Interactive Museum Guide: Fast and Robust Recognition of Museum Objects

Herbert Bay, Beat Fasel and Luc Van Gool

Computer Vision Laboratory (BIWI), ETH Zurich
Sternwartstr. 7, 8092 Zurich, Switzerland
{bay,bfasel,vangool}@vision.ee.ethz.ch

Abstract. In this paper, we describe the application of the novel SURF (Speeded Up Robust Features) algorithm [1] for the recognition of objects of art. For this purpose, we developed a prototype of a mobile interactive museum guide consisting of a tablet PC that features a touchscreen and a webcam. This guide recognises objects in museums based on images taken by the visitor. Using different image sets of real museum objects, we demonstrate that both the object recognition performance as well as the speed of the SURF algorithm surpasses the results obtained with SIFT, its main contender.

1 Introduction

Many museums still present their exhibits in a rather passive and non-engaging way. The visitor has to search through a booklet in order to find descriptions about the objects on display. However, looking for information in this way is a quite tedious procedure. Moreover, the information found does not always meet the visitor's specific interests. One possibility of making exhibitions more attractive to the visitor is to improve the interaction between the visitor and the objects of interest by means of a guide. In this paper, we present an interactive museum guide that is able to automatically find and instantaneously retrieve information about the objects of interest using a standard tablet PC. Undoubtedly, technological developments will lead to less heavy and downsized solutions in the near future. The focus of this paper is on the vision component used to recognise the objects.

1.1 Related Work

Recently, several approaches have been proposed that allow visitors to interact via an automatic museum guide. Kusunoki *et al.* [2] proposed a system for children that uses a sensing board, which can rapidly recognise type and locations of multiple objects. It creates an immersing environment by giving audio-visual feedback to the children. Other approaches include robots that guide users through museums [3, 4]. However, such robots are difficult to adapt to different environments, and they are not appropriate for individual use. An interesting

approach using hand-held devices, like mobile phones, was proposed by [5], but their recognition technique seems not to be very robust to viewing angle or lighting changes.

Various object recognition methods have been investigated in the last two decades. More recently, SIFT [6] and its variants such as PCA-SIFT [7] and GLOH [8] have been successfully applied for many image matching applications. In this paper, we show that the new SURF (Speeded Up Robust Features) algorithm [1] surpasses SIFT in both speed and recognition accuracy.

1.2 Interactive Museum Guide

The proposed interactive, image-based museum guide is invariant to changes in lighting, translation, scale, rotation and viewpoint variations. Our object recognition system was implemented on a Tablet PC using a conventional USB webcam for image acquisition, see Figure 1. This hand-held device allows the visitor to simply take a picture of an object of interest from any position and is provided, almost immediately, with a detailed description of the latter.

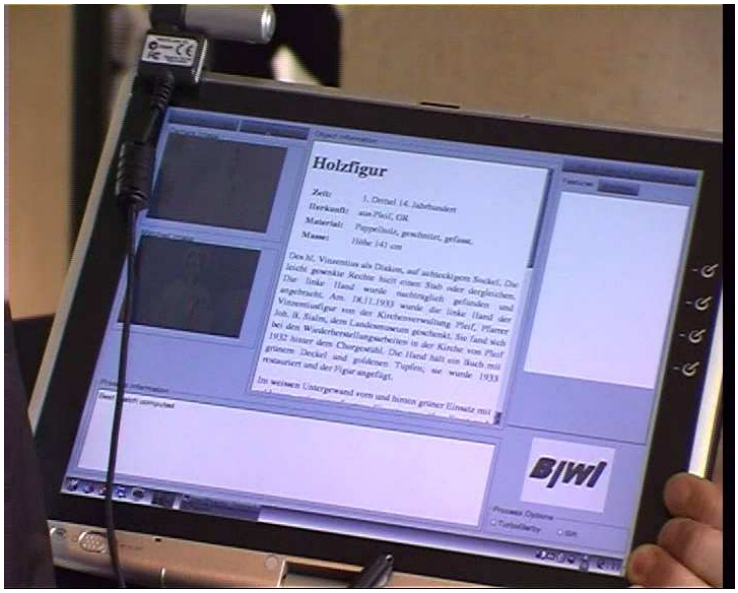

Fig. 1. Tablet PC with the USB webcam fixed on the screen. The interface of the object recognition software is operated via a touchscreen.

An early prototype of this museum guide was shown to the public during the 150 years anniversary celebration of the Federal Institute of Technology (ETH) in Zurich, Switzerland [9]. The descriptions of the recognised objects of art are

read to the visitors by a synthetic computer voice. This enhances the convenience of the guide as the visitors can focus on the objects of interest instead of reading the object descriptions on the screen of the guide.

In order to demonstrate the recognition capabilities of our latest implementation, we created a database with objects on display in the Landesmuseum. A sample image of each of the 20 chosen objects is shown in Figure 2.

Fig. 2. Sample images of the 20 chosen art objects from the Landesmuseum.

The remainder of this paper is organised as follows. First, we introduce our object recognition system in detail (Section 2). Then, we present and discuss results obtained for a multi-class task (Section 3), and finally conclude with an overall discussion and some final remarks (Section 4).

2 Object Recognition System

We developed an object recognition system that is based on interest point correspondences between individual image pairs. Input images, taken by the user, are compared to all model images in the database. This is done by matching their respective interest points. The model image with the highest number of matches with respect to the input image is chosen as the one which represents the object the visitor is looking for.

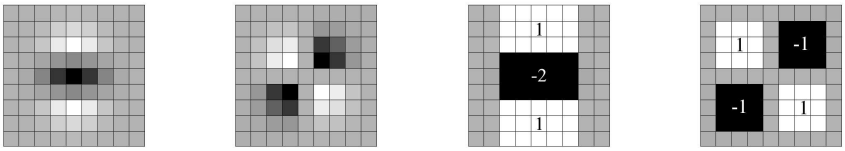

Fig. 3. Left to right: the (discretised and cropped) Gaussian second order partial derivatives in y-direction and xy-direction, and our approximations thereof using box filters. The grey regions are equal to zero.

Furthermore, we propose a new object identification strategy based on the mean Euclidean distance between all matching pairs. The latter proved to yield better results than the aforementioned traditional approach.

In the following sub-sections we shortly describe the SURF algorithm. Then, we present the new object selection strategy.

2.1 Fast Interest Point Detection

The SURF feature detector is based on the Hessian matrix. Given a point $\mathbf{x} = (x, y)^\top$ in an image I, the Hessian matrix $\mathcal{H}(\mathbf{x}, \sigma)$ in $\mathbf{x}$ at scale σ is defined as follows

$$\mathcal{H}(\mathbf{x}, \sigma) = \begin{bmatrix} L_{xx}(\mathbf{x}, \sigma) & L_{xy}(\mathbf{x}, \sigma) \\ L_{xy}(\mathbf{x}, \sigma) & L_{yy}(\mathbf{x}, \sigma) \end{bmatrix}, \tag{1}$$

where $L_{xx}(\mathbf{x}, \sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2} g(\sigma)$ with the image I in point $\mathbf{x}$, and similarly for $L_{xy}(\mathbf{x}, \sigma)$ and $L_{yy}(\mathbf{x}, \sigma)$.

In contrast to SIFT, which approximates Laplacian of Gaussian (LoG) with Difference of Gaussians (DoG), SURF approximates second order Gaussian derivatives with box filters, see Figure 3. Image convolutions with these box filters can be computed rapidly by using integral images as defined in [10]. The entry of an integral image $I_\Sigma(\mathbf{x})$ at location $\mathbf{x} = (x, y)^\top$ represents the sum of all pixels in the base image I of a rectangular region formed by the origin and $\mathbf{x}$.

$$I_\Sigma(\mathbf{x}) = \sum_{i=0}^{i \leq x} \sum_{j=0}^{j \leq y} I(i, j) \tag{2}$$

Once we have computed the integral image, it is strait forward to calculate the sum of the intensities of pixels over any upright, rectangular area.

The location and scale of interest points are selected by relying on the determinant of the Hessian. Hereby, the approximation of the second order derivatives is denoted as D_{xx}, D_{yy}, and D_{xy}. By choosing the weights for the box filters adequately, we find as approximation for the Hessian's determinant

$$\det(\mathcal{H}_{\text{approx}}) = D_{xx} D_{yy} - (0.9 D_{xy})^2. \tag{3}$$

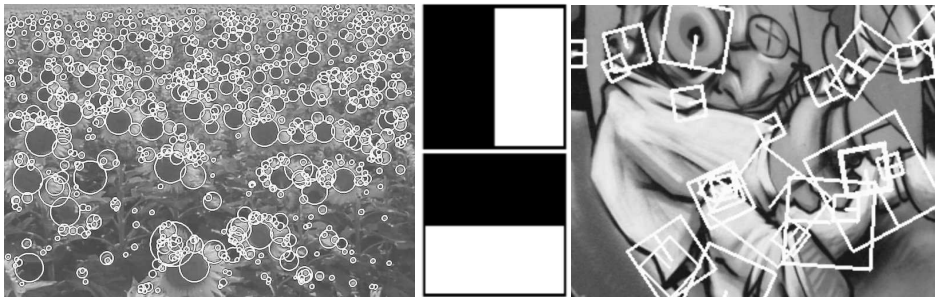

Fig. 4. Left: Detected interest points for a Sunflower field. This kind of scenes show clearly the nature of the features obtained from Hessian-based detectors. Middle: Haar wavelet filters used with SURF. Right: Detail of the Graffiti scene showing the size of the descriptor window at different scales.

For more details, see [1]. Interest points are localised in scale and image space by applying a non-maximum suppression in a $3 \times 3 \times 3$ neighbourhood. Finally, the found maxima of the determinant of the approximated Hessian matrix are interpolated in scale and image space.

2.2 Interest Point Descriptor

In a first step, SURF constructs a circular region around the detected interest points in order to assign a unique orientation to the former and thus gain invariance to image rotations. The orientation is computed using Haar wavelet responses in both x and y direction as shown in the middle of Figure 4. The Haar wavelets can be easily computed via integral images, similar to the Gaussian second order approximated box filters. Once the Haar wavelet responses are computed, they are weighted with a Gaussian with $\sigma = 2.5s$ centred at the interest points. In a next step the dominant orientation is estimated by summing the horizontal and vertical wavelet responses within a rotating wedge, covering an angle of $\frac{\pi}{3}$ in the wavelet response space. The resulting maximum is then chosen to describe the orientation of the interest point descriptor.

In a second step, the SURF descriptors are constructed by extracting square regions around the interest points. These are oriented in the directions assigned in the previous step. Some example windows are shown on the right hand side of Figure 4. The windows are split up in 4×4 sub-regions in order to retain some spatial information. In each sub-region, Haar wavelets are extracted at regularly spaced sample points. In order to increase robustness to geometric deformations and localisation errors, the responses of the Haar wavelets are weighted with a Gaussian, centred at the interest point. Finally, the wavelet responses in horizontal d_x and vertical directions d_y are summed up over each sub-region. Furthermore, the absolute values $|d_x|$ and $|d_y|$ are summed in order to obtain information about the polarity of the image intensity changes. Hence,

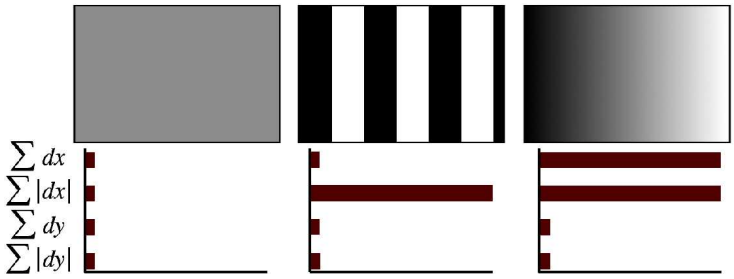

Fig. 5. The descriptor entries of a sub-region represent the nature of the underlying intensity pattern. Left: In case of a homogeneous region, all values are relatively low. Middle: In presence of frequencies in x-direction, the value of $\sum |d_x|$ is high, but all others remain low. If the intensity is gradually increasing in x-direction, both values $\sum d_x$ and $\sum |d_x|$ are high.

the underlying intensity pattern of each sub-region is described by a vector

$$\mathbf{v} = (\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|). \tag{4}$$

The resulting descriptor vector for all 4×4 sub-regions is of length 64. See Figure 5 for an illustration of the SURF descriptor for three different image intensity patterns. Notice that the Haar wavelets are invariant to illumination bias and additional invariance to contrast is achieved by normalising the descriptor vector to unit length.

Rotation-invariant object recognition is not always necessary. Therefore, a scale-invariant-only version of the SURF descriptor was introduced in [1] and denoted 'Upright SURF' (U-SURF). Indeed, in the scenario of a hand-held interactive museum guide, where the museum visitor holds the device in both hands, it is save to assume that images of objects are mostly taken in an upright position. Therefore, U-SURF can be used as an alternative descriptor with the benefit of both increased speed and discrimination power. U-SURF is faster than SURF as it does not perform the orientation related computations.

In this paper, we compare the results for SURF, referred to as SURF-64, and some alternative version (SURF-36, SURF-128) as well as for the upright counterparts (U-SURF-64, U-SURF-36, U-SURF-128) that are not invariant to image rotation. The difference between SURF and its variants lies in the dimension of the descriptor. SURF-36 extracts the descriptor vector from equation (4) for only 3×3 subregions. SURF-128 is an extended version of SURF that treats sums of d_x and $|d_x|$ separately for $d_y < 0$ and $d_y \geq 0$. Similarly, the sums of d_y and $|d_y|$ are split up according to the sign of d_x. This doubles the number of features (128 instead of 64) resulting in a more distinctive descriptor, which is not much slower to compute, but slower to match due to its higher dimensionality (but still faster to match than SIFT). The fast matching speed for all SURF versions is achieved by a single step added to the indexing based on the sign of

the Laplacian (trace of the Hessian matrix) of the interest point. The sign of the Laplacian distinguishes bright blobs on a dark background from the inverse situation. 'Bright' interest points are only matched against other 'bright' interest points and similarly for the 'dark' ones. This minimal information permits to almost double the matching speed and it comes at no computational costs, as it has already been computed in the interest point detection step.

2.3 Object Recognition

Traditional object recognition methods rely on model images, each representing a single object in isolation. In practice, however, the necessary segmentation is not always affordable or even possible. For our object recognition application, we use model images where the objects are not separated from the background. Thus, the background also provides features for the matching task. In any given test image, only one object or object group that belongs together is assumed. Hence, object recognition is achieved by image matching. Extracted interest points of the input image are compared to the interest points of all model images. In order to create a set of interest point correspondences M, we used the nearest neighbour ratio matching strategy [11, 6, 12]. This states that a matching pair is detected if its Euclidean distance in descriptor space is closer than 0.8 times the distance to the second nearest neighbour.

The selected object is the one figuring in the model image with the highest recognition score S_R. This score is traditionally the number of total matches in M. However, the presence of mismatches often lead to false detections. This can be avoided with the help of the following new alternative for the estimation of the recognition score. Hereby, we calculate the mean Euclidean distance to the individual nearest neighbours for each image pair. This value is typically smaller for corresponding image pairs than for non-corresponding ones, and it does not depend on the number of extracted features in the individual images. Hence, we maximise the following recognition score

$$S_R = \underset{i}{argmax}\left(\frac{N_i}{\sqrt{\sum_{j=1}^{N_i} d_{ij}^2}}\right) \tag{5}$$

and chose the object for which the mean distance of its matches is smallest. N_i denotes the number of matches in image i. Furthermore, d_{ij} is the Euclidean distance in the descriptor space between a matching pair of keypoints. The matching criteria is that this distance is closer than 0.8 times the distance to the second nearest neighbour.

3 Experimental Results

For each of the 20 objects of art in our database, images of size 320×240 were taken from different viewing angles. This allows for some degree of view-point independence. The database includes a total of 205 model images. These are

grouped in two model sets (M1 and M2) with 105 and 100 images, respectively. The reasons for the choice of two different model sets are the use of two different cameras and the presence of different lighting conditions. Moreover, less model images for a given object represents a more challenging situation for object recognition.

For similar reasons, we built 3 different test sets (T1-T3) with a total of 116 images (42, 34, 40). Each set contains one or more images of all objects. These objects of art are made of different materials, have different shapes and encompass wooden statues, paintings, metal and stone items as well as objects enclosed in glass cabinets which produce interfering reflections. The images were taken from substantially different viewpoints under arbitrary scale, rotation and varying lighting conditions.

The test image sets were evaluated on each of the model sets. The obtained recognition results are shown in Table 1 and 2. Listed are the results for the

Method	**Time**	**Recognition Rate**						**Total**
	D(s)+M(s)	T1/M1	T2/M1	T3/M1	T1/M2	T2/M2	T3/M2	(%)
SURF-36	19+26	81	79	85	71	94	78	81.0
SURF-64	19+38	88	79	90	69	100	78	83.6
SURF-128	19+59	81	91	90	71	97	75	83.5
U-SURF-36	16+26	74	79	90	74	91	75	80.2
U-SURF-64	16+38	86	85	88	74	94	78	83.8
U-SURF-128	16+59	83	94	95	76	94	80	86.5
SIFT	136+83	79	88	90	76	91	75	82.7

Table 1. Image matching results for different SURF versions and SIFT. Listed are both the total detection D(s) and matching time M(s) for all 3 test sets combined with the model sets.

standard recognition score based on the maximum number of matches (Table 1) and the mean Euclidean distance (Table 2) as described in Equation (5). It can be seen that most versions of SURF outperform SIFT for most test sets while being substantially faster for both computation and matching. The recognition rates for the new recognition score, based on the mean Euclidean distance, increase up to 10%. Note that both the SIFT and SURF descriptors were applied on the same interest points for all experiments. The reported computation times were achieved on a Linux Tablet PC equipped with an Intel Pentium M processor running at 1.7 GHz.

Figures 6 and 7 show cases where SURF and SIFT fail to recognise the same foreground objects. On the bottom of Figure 6, two image pairs are displayed where the foreground object is not correctly recognised by the SURF algorithm. Note however, that a correct match was found for valid objects that are visible in the background. In contrast, SIFT did not find enough matches to allow for

Method	Time	Recognition Rate						Total
	D(s)+M(s)	T1/M1	T2/M1	T3/M1	T1/M2	T2/M2	T3/M2	(%)
SURF-36	19+26	86	88	90	76	97	73	84.5
SURF-64	19+38	83	91	88	83	97	83	87.1
SURF-128	19+59	88	85	93	79	100	85	88.0
U-SURF-36	16+26	86	100	98	81	100	85	91.1
U-SURF-64	16+38	86	94	93	81	100	85	89.4
U-SURF-128	16+59	86	94	95	86	100	90	91.5
SIFT	136+83	83	91	100	76	94	80	86.9

Table 2. Image matching results for different SURF versions and SIFT with the new matching strategy. Listed are both the total detection D(s) and matching time M(s) for all test sets combined with the model sets.

a correct recognition of model objects situated either in the foreground or the background of the depicted test images.

Figures 8 and 9 show cases, where either SIFT or SURF fail to recognise the correct foreground object. Note that the goblet shown in the top row of Figure 8 was twice not correctly recognised by SIFT. Not a single match was found on the object itself, but many on the enclosing showcase. However, many model objects contained in the database are enclosed in showcases and can thus lead to false matches when it comes to the recognition of the foreground object of interest.

Figure 9 (left) shows a case where only SURF produces a false recognition. Notice that many false matches were found between the object of interest and a background object that is not part of the model database. Hence, test objects can be falsely recognised due to model images that contain similar arbitrary background objects that are not part of the objects of interest.

Finally, Figure 9 (right) shows a successfully recognised object. In that specific case, the background information was helpful for the recognition of the object.

4 Discussion and Conclusion

In this paper, we described the functionality of an interactive museum guide, which allows to robustly recognise museum exhibits under difficult environmental conditions. Our guide is robust to scale (SURF, U-SURF) and rotation (SURF). Changes of the viewing angle are covered by the overall robustness of the descriptor up to some extent. This museum guide is running on standard low-cost hardware.

4.1 Object Recognition

With the computational efficiency of SURF, object recognition can be performed instantaneously for the 20 objects on which we tested the different schemes. The

images were taken with a low-quality webcam. However, this affected the results only up to a limited extent. Note that in contrast to the approach described in [5], all the tested schemes do not use colour information for the object recognition task. This is one of the reasons for the above-mentioned recognition robustness under various lighting conditions. We experimentally verified that illumination variations, caused by artificial and natural lighting, lead to low recognition results when colour was used as the only source of information.

The fact that our model images include background information can be helpful for the recognition of objects. Especially in cases where the objects of interest are too similar or do not provide enough robust and discriminant features, background information may allow to recognise the object successfully. However, if a dominating background object is present in the test image, our recognition methods find more matches on the object in the background rather than on the one in the foreground and this leads to a false recognition, see Figure 6.

4.2 Automatic Room Detection

With a larger number of objects to be recognised, the matching accuracy and speed decrease. Also, additional background clutter can enter the database that may generate mismatches and thus lead to false detections. However, in a typical museum the proposed interactive museum guide has to be able to cope with ten-thousands of objects with possibly similar appearance. A solution to this problem would be to determine the visitor's location by adding a Bluetooth receiver to the interactive museum guide that can pick up signals emitted from senders placed in different exhibition rooms of the museum [9]. This information can then be used to reduce the search space for the extraction of relevant objects. Hence, the recognition accuracy is increased and the search time reduced. Moreover, this information can be used to indicate the user's current location in the museum.

5 Acknowledgements

The authors gladly acknowledge the financial support provided by the Toyota corporation. We also gratefully acknowledge the support by the Swiss National Museum in Zurich, Switzerland.

References

1. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: ECCV. (2006)
2. Kusunoki, F., Sugimoto, M., Hashizume, H.: Toward an interactive museum guide with sensing and wireless network technologies. In: WMTE2002, Vaxjo, Sweden. (2002) 99–102
3. Burgard, W., Cremers, A., Fox, D., Hähnel, D., Lakemeyer, G., Schulz, D., Steiner, W., Thrun, S.: The interactive museum tour-guide robot. In: Fifteenth National Conference on Artificial Intelligence (AAAI-98). (1998)

4. Thrun, S., Beetz, M., Bennewitz, M., Burgard, W., Cremers, A., Dellaert, F., Fox, D., Hähnel, D., Rosenberg, C., Roy, N., Schulte, J., Schulz, D.: Probabilistic algorithms and the interactive museum tour-guide robot minerva. International Journal of Robotics Research **19**(11) (2000) 972–999
5. Föckler, P., Zeidler, T., Bimber, O.: Phoneguide: Museum guidance supported by on-device object recognition on mobile phones. Research Report 54.74 54.72, Bauhaus-University Weimar, Media Faculty, Dept. Augmented Reality (2005)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints, cascade filtering approach. International Journal of Computer Vision **60**(2) (2004) 91–110
7. Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2004) 506–513
8. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. PAMI **27**(10) (2005) 1615–1630
9. Bay, H., Fasel, B., Van Gool, L.: Interactive museum guide. In: The Seventh International Conference on Ubiquitous Computing UBICOMP, Workshop on Smart Environments and Their Applications to Cultural Heritage. (2005)
10. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition. (2001)
11. Baumberg, A.: Reliable feature matching across widely separated views. In: Computer Vision and Pattern Recognition. (2000) 774–781
12. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. In: Computer Vision and Pattern Recognition. Volume 2. (2003) 257–263

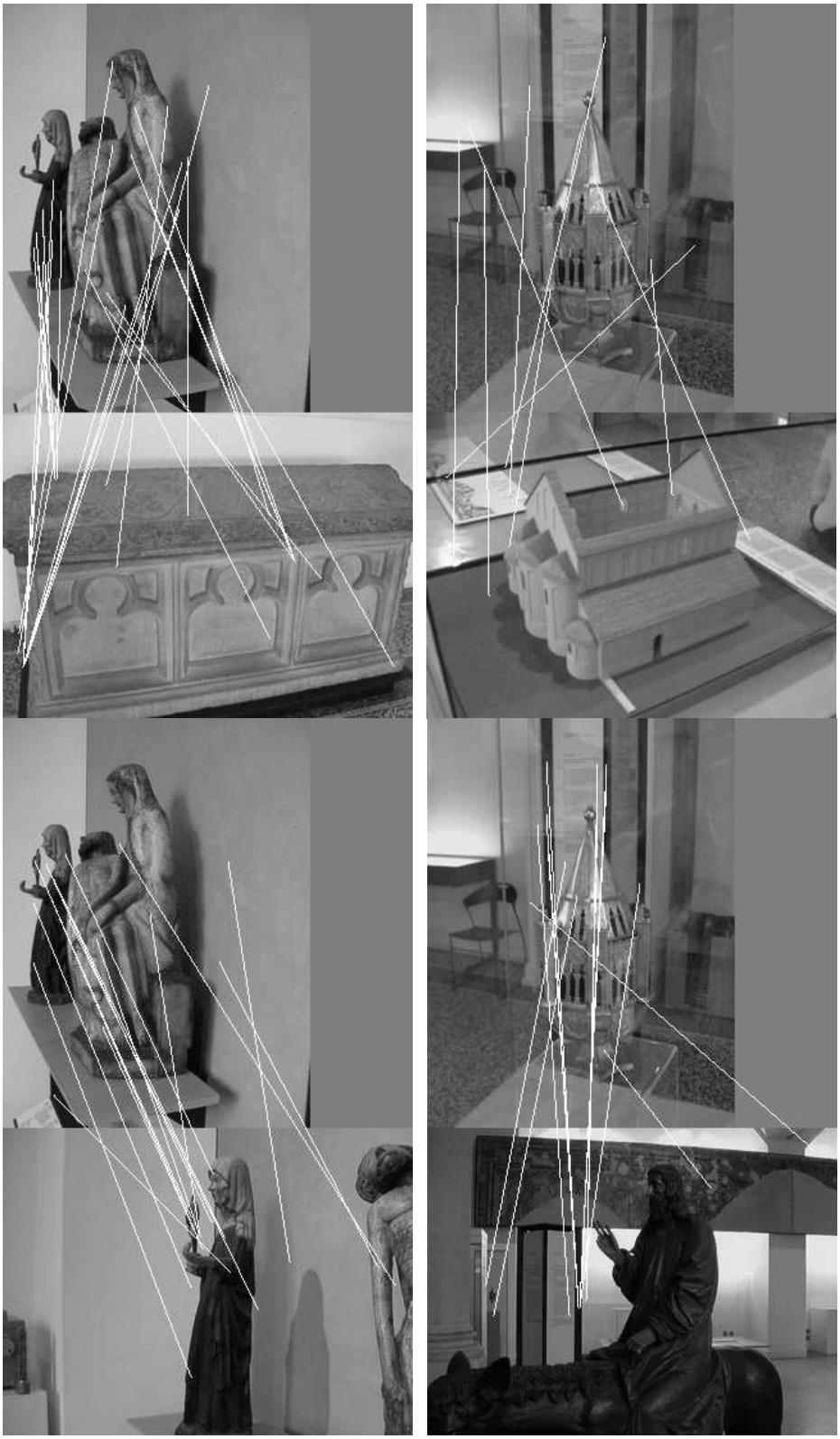

Fig. 6. Common image matching mistakes. Both SIFT (top row) and SURF (bottom row) fail to recognise the same test objects. In each of the four two-image combinations, test images are shown on the top and matched model images on the bottom.

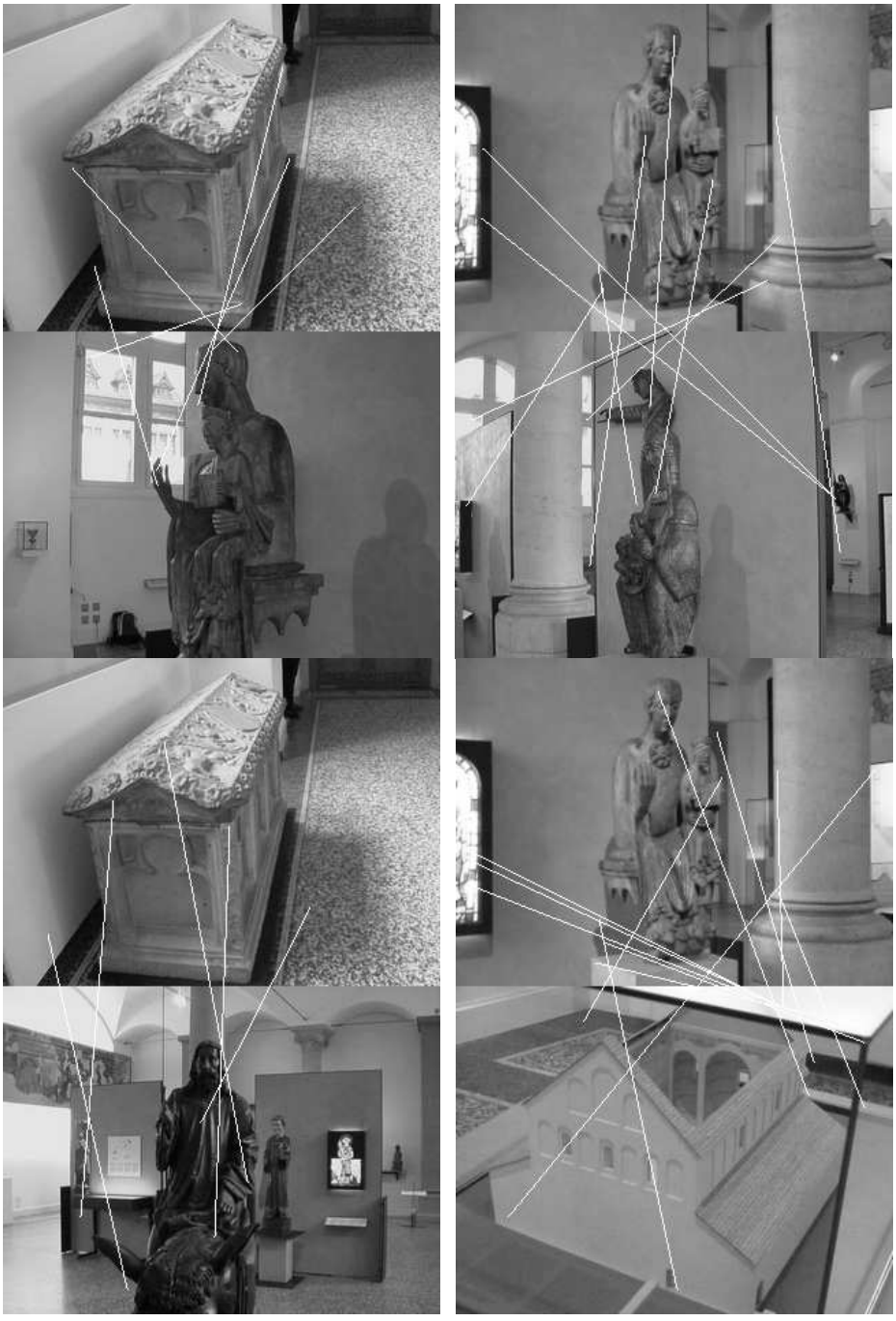

Fig. 7. Common image matching mistakes. Both SIFT (top row) and SURF (bottom row) fail to recognise the same test object. In each of the four two-image combinations, test images are shown on the top and matched model images on the bottom.

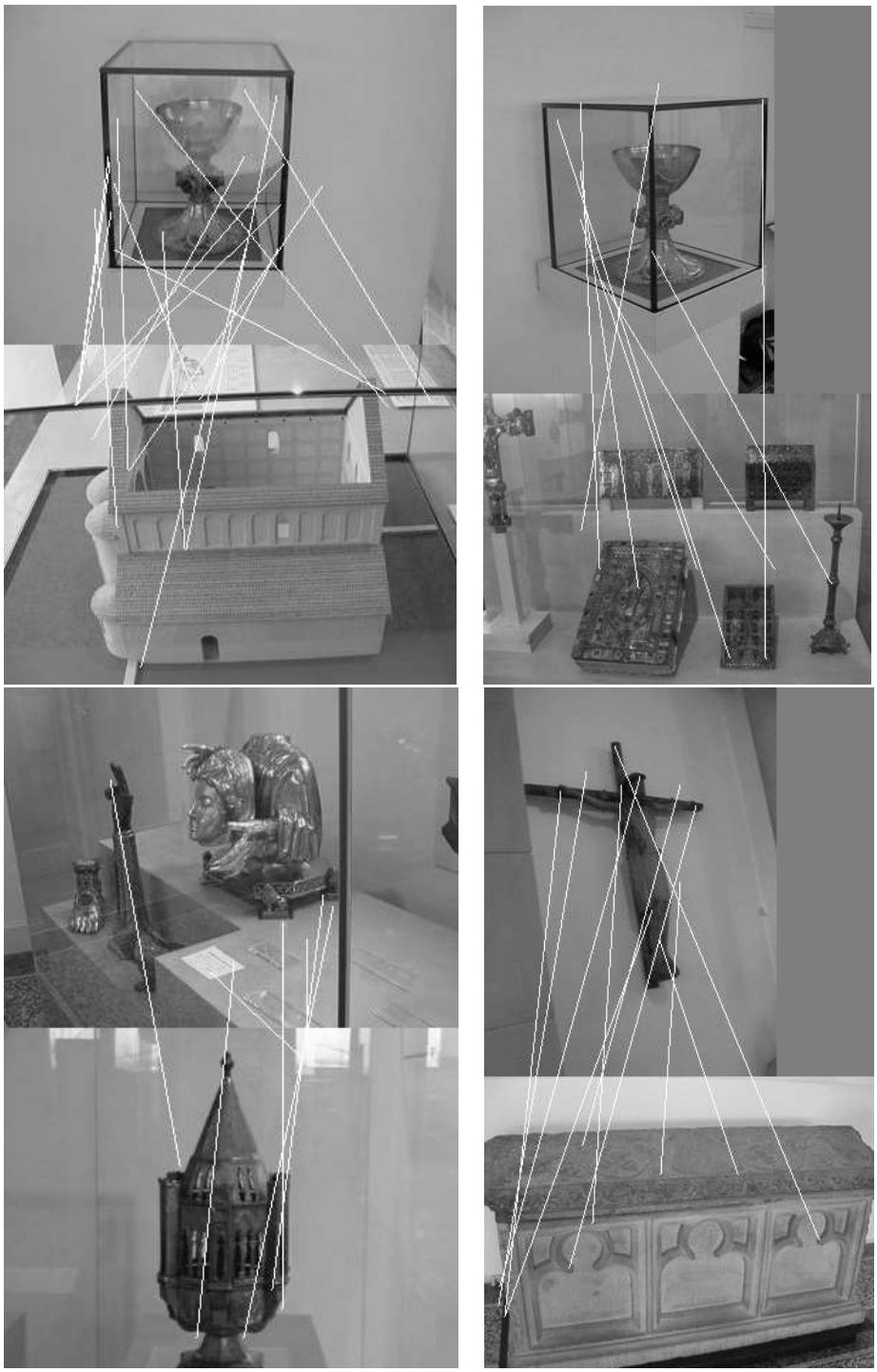

Fig. 8. Individual image matching mistakes produced by SIFT. In each of the four image combinations, test images are shown in the top row and the matched model image in the bottom row.

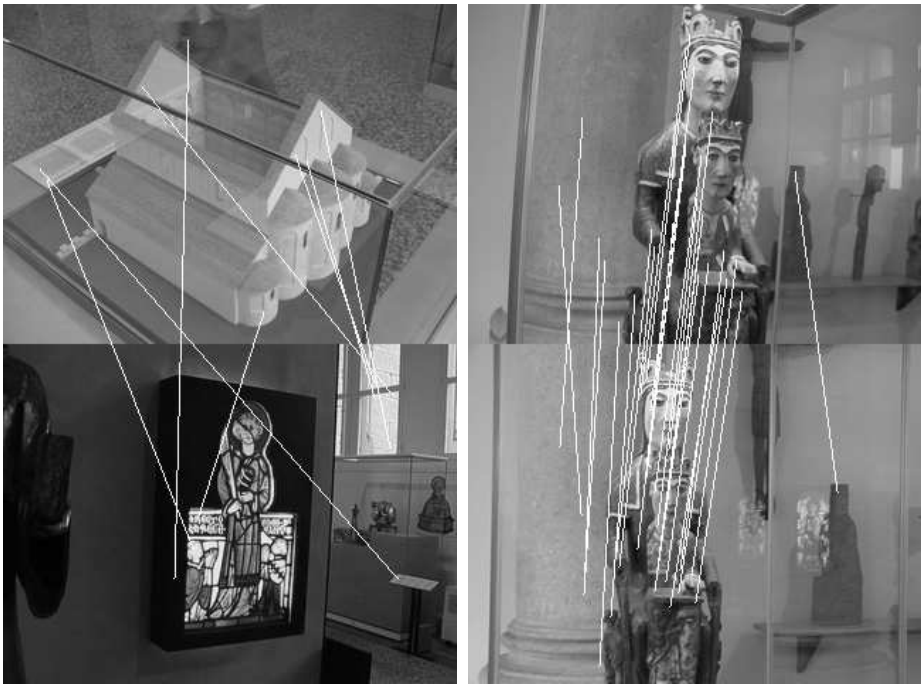

Fig. 9. Individual image matching mistake produced by SURF (left) and a successfully recognised object (right). The test image is shown on the top and the matched model image on the bottom.