

# A Robust Forensic Hash Component for Image Alignment

S. Battiato, G. M. Farinella, E. Messina, G. Puglisi  
{battiato, gfarinella, emessina, puglisi}@dmi.unict.it

Image Processing Laboratory  
Department of Mathematics and Computer Science  
University of Catania  
Viale A. Doria 6 - 95125 Catania, Italia  
<http://iplab.dmi.unict.it>

**Abstract.** The distribution of digital images with the classic and newest technologies available on Internet (e.g., emails, social networks, digital repositories) has induced a growing interest on systems able to protect the visual content against malicious manipulations that could be performed during their transmission. One of the main problems addressed in this context is the authentication of the image received in a communication. This task is usually performed by localizing the regions of the image which have been tampered. To this aim the received image should be first registered with the one at the sender by exploiting the information provided by a specific component of the forensic hash associated with the image. In this paper we propose a robust alignment method which makes use of an image hash component based on the Bag of Visual Words paradigm. The proposed signature is attached to the image before transmission and then analyzed at destination to recover the geometric transformations which have been applied to the received image. The estimator is based on a voting procedure in the parameter space of the geometric model used to recover the transformation occurred to the received image. Experiments show that the proposed approach obtains good margin in terms of performances with respect to state-of-the-art methods.

**Keywords:** Image forensics, Forensic hash, Bag of Visual Word, Tampering, Geometric transformations, Image validation and authentication.

## 1 Introduction and Motivations

The growing demand of techniques useful to protect digital visual data against malicious manipulations is induced by different episodes that make questionable the use of visual content as evidence material [1]. Methods useful to establish the validity and authenticity of a received image are needed in the context of Internet communications. To this aim different solutions have been recently proposed in literature [2–6]. Most of them share the same basic scheme: i) a hash code based on the visual content is attached to the image to be sent; ii) the hash is analyzed at destination to verify the reliability of the received image.

An image hash is a distinctive signature which represents the visual content of the image in a compact way (usually just few bytes). The image hash should

be robust against allowed operations and at the same time it should differ from the one computed on a different/tampered image. Image hashing techniques are considered extremely useful to validate the authenticity of an image received through the Internet. Although the importance of the binary decision task related to the image authentication, this is not always sufficient. In the application context of Forensic Science is fundamental to provide scientific evidences through the history of the possible manipulations applied to the original image to obtain the one under analysis. In many cases, the source image is unknown, and, as in the application context of this paper, all the information about the manipulation of the image should be recovered through the short image hash signature, making more challenging the final task. The list of manipulations provides to the end user the information needed to decide whether the image can be trusted or not.

In order to perform tampering localization<sup>1</sup>, the receiver should be able to filter out all the geometric transformations (e.g., rotation, scaling) added to the tampered image by aligning the received image with the one at the sender [6]. The alignment should be done in a semi-blind way: at destination one can use only the received image and the image hash to deal with the alignment problem since the reference image is not available. The challenging task of recovering the geometric transformations occurred on a received image from its signature motivates this paper.

Despite different robust alignment techniques have been proposed by computer vision researchers [7], these techniques are unsuitable in the context of forensic hashing, since a fundamental requirement is that the image signature should be as “compact” as possible to reduce the overhead of the network communications. To fit the underlying requirements, authors of [5] have proposed to exploit information extracted through Radon transform and scale space theory in order to estimate the parameters of the geometric transformations. To make more robust the alignment phase with respect to manipulations such as cropping and tampering, an image hash based on robust invariant features has been proposed in [6]. The latter technique extended the idea previously proposed in [4] by employing the Bag of Visual Words (BOVW) model to represent the features to be used as image hash. The exploitation of the BOVW representation is useful to reduce the space needed for the image signature, by maintaining the performances of the alignment component.

Building on the technique described in [6], we propose a new method to detect the geometric manipulations occurred on an image starting from the hash computed on the original one. Differently than [6], we exploit replicated visual words and a voting procedure in the parameter space of the transformation model employed to establish the geometric parameters (i.e., rotation, scale, translation). As pointed out by the experimental results, the proposed approach obtains the

---

<sup>1</sup> Tampering localization is the process of localizing the regions of the image that have been manipulated for malicious purposes to change the semantic meaning of the visual message.

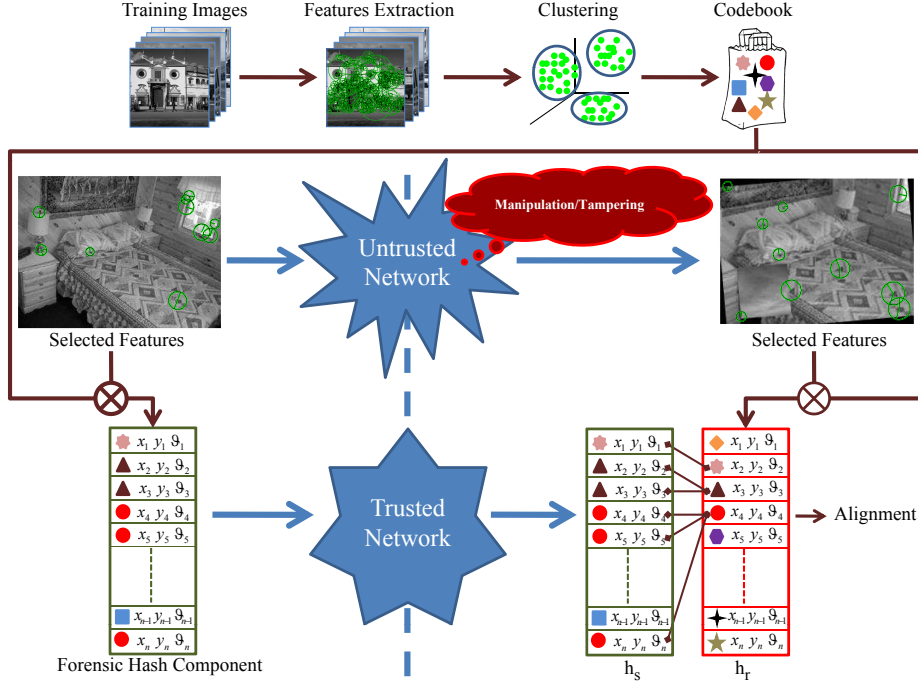


Fig. 1. Schema of the proposed approach.

best results with a significant margin in terms of estimation accuracy with respect to the approach proposed in [6].

The remainder of the paper is organized as follows: Section 2 presents the proposed framework. Section 3 reports the experiments and discusses the results. Finally, Section 4 concludes the paper with avenues for further research.

## 2 Proposed Approach

As stated in the previous section, one of the common steps of tampering detection systems is the alignment of the received image. Image registration is crucial since all the other tasks (e.g., tampering localization) usually assume that the received image is aligned with the original one, and hence could fail if the registration is not properly done. Classical registration approaches [7] cannot be directly employed in the considered context due the limited information that can be used (i.e., original image is not available and the image hash should be as short as possible).

The schema of the overall system is shown in Fig. 1. As in [6], we adopt a Bag of Visual Words based representations [8] to reduce the dimensionality of the feature descriptors to be used as hash component for the alignment. A codebook is generated by clustering the set of SIFT [9] extracted on training images. The pre-computed codebook is shared between sender and receiver. It should be noted

that the codebook is built only once, and then used for all the communications between a sender and a receiver (i.e., no extra overhead for each communication). Sender extracts SIFT features and sorts them in descending order with respect to their contrast values. Afterward, the top  $n$  SIFT are selected and associated to the  $id$  label corresponding to the closest visual word belonging to the shared codebook. Hence, the final signature for the alignment component is created by considering the  $id$  label, the dominant direction  $\theta$ , and the keypoint coordinates  $x$  and  $y$  for each selected SIFT (Fig. 1). The source image and the corresponding hash component ( $h_s$ ) are sent to the destination. The system assumes that the image is sent over a network consisting of possibly untrusted nodes, whereas the signature is sent upon request through a trusted authentication server which encrypts the hash in order to guarantee its integrity [3]. The image could be manipulated for malicious purposes during the untrusted communication.

Once the image reaches the destination, the receiver generates the related hash signature ( $h_r$ ) by using the same procedure employed by the sender. Then, the entries of the hashes  $h_s$  and  $h_r$  are matched by considering the  $id$  values (see Fig. 1). The alignment is hence performed by employing a similarity transformation of the keypoint pairs corresponding to matched hashes entries:

$$x_r = x_s \lambda \cos \alpha - y_s \lambda \sin \alpha + T_x \quad (1)$$

$$y_r = x_s \lambda \sin \alpha + y_s \lambda \cos \alpha + T_y \quad (2)$$

The above transformation is used to model the geometrical manipulations which have been done on the source image during the untrusted communication. The model assumes that a point  $(x_s, y_s)$  in the source image  $I_s$  is transformed in a point  $(x_r, y_r)$  in the image  $I_r$  at destination with a combination of rotation ( $\alpha$ ), scaling ( $\lambda$ ) and translation ( $T_x, T_y$ ). The aim of the alignment phase is the estimation of the quadruple  $(\hat{\lambda}, \hat{\alpha}, \hat{T}_x, \hat{T}_y)$  by exploiting the correspondences  $((x_s, y_s), (x_r, y_r))$  related to matchings between  $h_s$  and  $h_r$ . We propose to use a cascade approach to perform the parameter estimation. First the estimation of  $(\hat{\alpha}, \hat{T}_x, \hat{T}_y)$  is accomplished through a voting procedure in the parameter space  $(\alpha, T_x, T_y)$ . Such procedure is performed after filtering outlier matchings by taking into account the differences between dominant orientations of matched entries. Then the scaling parameter  $\hat{\lambda}$  is estimated by considering the parameters  $(\hat{\alpha}, \hat{T}_x, \hat{T}_y)$  which have been previously estimated on the reliable information obtained through the filtering. The proposed method is detailed in the following.

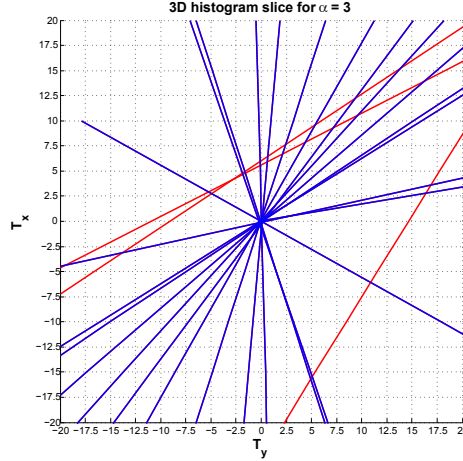
Moving  $T_x$  and  $T_y$  on the left side and making the ratio of (1) and (2) the following equation is obtained:

$$\frac{x_r - T_x}{y_r - T_y} = \frac{x_s \cos \alpha - y_s \sin \alpha}{x_s \sin \alpha + y_s \cos \alpha} \quad (3)$$

Solving (3) with respect to  $T_x$  we get the formula to be used in the voting procedure:

$$T_x = \left( \frac{x_s \cos \alpha - y_s \sin \alpha}{x_s \sin \alpha + y_s \cos \alpha} \right) (T_y - y_r) + x_r \quad (4)$$

Each pair of coordinates  $(x_s, y_s)$  and  $(x_r, y_r)$  in (4) represents a line in the parameter space  $(\alpha, T_x, T_y)$ . An initial estimation of  $(\widehat{\alpha}, \widehat{T}_x, \widehat{T}_y)$  is obtained by considering the densest bin of a 3D histogram in the quantized parameter space. This means that the initial estimation of  $(\widehat{\alpha}, \widehat{T}_x, \widehat{T}_y)$  is accomplished in correspondence of the maximum number of intersections between lines generated by matched keypoints (Fig. 2).



**Fig. 2.** A slices of the 3D histogram in correspondence of  $\alpha = 3$ , obtained considering an image manipulated with parameters  $(\lambda, \alpha, T_x, T_y) = (1, 3, 0, 0)$ . For a fixed rotational angle  $\bar{\alpha}$ , each pair of coordinates  $(x_s, y_s)$  and  $(x_r, y_r)$  votes for a line in the quantized 2D parameter space  $(T_x, T_y)$ . Lines corresponding to inliers (blue) intersect in the bin  $(T_x, T_y) = (0, 0)$ , whereas the remaining lines (red) are related to outliers.

As said before, to discard outliers (i.e., wrong matchings) the information coming from the dominant directions ( $\theta$ ) of the SIFT are used during the voting procedure. In particular  $\Delta\theta = \theta_r - \theta_s$  is a rough estimation of the rotational angle  $\alpha$ . Hence, for each fixed triplet  $(\bar{\alpha}, \bar{T}_x, \bar{T}_y)$  of the quantized parameter space, the voting procedure considers only the matchings between  $h_s$  and  $h_r$  such that  $|\Delta\theta - \bar{\alpha}| < t_\alpha$ . The threshold value  $t_\alpha$  is chosen to consider only matchings with a rough estimation  $\Delta\theta$  which is closer to the considered  $\bar{\alpha}$  (e.g., consider just matchings with a small initial error of  $\pm 3.5$  degree). The proposed approach is summarized in Algorithm 1.

The proposed method gives an estimation of rotation angle  $\widehat{\alpha}$ , and translation vector  $(\widehat{T}_x, \widehat{T}_y)$  by taking into account the quantized values used to build the 3D histogram into the parameter space. To refine the estimation we can use the  $m$  matchings which have been generated by the  $m$  lines which intersect in the selected bin. Specifically, for each pair  $(x_{s,i}, y_{s,i}), (x_{r,i}, y_{r,i})$  corresponding to the selected bin, we consider  $(\widehat{T}_{x,i}, \widehat{T}_{y,i}) = \left( \left( \frac{x_s \cos \bar{\alpha} - y_s \sin \bar{\alpha}}{x_s \sin \bar{\alpha} + y_s \cos \bar{\alpha}} \right) (\bar{T}_y - y_r) + x_r, \bar{T}_y \right)$ , with  $(\bar{\alpha}, \bar{T}_y)$  obtained through the voting procedure (see Algorithm 1), and use the equations (5) and (6).

**Algorithm 1:** Parameters estimation through voting procedure.

---

**Input:** The set  $M$  of matching pairs  $((x_s, y_s), (x_r, y_r))$   
**Output:** The estimated parameter  $(\hat{\alpha}, \hat{T}_x, \hat{T}_y)$

**begin**

*Initialize*  $Votes(i, j, k) := 0 \ \forall i, j, k;$

**for**  $\bar{\alpha} = -180, -179, \dots, 0, \dots, 179, 180$  **do**

$V_\alpha = \{((x_s, y_s), (x_r, y_r)) \mid |(\theta_r - \theta_s) - \bar{\alpha}| < t_\alpha\};$

**foreach**  $((x_s, y_s), (x_r, y_r)) \in V_\alpha$  **do**

**for**  $\bar{T}_y = \min_{T_y}, \min_{T_y} - 1, \dots, \max_{T_y} - 1, \max_{T_y}$  **do**

$T_x := \left( \frac{x_s \cos \bar{\alpha} - y_s \sin \bar{\alpha}}{x_s \sin \bar{\alpha} + y_s \cos \bar{\alpha}} \right) (\bar{T}_y - y_r) + x_r;$

$\bar{T}_x := \text{Quantize}(T_x);$

$(i, j, k) := \text{QuantizedValuesToBin}(\bar{\alpha}, \bar{T}_x, \bar{T}_y);$

$Votes(i, j, k) := Votes(i, j, k) + 1;$

$(i_{max}, j_{max}, k_{max}) = \text{SelectBin}(Votes);$

$(\hat{\alpha}, \hat{T}_x, \hat{T}_y) := \text{BinToQuantizedValues}(i_{max}, j_{max}, k_{max});$

**end**

---

$$x_{r,i} = x_{s,i} \lambda_i \cos \alpha_i - y_{s,i} \lambda_i \sin \alpha_i + \hat{T}_{x,i} \quad (5)$$

$$y_{r,i} = x_{s,i} \lambda_i \sin \alpha_i + y_{s,i} \lambda_i \cos \alpha_i + \hat{T}_{y,i} \quad (6)$$

Solving (5) and (6) with respect to  $a_i = \lambda_i \cos \alpha_i$  and  $b_i = \lambda_i \sin \alpha_i$  we obtain

$$\hat{a}_i = \frac{y_{r,i} y_{s,i} + x_{r,i} x_{s,i} - x_{s,i} \hat{T}_{x,i} - y_{s,i} \hat{T}_{y,i}}{x_{s,i}^2 + y_{s,i}^2} \quad (7)$$

$$\hat{b}_i = \frac{x_{s,i} y_{r,i} - x_{r,i} y_{s,i} + y_{s,i} \hat{T}_{x,i} - x_{s,i} \hat{T}_{y,i}}{x_{s,i}^2 + y_{s,i}^2} \quad (8)$$

Since the ratio  $\hat{b}_i / \hat{a}_i$  is by definition equals to  $\tan \alpha_i$ , we can estimate  $\hat{\alpha}_i$  with the following formula:

$$\hat{\alpha}_i = \arctan \left( \frac{x_{s,i} y_{r,i} - x_{r,i} y_{s,i} + y_{s,i} \hat{T}_{x,i} - x_{s,i} \hat{T}_{y,i}}{y_{r,i} y_{s,i} + x_{r,i} x_{s,i} - x_{s,i} \hat{T}_{x,i} - y_{s,i} \hat{T}_{y,i}} \right) \quad (9)$$

Once  $\hat{\alpha}_i$  is obtained, the following equation derived from (5) and (6) is used to estimate  $\hat{\lambda}_i$

$$\hat{\lambda}_i = \frac{1}{2} \left( \frac{x_{r,i} - \hat{T}_{x,i}}{x_{s,i} \cos \hat{\alpha}_i - y_{s,i} \sin \hat{\alpha}_i} + \frac{y_{r,i} - \hat{T}_{y,i}}{x_{s,i} \sin \hat{\alpha}_i + y_{s,i} \cos \hat{\alpha}_i} \right) \quad (10)$$

The above method produce a quadruple  $(\hat{\lambda}_i, \hat{\alpha}_i, \hat{T}_{x,i}, \hat{T}_{y,i})$  for each matching pair  $(x_{s,i}, y_{s,i}), (x_{r,i}, y_{r,i})$  corresponding to the bin selected with the voting procedure. The final transformation parameters  $(\hat{\lambda}, \hat{\alpha}, \hat{T}_x, \hat{T}_y)$  to be used for the alignment are computed by averaging over all the  $m$  produced quadruple:

$$\hat{T}_x = \frac{1}{m} \sum_i^m \hat{T}_{x,i} \quad \hat{T}_y = \frac{1}{m} \sum_i^m \hat{T}_{y,i} \quad \hat{\alpha} = \frac{1}{m} \sum_i^m \hat{\alpha}_i \quad \hat{\lambda} = \frac{1}{m} \sum_i^m \hat{\lambda}_i \quad (11)$$

It should be noted that some  $id$  values may appear more than once in  $h_s$  and/or in  $h_r$ . For example, it is possible that a selected SIFT has no unique dominant direction [9]; in this case the different directions are coupled with the same descriptor, and hence will be considered more than once by the selection process which generates many instance of the same  $id$  with different dominant directions.

As experimentally demonstrated in the next section, by retaining the replicated visual words the accuracy of the estimation increases, and the number of “unmatched” images decreases (i.e., image pairs that the algorithm is not able to process because there are no matchings between  $h_s$  and  $h_r$ ). Differently than [6], the described approach considers all the possible matchings in order to preserve the useful information. The correct matchings are hence retained but other wrong pairs could be generated. Since the noise introduced by considering correct and incorrect pairs can badly influence the final estimation results, the presence of possible wrong matchings should be considered during the estimation process. The approach described in this paper deals with the problem of wrong matchings combining in cascade a filtering strategy based on the SIFT dominant direction ( $\theta$ ) with a robust estimator based on a voting strategy on the parameters’ space. In this way the information of spatial position of keypoints and their dominant orientations are jointly considered. The scale factor is estimated only at the end of the cascade on reliable information. As shown in the experiments reported in the following section, replicated matchings help to better estimate the rotational parameter, whereas the introduced cascade approach allows robustness in estimating the scale factor.

### 3 Experimental Results

This section reports a number of experiments on which the proposed approach has been tested and compared with respect to [6]. The tests have been performed considering a subset of the fifteen scene category benchmark dataset [10]. The training set used in the experiments is built through a random selection of 150 images from the scene dataset. Specifically, ten images have been randomly sampled from each scene category. The test set consists of 5250 images generated through the application of different transformations on the training images<sup>2</sup>. Accordingly with [6], the following image manipulations have been applied (Tab. 1): cropping, rotation, scaling, tampering, JPEG compression. The considered transformations are typically available on image manipulation software. Tampering has been performed through the swapping of blocks ( $50 \times 50$ ) between two images randomly selected from the training set. Images obtained through various combinations of the basic transformations have been also included to make more challenging the task to be addressed. Taking into account the different parameter settings, for each training image there are 35 corresponding manipulated images into the test set.

<sup>2</sup> Training and test sets used for the experiments are available at [http://iplab.dmi.unict.it/download/ICIAP\\_2011/Dataset.rar](http://iplab.dmi.unict.it/download/ICIAP_2011/Dataset.rar)

**Table 1.** Image transformations.

Operations	Parameters
Rotation ( $\alpha$ )	3, 5, 10, 30, 45 degrees
Scaling ( $\sigma$ )	factor= 0.5, 0.7, 0.9, 1.2, 1.5
Cropping	19%, 28%, 36%, of entire image
Tampering	block size 50x50
Compression	JPEG Q=10
Various combinations of above operations	

To demonstrate the effectiveness of the proposed approach, and to highlight the contribution of the replicated visual words during the estimation, the comparative tests have been performed by considering our method and the approach proposed in [6]. Although Lu et al. [6] claim that further refinements are performed using the points that occur more than once, actually they do not provide any implementation detail. In our test we have hence considered two versions of [6] with and without replicated matchings. The approach proposed in [6] has been reimplemented. The Ransac thresholds used in [6] to perform the geometric parameter estimation have been set to 3.5 degrees for the rotational model and to 0.025 for the scaling one. These thresholds values have been obtained through data analysis (inliers and outliers distributions). In order to perform a fair comparison, the threshold  $t_\alpha$  used in our approach to filter the correspondences (see Section 2) has been set with the same value of the threshold employed by Ransac to estimate the rotational parameter in [6]. The value  $T_y$  needed to evaluate (4) has been quantized considering a step of 2.5 pixels (see Fig. 2). Finally, a codebook with 1000 visual words has been employed to compare the different approaches. The codebook has been learned through k-means clustering on the overall SIFT descriptors extracted on training images.

First, let us examine the typically cases in which the considered approaches are not able to work. Two cases can be distinguished: i) no matchings are found between the hash built at the sender ( $h_s$ ) and the one computed by the receiver ( $h_r$ ); ii) all the matchings are replicated. The first problem can be mitigated considering a higher number of features (SIFT). The second one is solved only allowing the replicated matchings (see Section 2). As reported in Tab. 2, by increasing the number of SIFT there is a decreasing of the number of unmatched images for both approaches. In all cases the percentage of images on which our algorithm is not able to work is lower than the one of [6].

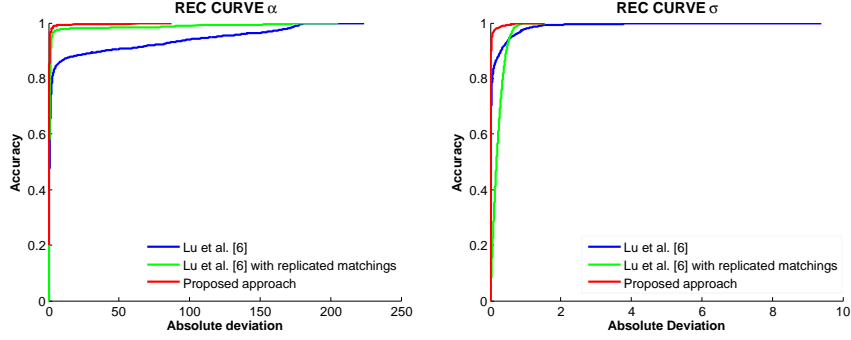
**Table 2.** Comparison with respect to unmatched images.

Number of SIFT	Unmatched Images			
	15	30	45	60
Lu et al. [6]	11.73%	3.60%	1.64%	0.91%
Proposed approach	4.44%	1.26%	0.57%	0.46%

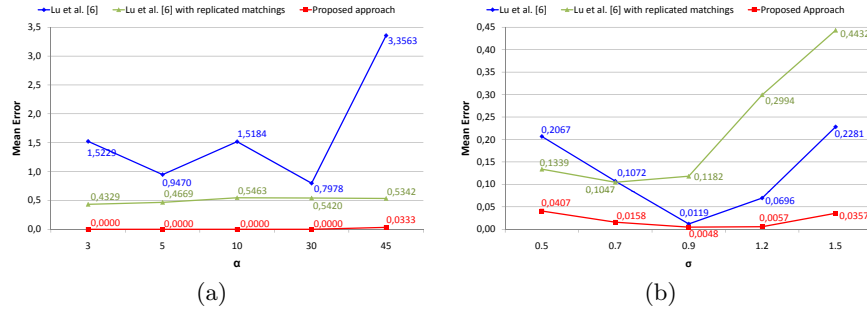
Tab. 3 shows the results obtained in terms of rotational and scale estimation through mean error. To properly compare the methods, the results have been computed taking into account only the images on which all approaches are able to work. Our approach outperforms [6] obtaining a considerable gain both in terms of rotational and scaling accuracy (Tab. 3). Moreover, the performance of

**Table 3.** Average rotational and scaling error.

Number of SIFT	Mean Error $\alpha$				Mean Error $\sigma$			
	15	30	45	60	15	30	45	60
Lu et al. [6]	12.8135	13.7127	13.2921	13.5840	0.1133	0.1082	0.1086	0.1124
Lu et al. [6] with replicated matchings	6.7000	4.1444	3.3647	2.8677	0.1522	0.1783	0.1981	0.2169
Proposed approach	2.2747	1.2987	0.6514	0.5413	0.0710	0.0393	0.0230	0.0183



**Fig. 3.** REC curves comparison.



**Fig. 4.** Comparison on single transformation (60 SIFT). (a) Average rotation error at varying of the rotation angle. (b) Average scaling error at varying of the scale factor.

our approach significantly improves with the increasing of the extracted feature points (SIFT). On the contrary, the technique in [6] is not able to exploit the additional information coming from the higher number of extracted points.

To further study the contribution of the replicated matchings we performed tests by considering the modified version of [6] in which replicated matchings have been allowed (Tab. 3). Although the modified approach obtains better performance with respect to the original one in terms of rotational accuracy, it is not able to obtain satisfactory results in terms of scaling estimation. The wrong pairs introduced by the replicated matchings cannot be handled by the method. Our approach deals with the problem of wrong pairs combining a filtering based on the SIFT dominant direction ( $\theta$ ) with a robust estimator based on voting.

To better compare the methods, the Regression Error Characteristic Curves (REC) have been employed (Fig. 3). The area over the curve is an estimation of the expected error of a model. The proposed approach obtains the best results.

Additional experiments have been performed to examine the dependence of the average rotational and scaling error with respect to the rotation and scale transformation parameters respectively. Results in Fig. 4(a) show that the rotational estimation error increases with the rotation angle. For the scale transformation, the error has lower values in the proximity of one (no scale change) and increases considering scale factors higher or lower than one (Fig. 4(b)). It should be noted that our approach obtains the best performances in all cases.

## 4 Conclusions and Future Works

The assessment of the reliability of an image received through the Internet is an important issue in nowadays society. This paper addressed the image alignment task in the context of distributed forensic systems. Specifically, a robust image alignment component which exploits an image signature based on the Bag of Visual Words paradigm has been introduced. The proposed approach has been experimentally tested on a representative dataset of scenes obtaining effective results in terms of estimation accuracy. Future works will concern the extension of the system to allow tampering detection. Moreover, a selection step able to takes into account the spatial distribution of SIFT will be addressed in order to avoid their concentration on high textured regions.

## References

1. Farid, H.: Digital doctoring: how to tell the real from the fake. *Significance* 3(4), 162–166 (2006)
2. Battiato, S., Farinella, G. M., Messina, E., Puglisi, G.: Understanding Geometric Manipulations of Images Through BOVW-Based Hashing. In: *IEEE International Workshop on Content Protection & Forensics*, held in conjunction with the *IEEE International Conference on Multimedia & Expo* (2011)
3. Lin, Y.-C., Varodayan, D., Girod, B.: Image authentication based on distributed source coding. In: *IEEE International Conference on Image Processing*, 3–8 (2007)
4. Roy, S., Sun, Q.: Robust hash for detecting and localizing image tampering. In: *IEEE International Conference on Image Processing*, 117–120 (2007)
5. Lu, W., Varna, A. L., Wu, M.: Forensic hash for multimedia information. In: *IS&T-SPIE Electronic Imaging Symposium - Media Forensics and Security* (2010)
6. Lu, W.J., Wu, M.: Multimedia forensic hash based on visual words. In: *IEEE International Conference on Image Processing*, 989–992 (2010)
7. Szeliski, R.: Image alignment and stitching: A tutorial. *Foundations and Trends in Computer Graphics and Computer Vision*, (2):1 1–104 (2006)
8. Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C.: Visual categorization with bags of keypoints. In: *ECCV International Workshop on Statistical Learning in Computer Vision* (2004)
9. Lowe, D.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2): 91–110 (2004)
10. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2169–2178 (2006)