_2]\).

The tone transformation is given by:

\[ \tau_{i,u}(X) = (t_{q,r,s})_{q \in \{1,2,3\}, r \in \{1,\ldots,M\}, s \in \{1,\ldots,N\}} \] (11)
\[ t_{q,r,s} = c_q x_{q,r,s} \] (12)

where \( c_1, c_2 \) and \( c_3 \) are three random numbers obtained from the uniform probability distribution on the interval of real numbers \([C_1, C_2]\).

Fourthly, the color clipping is computed as follows:

\[ \kappa_{i,u}(X) = (k_{q,r,s})_{q \in \{1,2,3\}, r \in \{1,\ldots,M\}, s \in \{1,\ldots,N\}} \] (13)
\[ k_{q,r,s} = \min(\max(x_{q,r,s}, 0), 255) \] (14)

The color transformation considered here was previously considered by Nguyen et al. [28].

The introduction of random color transformations that has been proposed above implies that many samples of estimations of the true homography can be drawn as needed. As the number \( H \) of sample homographies grows, it is expected that the aggregated homography which combines them all is closer to the true homography. In other words, the accuracy of the homography estimation procedure is expected to enhance as compared to the application of the deep homography estimator to the original input pair of images.

The operation of the proposed methodology is summarized in Figure 1. First, an input pair of images \((I^a, I^b)\) are given. Four points from one of these images \((I^a)\) are also provided, so that, the four corners to be predicted in the other image \((I^b)\). Please, note that these points are not illustrated in the figure in order to show it in a clearer visualization. This way, the input pair of images (and also the four corners) are the input of the proposed model. Then, \( H - 1 \) pair of randomly color transformed images are generated from the raw pair of images. So that, \( H \) pair of images are considered. The color shift process consists of a gamma, brightness and tone transformations and a color clipping step, in that order. After that, a homography estimation method used as a base is employed to predict the homography between a pair of images. Each pair of images is given to the estimator. Thus, \( H \) predictions are obtained where \( h_1 \) corresponds to the prediction of the raw pair of images and \( h_2, \ldots, h_H \) corresponds to the predictions of the random color transformed pair of images. Finally, all of these predictions are suitable to be the input of a selected aggregation function. The output of this function is the predicted homography result \((\hat{h})\) of the proposed methodology. Please, note that the use of a single homography in the proposed methodology (so that \( H = 1 \)) produces the same result than the obtained prediction offered by the homography estimation method.

A. AGGREGATION FUNCTIONS

Several kinds of aggregation functions have been considered. In order to define each aggregation function, let \( S \) be the set of homographies, as defined in equation (6). Let us remember that \( H \) is the number of homographies to be aggregated by the \( \psi \) aggregation function, as it was previously defined.

- **Mean**
  \[ \psi_{\text{Mean}} = \text{mean}(S) \] (15)

- **Median**
  \[ \psi_{\text{Median}} = \text{median}(S) \] (16)

- **Mean except the worst (Mean-1)**
  As it was previously defined, a homography \( h_j \) is composed by 4 points. Let \( \mu_j \) be the mean of the \( j \)-th points of the homographies \( S \), with \( j \in \{1, \ldots, 4\} \), so that \( \mu_j \) is a vector of dimension \( 2 \times 1 \):

  \[ \mu_j = \text{mean} \left\{ \{ q_{i,j} | i \in \{1, \ldots, H\} \} \right\} \] (17)
  \[ q_{i,j} = \left( \frac{h_{i,j-1}}{h_{i,j}} \right) \] (18)

  For each point \( q_{i,j} \), the Euclidean distance between \( \mu_j \) and that point is calculated.
Let $h_{c}$ be the homography with the furthest $j$-th point from $\mu_{j}$. After that, the Mean-1 aggregation function is defined as follows:

$$\psi_{Mean-1} = mean(S - \{h_{c}\})$$  \hspace{1cm} (19)

- Median except the worst (Median-1) The process is the same than Mean-1, except for the following equation:

$$\psi_{Median-1} = median(S - \{h_{c}\})$$  \hspace{1cm} (20)

- Geometric Mean (GMean)

$$\psi_{GMean} = \left( \prod_{i=1}^{H} h_{i,k} \right)^{\frac{1}{n}}$$  \hspace{1cm} (21)

- Geometric Median (GMedian)

The algorithm for the geometric median presented in [30] has been implemented.

- Mean and Geometric Mean, and then Mean (GMean + Mean)

$$\psi_{GMean+Mean} = mean(mean(S), Gmean(S))$$  \hspace{1cm} (22)

- Mahalanobis per point (MahaPoint)

Let $C_{j}$ be the covariance matrix of the $j$-th points, so that $C_{j}$ is a matrix of dimensions $2 \times 2$:

$$C_{j} = mean \left( \left\{ (q_{i,j} - \mu_{j})(q_{i,j} - \mu_{j})^{T} \mid i \in \{1, \ldots, H\} \right\} \right)$$  \hspace{1cm} (23)

where the operator $T$ is the transpose of a vector.

For each $j$-th point $q_{i,j}$, the Mahalanobis distance for that point is calculated:

$$distMaha(q_{i,j}, \mu_{j}) = \sqrt{(q_{i,j} - \mu_{j})^{T}C_{j}^{-1}(q_{i,j} - \mu_{j})}$$  \hspace{1cm} (24)

where $C_{j}^{-1}$ stands for the inverse of the matrix $C_{j}$.

Let $T$ be the set of the $H(p - 1)$ homographies with the furthest $j$-th points from their means $\mu_{j}$, where $p$ is a given real number with $p \in [0, 1]$, which represents the fraction of homographies that will be considered to calculate the result of the aggregation function. Therefore, $p$ is a percentile, while $1 - p$ is the fraction of the homographies that will be excluded.

After that, the MahaPoint aggregation function is defined as follows:

$$\psi_{MahaPoint} = mean(S - T)$$  \hspace{1cm} (25)

- Mahalanobis per homography (MahaHomo)

The procedure is similar to the presented in MahaPoint. While the MahaPoint aggregation function considers the points individually and discards the worst points according to their distance respecting to the mean $\mu_{j}$, MahaHomo considers homographies instead of points. So that, this aggregation function discards the worst homographies.

For each point $q_{i,j}$ with $i \in \{1, \ldots, H\}$ and $j \in \{1, \ldots, 4\}$, the Mahalanobis distance between the mean of the $j$-th points of the homographies $\mu_{j}$ and $q_{i,j}$ is calculated, following Eq. 24.

After that, for each homography $h_{j}$, the sum of distances $D_{i}$ between its points $h_{i,j}$ and their means $\mu_{j}$ is calculated. So that:

$$D_{i} = \sum_{j=1}^{4} DistMaha(q_{i,j}, \mu_{j})$$  \hspace{1cm} (26)

Let $T$ be the set of the $H(p - 1)$ furthest homographies (according to $D_{i}$) from the homography composed by $\mu_{j}$, where $p$ is a given real number where $p \in [0, 1]$ which represents the fraction of homographies that will be considered to calculate the result of the aggregation function. So that, $p$ is a percentile, while $1 - p$ is the fraction of the homographies that will be excluded.

After that, the MahaHomo aggregation function is defined as follows:

$$\psi_{MahaHomo} = mean(S - T)$$  \hspace{1cm} (27)

- Euclidean per point (EuclideanPoint)

This aggregation function is defined in the same way that MahaPoint. The difference respecting that aggregation function is the definition of the distance between two given points, where its definition is as follows:

$$distEuclidean(q_{i,j}, \mu_{j}) = \sqrt{(q_{i,j} - \mu_{j})^{2}}$$  \hspace{1cm} (28)

- Euclidean per homography (EuclideanHomo)

This aggregation function is defined in the same way that MahaHomo. The difference respecting that aggregation function is the definition of the distance between two given points, where it is defined with Eq. 28.

- Manhattan per point (ManhattanPoint)

This aggregation function is defined in the same way that MahaPoint. The difference respecting that aggregation function is the definition of the distance between two given points, where its definition is as follows:

$$distManhattan(q_{i,j}, \mu_{j}) = \sqrt{(q_{i,j} - \mu_{j})^{2}}$$  \hspace{1cm} (29)

- Manhattan per homography (ManhattanHomo)

This aggregation function is defined in the same way that MahaHomo. The difference respecting that aggregation function is the definition of the distance between two given points, where it is defined with Eq. 29.

Next, a brief discussion of the proposed aggregation functions is carried out. Many of them are based on the sample mean, which is a fast way to obtain an unbiased estimator of the distribution mean. Nevertheless, in the presence of outliers, the functions based on the sample median can yield more accurate results, at the expense of a slightly higher computational load. The introduction of the Mahalanobis distance in some of the functions compensates for the different scaling that may arise in the dimensions of the homographies. Finally, in several functions, the most extreme observations are removed prior to the computation of the final result, in order to filter out outliers.
TABLE 1. Tuned configurations used in the experiments.

| Parameter                  | Values         |
|----------------------------|----------------|
| Number of homographies, \(H\) | \{1, 30\}     |
| Aggregation function, \(\psi\) | See Section II-A |
| 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, \(\rho\) | 45            |
| Percentile, \(p\) | \{0.96, 0.92, 0.88, 0.84, 0.80\} |

The computational complexity of our approach is linear with respect to the number of color transformations \(H\), \(O(H)\), except for the aggregation functions that are based on the median, which have complexity \(O(H \log(H))\). Nevertheless, the computation of the median takes a negligible amount of the overall runtime, so in practice, the complexity is linear with respect to \(H\).

III. EXPERIMENTAL RESULTS

This section depicts the experiments that have been carried out and the obtained results. First, the software and hardware which have been employed are described in Subsection III-A. Then, Subsection III-B specifies the image dataset used. After that, the parameter selection of our proposed model is depicted in Subsection III-C. Finally, the results which have been obtained are reported in Subsection III-D.

A. METHODS

A well-known recent method from the state-of-the-art has been selected in order to test the approach. The homography estimation model chosen for that purpose is the method (noted as Base) proposed in [28]. Given two input images, this method predicts the homography between them. The code is written in python by using TensorFlow and OpenCV libraries. The code of this method can be downloaded from its website.

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

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 an NVIDIA GeForce GTX 1080 Ti.

B. DATASET

The COCO dataset [31] has been used in the experiments. This dataset is a public one with the purpose of large-scale object detection, segmentation, and captioning. The version that we have used is the 2014 Test, which is formed by 40,775 images. It can be downloaded from its website.

In order to test the performance of the homography estimation method competitors, a synthetic dataset has been created from the COCO dataset. The generation of this dataset is similar to the used in other articles from the state-of-the-art [21], [28], [32]–[34]. This way, 5,000 pair of images have been generated to compose the testing dataset where each pair of images is formed by two images: an image chosen from the COCO dataset (the first 5,000 images from the test set have been selected) and a synthetic image obtained by a random transformation of the chosen image. The applied transformation consists of a random color, brightness, and gamma shifts that are injected. Additionally, the amount of image overlap is also controlled by a point perturbation parameter \(\rho\). The objective of this generation image pairs process is to produce a dataset that covers a wide range of adverse conditions (illumination variation, large image displacement...).

C. PARAMETER SELECTION

A wide range of parameter values has been chosen in order to test the approach. As previously presented in Section II, 1

1https://github.com/tynguyen/unsupervisedDeepHomographyRAL2018
2http://cocodataset.org/
3http://images.cocodataset.org/zips/test2014.zip
several aggregation functions have been considered to compute the output of the ensemble. The number of CNNs which form the ensemble $H$ is also analyzed, while the color transformation parameters, as well as the point perturbation used to generate the synthetic dataset, are detailed in Table 1.

**D. RESULTS**

In order to test the performance of the proposed approach, a comparison has been established. The results of these experiments are shown from a qualitative and quantitative point of view.

1) **QUALITATIVE RESULTS**

Figure 2 exhibits an example of the operation of the proposal. First of all, a pair of two input images are presented (a), where the aim is to predict the homography between them, as it was detailed in Section II. Secondly, this pair of images are given to the homography estimator that returns the prediction of the homography between both images as output (b, first row).

After that, a random color transformation is applied to the pair of images. Then these transformed images are supplied to the homography estimation method, where a new prediction is acquired as output (b, second row). This process of random color transformation and prediction of the homography is repeated $H - 1$ times. Thus, $H$ predictions are obtained at the end of this process (in this case, $H = 4$). Each prediction output is composed of two subimages, where points (and their connection) to be predicted are shown in the left subimage while the ground truth and the predicted homographies are shown in the right subimage. Finally, all predicted homographies are combined by using an aggregation function, and a consensed homography is provided as output (c). In this case, the output of the *Base* and the *Mean* and the *Median* aggregation functions are shown in order to observe how different the output can be.

Figure 3 reports several qualitative results. In this figure, homographies produced by the *Mean* and *Median* consensus proposals are compared against the *Base* method and the ideal

---

**TABLE 2.** Performances of best methods over the considered dataset (5,000 pairs of images). Each column shows the mean and the standard deviation performance of each measure. Best results are highlighted in bold.

| Method             | EDPE   | RMSE   | MAE   |
|--------------------|--------|--------|-------|
| Base               | 119.137 ± 56.463 | 23.029 ± 10.570 | 18.976 ± 9.115 |
| Median             | 109.123 ± 40.742 | 21.067 ± 7.606  | 17.416 ± 6.619 |
| Mean               | 113.895 ± 48.002 | 22.116 ± 8.978  | 18.090 ± 7.751 |
| Mean-1             | 110.101 ± 42.291 | 21.267 ± 7.902  | 17.567 ± 6.864 |
| Median-1           | 114.732 ± 48.745 | 22.284 ± 9.117  | 18.221 ± 7.874 |
| GMean              | 109.098 ± 40.993 | 21.047 ± 7.642  | 17.399 ± 6.652 |
| Gmedian            | 113.519 ± 47.960 | 22.026 ± 8.964  | 18.105 ± 7.756 |
| GMean+Mean         | 108.859 ± 40.782 | 21.005 ± 7.606  | 17.370 ± 6.623 |
| MahaPoint          | 110.145 ± 41.941 | 21.272 ± 7.832  | 17.575 ± 6.806 |
| MahaHomo           | 110.193 ± 41.962 | 21.269 ± 7.833  | 17.581 ± 6.811 |
| EuclideanPoint     | 110.092 ± 42.289 | 21.265 ± 7.901  | 17.566 ± 6.864 |
| EuclideanHomo      | 110.166 ± 42.088 | 21.266 ± 7.857  | 17.578 ± 6.832 |
| ManhattanPoint     | 110.046 ± 42.252 | 21.256 ± 7.894  | 17.560 ± 6.859 |
| ManhattanHomo      | 110.150 ± 42.070 | 21.262 ± 7.853  | 17.575 ± 6.829 |

*FIGURE 3.** Qualitative results for some benchmark input pairs of images. For each pair, the left image shows the selected four points to be predicted, while the right image exhibits the ground truth (ideal result) and the predicted homography obtained by *Base, Mean* and *Median*. From left to right and up to bottom, the input pair of images generated from the COCO test images are 127, 162, 3536 and 4160.
result (ground truth) in a visual way. These Mean and Median aggregation functions have been chosen because they are the simplest ones to be implemented considered in this work.

Selected pair of images represent a wide range of situations where Base does not achieve a high performance while Mean and Median (especially Mean) overpass it. Usually, the predicted homographies do not exactly match with the ideal result. However, in general, the predictions produced by the consensus offer a better performance than the Base method. As it can be observed, the consensus homography predictions are close to the ground truth homography. In particular, it is shown how Mean yields better than Median, where its predictions are closer to the ideal results than Median.

2) QUANTITATIVE RESULTS
The performance of the selected methods has also been compared from a quantitative point of view. As was detailed in Section II, a homography composed of 4 points with 2 coordinates per point can be defined as an 8-dimension vector. This way, a quantitative performance can be obtained from the comparison of the predicted and the ideal homographies. This comparison has been established according to the performance indicated by several well known measures which have been chosen for that purpose. The selected measures are about the error between the predicted and the ideal result of each pair of images, by providing a real value from 0 onwards where lower is better. All of these measures are defined as follows:

- **Euclidean Distance Point Error (EDPE)**
  \[
  EDPE = \sum_{i=1}^{K/2} \sqrt{(\bar{h}_{2i-1} - \hat{h}_{2i-1})^2 + (\bar{h}_{2i} - \hat{h}_{2i})^2}
  \]

- **Root Mean Squared Error (RMSE)**
  \[
  RMSE = \sqrt{\frac{1}{K} \sum_{i=1}^{K} (\hat{h}_i - \bar{h}_i)^2}
  \]

- **Mean Absolute Error (MAE)**
  \[
  MAE = \frac{1}{K} \sum_{i=1}^{K} |(\hat{h}_i - \bar{h}_i)|
  \]

where \(\bar{h}, \hat{h} \in \mathbb{R}^8\) are the ground truth and predicted homographies, respectively and \(K\) is the dimension of the homography vectors (\(K = 8\)).

Table 2 shows the mean performance yielded by each approach over the dataset. Due to a large number of tuned configurations, only the best configuration is reported for several of the considered aggregation functions. The first key point is that all proposed approaches improve the performance of the Base method. As it can be observed, not only the mean errors are lower, the standard deviation of the errors are also lower. Thus, the proposed methodology by using an aggregation function is more trustworthy than the employment of a single base method. Moreover, GMean + Mean proposal achieves the best performance. Additionally, GMean and Mean are also good in terms of the yielded performance. It is interesting to see how approaches that use
the median function to compute their prediction are worse than those that use the mean function.

It is also interesting to analyze the behavior of the performance of the proposal according to the number of members that belong to the ensemble, i.e., the parameter $H$. Figure 4 reports the performance yielded by several selected approaches by considering the different tuned values for the parameter $H$. In order to show the behavior of the performance in a clearer way in the figure, this way, MahahaHomo, EuclideanPoint, EuclideanHomo, ManhattanPoint and ManhattanHomo approaches have not been included in that figure since they exhibit a similar behavior. Note that $H = 1$ represents the Base method. The shape of the figure is very similar for the considered measures. Additionally, for the best proposals, the higher the value of $H$, the better their performance. However, this effect is not present in the case of median approaches, where the error slightly swings. It must be highlighted that the improvement of the proposal when $H$ is increased is not as prominent as when $H > 20$.

**IV. CONCLUSION**

In this paper, a new ensemble-based strategy is considered to improve the homography estimation methods. This proposal combines the outputs of a deep convolutional network applied to homography estimation, whose inputs (pair of images) are perturbing after applying color function transformations. This combination is carried out by several aggregation functions. Among them, the geometric mean must be highlighted, whose results according to different metrics (EDPE, RMSE, and MAE) present a significant improvement with regard to a single execution of the neural network homography estimator. Additionally, as the number of combined outputs grows, the quality of the homography estimate increases consistently. This fact proves the viability and stability of the proposal presented.

**ACKNOWLEDGMENT**

The authors would like to thank the computer resources, technical expertise, and assistance provided by the Supercomputing and Bioinformatics (SCBI) Center, University of Malaga. They would also like to thank the NVIDIA Corporation with the donation of two Titan X GPUs used for this research. They also acknowledge the grant of the Universidad de Malaga.

**REFERENCES**

[1] Y. Huang, C. Wang, and C. Li, “Translucent image recoloring through homography estimation,” Comput. Graph. Forum, vol. 37, no. 7, pp. 421–432, Oct. 2018.

[2] R. Juarez-Salazar and V. H. Diaz-Ramirez, “Homography estimation by two PClines Hough transforms and a square-radial checkerboard pattern,” Appl. Opt., vol. 57, no. 12, pp. 3316–3322, Apr. 2018.

[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 J. Ocean. Eng., vol. 44, no. 2, pp. 386–400, Apr. 2019.

[4] Y. Shen, Y. Dui, and Z. Zhu, “Efficient line matching with homography,” Meas. Sci. Technol., vol. 29, no. 3, Mar. 2018, Art. no. 035001.

[5] L. Zeiniz-Manor and M. Irani, “Multiview constraints on homographies,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 2, pp. 214–223, 2002.

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

[7] H. Wang, D. Ofteat, J. Verbeeck, and C. Schmid, “A robust and efficient video representation for action recognition,” Int. J. Comput. Vis., vol. 119, no. 3, pp. 219–238, Sep. 2016.

[8] F. Kahl and R. Hartley, “Multiple-view geometry under the $\ell_\infty$-norm,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 9, pp. 1603–1617, Sep. 2008.

[9] F. Kahl, S. Agarwal, M. K. Chandraker, F. Kahl, S. Agarwal, M. K. Chandraker, D. Kriegman, and S. Belongie, “Practical global optimization for multiview geometry,” Int. J. Comput. Vis., vol. 79, no. 3, pp. 271–284, Sep. 2008.

[10] 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 Trans. Circuits Syst. Video Technol., vol. 28, no. 10, pp. 2993–3006, Oct. 2018.

[11] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy, “Learning to track and identify players from broadcast sports videos,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 7, pp. 1704–1716, Jul. 2013.

[12] 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 Process., vol. 12, no. 4, pp. 552–562, Apr. 2018.

[13] H. Wang, T.-J. Chin, and D. Suter, “Simultaneously fitting and segmenting multiple-structure data with outliers,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 6, pp. 1177–1192, Jun. 2012.

[14] C. Harris and M. Stephens, “A combined corner and edge detector,” in Proc. 4th Alvey Vis. Conf., 1988, pp. 147–151.

[15] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, Nov. 2004.

[16] S. Choi, T. Kim, and W. Yu, “Performance evaluation of RANSAC family,” in Proc. Brit. Mach. Vis. Conf., 2009, p. 1.

[17] O. Chum, J. Matas, and J. Kittler, “Locally optimized RANSAC,” in Proc. Joint Pattern Recognit. Symp., in Lecture Notes in Computer Science, vol. 2781, 2003, pp. 236–243.

[18] H. K. Tang and K. R. Ramakrishnan, “A probabilistic analysis of a common RANSAC heuristic,” Mach. Vis. Appl., vol. 30, no. 1, pp. 71–89, Feb. 2019.

[19] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in Proc. 7th Int. Joint Conf. Artif. Intell., vol. 2, 1981, pp. 674–679.

[20] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” in Proc. Int. Conf. Comput. Vis., Nov. 2011, pp. 2564–2571.

[21] D. DeTone, T. Malisiewicz, and A. Rabinovich, “Deep image homography estimation,” 2016, arXiv:1606.03798. [Online]. Available: http://arxiv.org/abs/1606.03798.

[22] S. Baker, A. Datta, and T. Kanade, “Parameterizing homographies,” Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-RI-TR-06-11, 2006.

[23] Q. Zhou and X. Li, “$\text{STN}$-homography: Direct estimation of homography parameters for image pairs,” Appl. Sci., vol. 9, no. 23, p. 5187, 2019.

[24] F. E. Nowruz, R. Lagnaniere, and N. Japkowicz, “Homography estimation from image pairs with hierarchical convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2017, pp. 904–911.

[25] L. Kang, Y. Wei, Y. Xie, J. Jiang, and Y. Guo, “Combining convolutional neural network and photometric refinement for accurate homography estimation,” IEEE Access, vol. 7, pp. 109460–109473, 2019.

[26] G. Wang, Z. You, P. An, J. Yu, and Y. Chen, “Efficient and robust homography estimation using compressed convolutional neural network,” in Digital TV and Multimedia Communication (Communications in Computer and Information Science), vol. 1009, G. Zhai, J. Zhou, P. An, and X. Yang, Eds. Singapore: Springer, 2019.

[27] X. Wang, C. Wang, X. Bai, Y. Liu, and J. Zhou, “Deep homography estimation with pairwise invertibility constraint,” in Structural, Syntactic, and Statistical Pattern Recognition (Lecture Notes in Computer Science), vol. 11004. X. Bai, E. Hancock, T. Ho, R. Wilson, B. Biggio, and A. Robles-Kelly, Eds. Cham, Switzerland: Springer, 2018.

[28] T. Nguyen, S. W. Chen, S. S. Shikumakura, C. J. Taylor, and V. Kumar, “Unsupervised deep homography: A fast and robust homography estimation model,” IEEE Robot. Autom. Lett., vol. 3, no. 3, pp. 2346–2353, Jul. 2018.
[29] M. A. Molina-Cabello, R. M. Luque-Baena, E. López-Rubio, and K. Thurnhofer-Hemsi, “Vehicle type detection by ensembles of convolutional neural networks operating on super resolved images,” Integr. Comput.-Aided Eng., vol. 25, no. 4, pp. 321–333, Sep. 2018.

[30] Y. Vardi and C.-H. Zhang, “The multivariate $L_1$-median and associated data depth,” Proc. Nat. Acad. Sci. USA, vol. 97, no. 4, pp. 1423–1426, 2000.

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2014, pp. 740–755.

[32] C. Wang, X. Wang, X. Bai, Y. Liu, and J. Zhou, “Self-supervised deep homography estimation with invertibility constraints,” Pattern Recognit. Lett., vol. 128, pp. 355–360, Dec. 2019.

[33] F. E. Nowruzi, R. Laganier, and N. Japkowicz, “Homography estimation from image pairs with hierarchical convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2017, pp. 913–920.

[34] M. A. Molina-Cabello, D. Elizondo, R. M. Luque-Baena, and E. López-Rubio, “Homography estimation with deep convolutional neural networks by random color transformations,” in Proc. Brit. Mach. Vis. Conf. (BMVC), 2019, pp. 1–11.

MIGUEL A. MOLINA-CABELLO received the M.Sc. and Ph.D. degrees in computer engineering from the University of Málaga, Spain, in 2015 and 2018, respectively. He joined the Department of Computer Languages and Computer Science, University of Málaga, in 2015, where he has a teaching and researching position. He also keeps pursuing research activities in collaboration with other Universities. His research interests include visual surveillance, image/video processing, computer vision, neural networks, and pattern recognition.

DAVID A. ELIZONDO (Senior Member, IEEE) received the B.A. degree in computer science from Knox College, Galesburg, IL, USA, the M.S. degree in artificial intelligence from the Department of Artificial Intelligence and Cognitive Computing, University of Georgia, Athens, GA, USA, and the Ph.D. degree in computer science from the University of Strasbourg, France, in cooperation with the Swiss Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP). He is currently a Professor in intelligent transport systems with the Department of Computer Technology, De Montfort University, U.K.

RAFAEL MARCOS LUQUE-BAENA received the M.S. and Ph.D. degrees in computer engineering from the University of Málaga, Spain, in 2007 and 2012, respectively. He moved to Mérida, Spain, in 2013, as a Lecturer with the Department of Computer Engineering, Centro Universitario de Mérida, University of Extremadura. He currently come back to the University of Málaga, in 2016, where he has been a teaching position with the Department of Languages and Computer Science. He also keeps pursuing research activities in collaboration with other Universities. His current research interests include visual surveillance, image/video processing, neural networks, and pattern recognition.

EZEQUIEL LÓPEZ-RUBIO was born in 1976. He received the M.Sc. and Ph.D. (Hons.) degrees in computer engineering from the University of Málaga, Spain, in 1999 and 2002, respectively. He joined the Department of Computer Languages and Computer Science, University of Málaga, in 2000, where he is currently a Full Professor of computer science and artificial intelligence. His research interests include deep learning, pattern recognition, image processing, and computer vision.