Incremental models based on features persistence for object recognition

Pattern Recognition Letters

Authorship Confirmation

Please save a copy of this file, complete and upload as the "Confirmation of Authorship" file.

As corresponding author I, Marinella Iole Cadoni, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature Marinella Cadoni Date 23-05-2018

List any pre-prints:

Relevant Conference publication(s) (submitted, accepted, or published):

Justification for re-publication:

Research Highlights

- A novel view-based object recognition method is proposed
- An incremental model of an object based on SIFT features is constructed
- The new concept of features persistence is used to refine the model
- Good recognition rates are achieved on a dataset of varied cultural objects

Pattern Recognition Letters
journal homepage: www.elsevier.com

Incremental models based on features persistence for object recognition

Marinella Cadoni[a,**], Andrea Lagorio[a], Enrico Grosso[a]

[a] *University of Sassari, Computer Vision Laboratory, Porto Conte Ricerche, Alghero, Italy*

ABSTRACT

Object recognition has regained a high level of attention in recent years, with the application of deep convolutional neural networks to classification tasks. However, the problem of recognising objects for which a limited number of images is available is still open. In this paper, we propose a view-based object recognition method which can deal with objects represented by a few images. To build a model of the object, salient points are extracted from the images and a persistence value is defined for each point and updated as new images are added. The model that describes the object is refined on the basis of points persistence, where points with high persistence have priority over low persistency ones. The method is validated on a collected a dataset of objects of cultural interest. Recognition rates reach 86.4% at rank one.

1. Introduction

Object recognition is the subject of a new sprout of interest in Computer Vision and Artificial Intelligence. As discussed in the survey by Han and Vasconcelos (2014), recent advances in computational neuroscience have been often encoded in novel object recognition/detection models, such as the HMAX of Riesenhuber and Poggio (1999); Serre et al. (2007); Mutch and Lowe (2008), the convolutional networks of Pinto et al. (2008), Jarrett et al. (2009) and a number of deep learning models (Hinton et al. (2006); Krizhevsky et al. (2012)). These models outperform other detection algorithms in tasks like locating cars or pedestrians in an image, and even humans in classifying objects into fine-grained categories. However, they suffer from major drawbacks that make a completely unsupervised solution out of reach. As stated in Weber et al. (2000), three main problems arise: - Which objects are to be recognized and where do they appear in the training images? - Which object parts are distinctive and stable? - What are the parameters of the global geometry or shape and of the appearance of the individual parts that best describe the training data? While variations in position (of the object in the image) and viewpoint (of the observer) are dealt with very well by CNNs (van Noord and Postma (2017)), size variations pose a particular challenge (Xu et al. (2014)). Moreover, while in the initial layers of the CNN the visual features usually correspond to oriented edges or color transitions, in deep layers they are typically more complex and the extraction of features from the deep structure may not be appropriate in order to identify stable parts or representing shapes. Another drowback is the requirement of a large number of templates (image examples) and computational resources for training. While computational requirements are sometimes met by graphic processing units (GPUs), DSPs, or other silicon architectures, it is often necessary to describe an object (or a class of objects) from a very limited number of images. In this case the performance of deep learning and CNNs drastically drops, and the methods are of no use.

To also deal with objects for which we do not have a sufficient number of images for training, we look amongst earlier approaches to object recognition, and in particular, at view-based approaches, which represent an object as a collection of 2D views, sometimes called “aspects” or “characteristic” views (Koenderink and van Doorn (1979)). The advantage of this kind of approach is that it does neither require constructing a 3D model of an object nor making 3D inferences from 2D features (infering the depth from 2D features). View-based models can be generated both starting from features (Pope and Lowe (1993)) or images (Murase and Nayar (1995)). Nevertheless, the lack of abstraction from raw image data to the model means that the model essentially defines a set of specific object instances; therefore, these approaches pay the price of high sensi-

**Corresponding author.
e-mail: maricadoni@uniss.it (Marinella Cadoni)

tivity to lighting conditions, perspective transformations (translation, rotation, depth), occlusions etc.

In feature generated view-based model, the choice of 2D features plays a crucial role. Indeed, in order to perform tasks of visual recognition or indexing, salient points should be related to objects and space-variant transformations should be taken into account. This approach has been investigated by various researchers. For example Lindeberg (1993) has based this selection process on a quantitative analysis of gray-level blobs in scale space (thus trying to identify points that maintain at different scales similar relevance), Wiskott et al. (1997) proposed the use of Gabor wavelet jets to extract salient image features and the creation of grid-like planar graphs that (if coarsely aligned in scale and image rotation) can be compared by elastic graph matching techniques. Siddiqi et al. (1999) extended this approach to multiple scales introducing saliency map graphs (SMG). This representation turned out to be highly invariant to translation, rotation and scaling, and of practical use for occluded object recognition.

Building on view based models with a focus on the selection of 2D features, we propose the construction of a model based on salient points that persist throughout the views. The model is constructed iteratively, starting with the selection of salient points from a view and building up as more views are added to it. Salient points are extracted using SIFT descriptors on each single view and a persistency value is assigned to each point and it is updated each time a new image goes into the model. After all images are processed, the model is refined based on persistency values: points with high persistency are kept in the model, while those with low persistency are discarded. Objects are therefore represented by models made of persistent SIFT points. A sigle image of the object, or another model made up of images, can then be compared to the model for matching. The novelty of the proposed method resides in the construction of an incremental model that retains significant points, invariant to scale-space transformations, from different views, so a higher level of abstraction is reached, as some of the 3-dimensional nature of the object is retained. As we will see, the models prove to be robust enough for object recognition tasks. The method is validated on a database of cultural heritage objects, mostly made up of images collected from the web. A baseline experiment was also carried out using deep CNN trained on ImageNet, without any subsequent fine tuning due to the scarcity of images per object in the used dataset. The remaining of the paper is organized as follows: in section 2 related works that make use of SIFT descriptors for object recongnition or traking are summarized; in section 3 the construction of the model is detailed and in section 4 the experiments carried out to validate the method are illustrated.

2. Related work

Several works in literature have proven the efficacy of SIFT descriptors for object recognition (Lowe (2001)), face recognition (Cadoni et al. (2016)) and object tracking (Zhou et al. (2009)). In particular, the method we propose has some analogies with the one in Lowe (2001) where a model of an object is generated starting from a number of images of the object. The starting model consists, as in our work, of the SIFT points of the first image. The model is then compared to the next input image. If there are enough correspondences, the SIFT points coordinates are transformed so that model and image SIFT points all belong to the same coordinate system and their relative position is used to discard false correspondences. The model is then updated with the image SIFT that did not match with any SIFT of the model, while the ones that have close correspondences in the model are discarded. So in Lowe (2001) the model is enriched with several different SIFT points from different images, while in our proposed method we look for SIFT persistence which we think leads to highly meaningful key-points.

A similar key-point persistence is explored in Sabatta (2008), where a topological mapping is built for autonomous robot applications. SIFT points are extracted from panoramic images of the ambient around a robot, and a dynamic array of SIFT points is used to build a topological map of the ambient. The localization of the robot is obtained by comparing SIFT points extracted in real time with those of the ambient.

Another relevant work is that by Liu et al. (2011), in which a SIFT Flow is defined similarly to an optical flow with the difference that in the SIFT flow the matching is done on the SIFT calculated on each pixel, rather than on the pixels themselves. Although in this work no model accretion is considered, we would like to highlight the parallel between SIFT flow and persistence as the latest has the effect of tracking highly relevant points.

3. Model construction and matching

In this section the construction of the model of an object is illustrated. The model is first generated from the available images and is then refined on the basis of the persistence of the salient points selected from each image.

3.1. Model generation

The model of an object can be generated from a variable number of images, starting from one. So assume we have $I_1, \ldots, I_n$ images, with $n \geq 1$, of an object O. From each image I_k, $k = 1, \ldots, n$, we extract the SIFT points using the method by Lowe (2004). Let us denote by $S_k = \{s_{k1}, \ldots, s_{ki_k}\}$ the set of SIFT key-points extracted from image I_k. For each SIFT point s_{kj} of image I_k, $j = 1, \ldots, i_k$, the 128 entries descriptor vector v_{kj} is calculated using the method in Lowe (2004), so we get a set $V_k = \{v_{k1}, \ldots, v_{ki_k}\}$ of descriptor vectors that are in a one to one correspondence with the the SIFT points in the set S_k. For each SIFT point s_{kj}, we also initialise the persistence value ρ_{kj} to zero, and save all these values in a set $P_k = \{\rho_{k1}, \ldots, \rho_{ki_k}\}$. So each image I_k is converted into a triplet of sets $\tilde{I}_k = \{S_k, V_k, P_k\}$ containing the SIFT points, their descriptor vectors and their persistence values. After converting all images, one of them is taken at random to initialize the model, for simplicity of notation we can assume it to be the first image $\tilde{I}_1$. At this first step the model is defined as $M_1 = \{S_{M1}, V_{M1}, P_{M1}\} = \tilde{I}_1$, with $S_{M1} = S_1$, $V_{M1} = V_1$, $P_{M1} = P_1$. If $n = 1$ and the only image is denoted by I, then the model is simply $M = \tilde{I} = \{S_M, V_M, P_M\}$.

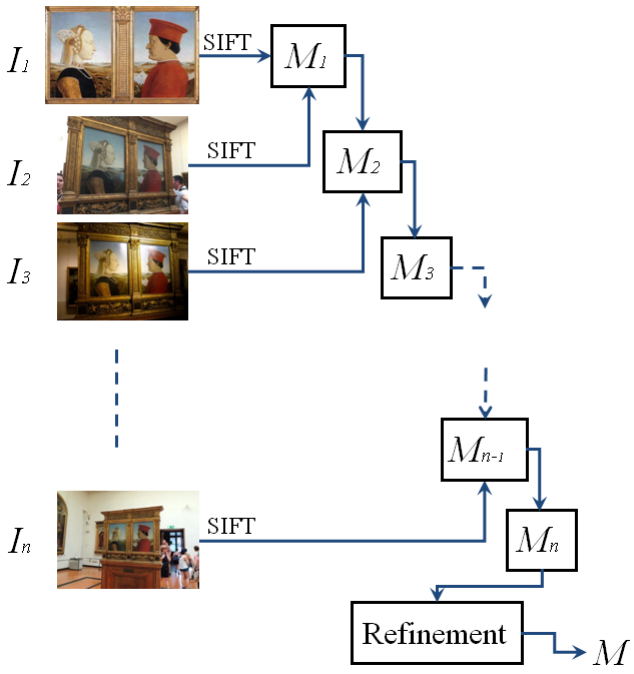

Fig. 1. Scheme of the model construction from a set of images

If $n > 1$, a second image is selected at random amongst the $\tilde{I}_j$, let us assume it to be $\tilde{I}_2$. The model M_1 is updated on the basis of the matches between the SIFT points of the model M_1 and those of $\tilde{I}_2$. Let us see how. Each descriptor vector of the set V_2 is compared to all vectors of the set V_1 using the method in Lowe (2004). Given a vector $v_{2l} \in V_2$, there are two possible outcomes:

1 v_{2l} is not matched to any vectors in V_1. In this case, the point s_{2l}, together with its descriptor vector v_{2l} are added to the model, so that now $s_{2l} \in S_1$ and $v_{2l} \in V_1$, while its persistence is inizialized to 0 ($\rho_{2l} = 0$).

2 v_{2l} is matched to a vector v_{1m} of the set V_1. We then say that the SIFT points s_{2l} and s_{1m} are corresponding (or matching) points. The point s_{1m} is kept in the model and its persistence value is updated to one, $\rho_{1m} = 1$, while the point s_{2l} is discarded.

The reason why we do not include the matching SIFT points of $\tilde{I}_2$ resides in our guiding philosophy: we are aiming to rely, whenever possible, on points that are persistent, that can be identified in several images. By using the stringent thresholds as in Lowe (2004) to compare descriptor vectors, two corresponding points can be considered equivalent, more precisely, we can think of them as belonging to the same class and choose a representative of the class, which, for ease of algorithm implementation, we take as the point that was already contained in the model, together with its descriptor vector (in this case s_{1m} and v_{1m}, respectively). In this way we keep track of persistence with the advantage that we have just one point to compare in successive steps, rather than all the class members.

After comparing all descriptor vectors of V_2 to those in V_1, the model M_1 deriving from the first image is updated to $M_2 = \{S_{M2}, V_{M2}, P_{M2}\}$ according to the actions described in the previous points, so the set S_{M2} contains S_{M1} plus the SIFT points of V_2 that had no a correspondence in S_{M1}, the set V_{M2} contains V_{M1} plus the descriptor vectors of the points added to S_{M1}, and the set P_{M2} has the updated persistence values relative to the points of S_{M1} plus the persistence values (initialized to zero) of the newly added points. The model is updated in the same way up to the last image I_n, and the final output is $M_n = \{S_{Mn}, V_{Mn}, P_{Mn}\}$.

3.2. Model refinement

The number of SIFT points contained in the model M_n, i.e. the cardinality of the set S_{Mn}, cannot be established a priory, since, other than on the number of images, it depends on various factors such as how the images are related (e.g. if and how they overlap), their quality etc. In any case, the number of SIFT points will, on average, increases with the number of input images. We therefore need to ensure that the final model is of a manageable size while containing the most relevant information, but there is also another reason to remove SIFT points that are not particularly significant. Indeed, suppose we are comparing a SIFT point of a probe image to those of a model which actually contains the same SIFT point, and the best match is the one between those two points. The more SIFT points the model has, the more are the chances that the difference between the errors of the first and the second best match is so little that the best match is not considered to be robust and it is therefore dropped (see Lowe (2004) for more details on the threshold for comparing SIFT descriptors). To overcome these difficulties, we propose to refine the model on the basis of persistence: points with higher persistence will have priority over those of lower persistence, so if a model has a sufficient number of points with high persistence, we will use them to describe it, discarding the ones with lower persistence. On the other side, if a model contains only points with low persistence, which can happen when the images that generated the model are taken from distant viewpoints, or the object has great curvature variations so that even a modest change of viewpoint can cause occlusions etc., all points are preserved. In practice, we set two thresholds, θ_1 and θ_2 so that if a model has at least θ_1 SIFT points with persistence $\rho \geq 1$, up to θ_2 points with highest persistence are selected, and all the other points are discarded. On the other hand, if a model has a number of SIFT points with $\rho \geq 1$ that is less than θ_1, no point is discarded, and the model remains unchanged. Various thresholds were tested on a training dataset as described in section 4.2, and the best ones proved to be $\theta_1 = 20$ and $\theta_2 = 100$, so that is what we set for the experiments we carried out.

The model after refinement is simply called M, where $M = \{S_M, V_M, P_M\}$ with $S_M \subseteq S_{Mn}$, $V_M \subseteq V_{Mn}$ and $P_M \subseteq P_{Mn}$. The whole process of the model construction is described in figure 1: the first image I_1 generates the first model M_1, which is updated with the SIFT from the second image and so on, up to the last image I_n. The resulting model M_n is refined to obtain the final model M.

3.3. Image model comparison

Given an object O, the generated model M_O represents the object and will be used to recognise it. Let I_P be an image of the object that we assume was not given as input for the generation of the object model. The image can readily be compared to the model by comparing all SIFT points of the image to all SIFT points of the model, always using the threshold as in Lowe (2004). The number of SIFT correspondences is taken as similarity measure.

4. Experimental validation

4.1. Datasets description

To validate the proposed method, we collected two datasets: one is the the test dataset made of 55 objects of cultural or artistic value and the other is a smaller training dataset of 15 cultural objects. For the test set, the images of 42 of the 55 objects were downloaded from the web from various websites, so they were likely taken with different acquisition devices and, even within images of the same object, we can find a great variety of sizes (from 50KB to 4MB), resolutions and time spans between acquisitions, which can vary from a few minutes up to several years. All these factors can have an effect on the looks of the objects. For instance, in a long time span a building facade might have been cleaned/restored, while meteorological conditions such as snow can visibly alter a monument. Even the different times of day at which images are taken add complexity both for outdoor objects and for indoor ones, as the SIFT points detection is not robust to great light incidence differences when object have a variable curvature. To include some controlled object images into the test database, we acquired 6 objects with a Nikon D90 digital camera on a tripod, placing the objects on a rotating plate, and acquiring an image for every plate rotation of 10° starting from 0^{circ} and up to 360°. Five other objects were acquired using an iPhone 7 mobile phone, keeping the object fixed and moving the phone around it while taking pictures. Two videos of two objects were also taken, again turning the phone around the objects, and a selection of video frames has been used to build the models of the objects.

For the training dataset, all images were downloaded from the web.

In table 1 we can see the composition of the two datasets according to object classes (or object typology):

1. Interiors: the set includes frescos over architectural structures such as ceiling (Sistine Chapel by Michelangelo) or walls (Last Supper by Leonardo), so they can have large curvature variations or they can be planar.
2. Monuments: the set includes outdoor monuments such as churches (Pisa Cathedral, San Vigilio etc) or monuments such as Petra, the Parthenon etc.
3. Small-medium objects: includes cultural objects such as statuettes (such as the Venus of Willendorf), terracotta or ceramic vases, ship models and so on.
4. Paintings: includes paintings and altarpieces (Botticelli's Venus, Miro etc). Due to their planar nature they are the easiest objects to be recognised.

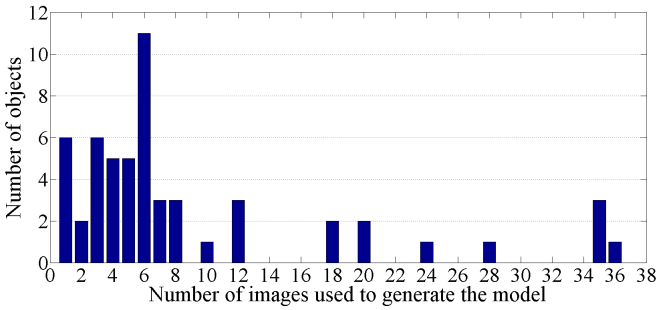

Fig. 2. Histogram of number of images that generate the models

5. Statues: includes mainly marble statues (Moses and David by Michelangelo, Love and Psyche by Canova) and some bronze ones (Dancer by Degas). They are amongst the most difficult object to recognise because of their uniform texture and variable curvature, so SIFT can easily locate different points under light changes, and even if they are able to locate the same point in different images, the descriptor vectors will likely be too different for them to match.

The number of images per object in the test set is variable (from 3 to 40) and the total number of images is 593, while in the train set there are 9 images for each object for a total of 135 images.

The train set was partitioned into a gallery set, consisting of 6 images per object and was used to generate the model, and a probe set, consisting of 3 images per object. The test set was partitioned in a similar way: when available, 4 images were set aside for the probe set (only for three objects, namely three vases, there were less than 5 images in total and so less than four were selected as probe), giving 206 probe images in total, and the remaining images (in variable number per object) were assigned to the gallery set, for the generation of the models.

In table 2 we can see a selection of the collected images of four objects, rescaled if their size was grater than 640×480. As it can be seen from the first two rows, images of large interior elements such as Cappella degli Scrovegni (first row) and of large outdoor monuments such as the Cathedral of Pisa (second row), tend to be quite distorted due to the use of wide angle camera lenses. This clearly adds difficulty to the recognition process. Images of both objects were downloaded from the web and are different in size and resolution, as are the images of the statue of Moses in the third row. All three objects were acquired at different time of day (possibly on different days or even years), as it can be seen from the different light incidences or meteo conditions. The Lateen sail in the third row was acquired with a mobile phone camera, and it is the only object in figure whose images are taken by the same camera. The images in the column "Probe image" were used as probe for each one of the objects, and were compared to the models generated using the images of the objects in the first column. Notice that for the Lateen sail and the Pisa Cathedral, some additional images to the ones shown in figure were used to generate the model.

As mentioned earlier, each model can be generated by $n \geq 1$ images. In the collected dataset, the 55 objects do not have the

Table 1. Datasets partition according to object classes.

	Interiors	Monuments	Small-medium objects	Paintings	Statues	Total
N. of objects in test set	5	11	21*	12	6	55
N. of objects in train set	3	3	3	3	3	15

* of which 6 taken in a controlled setting

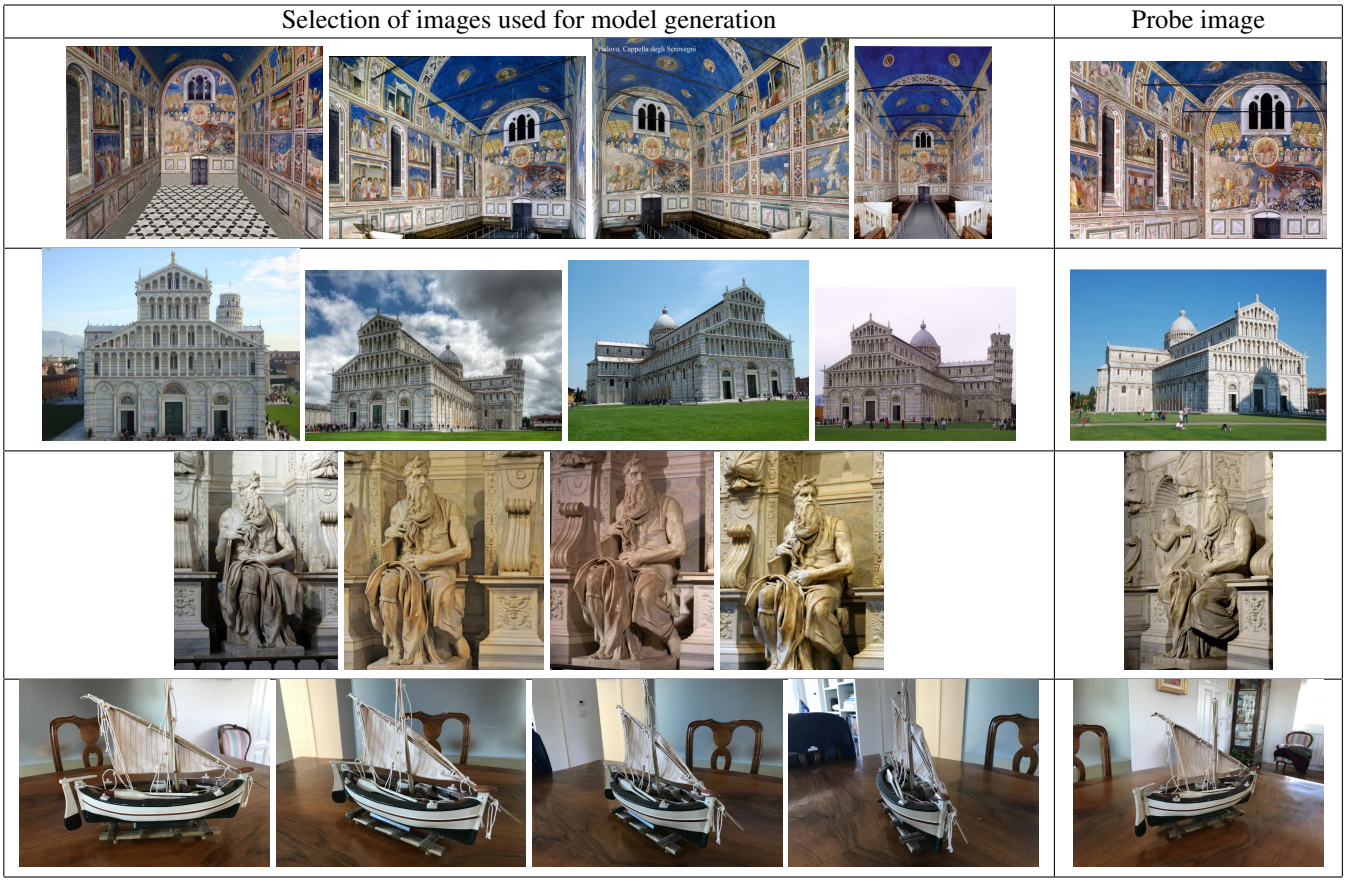

Table 2. A selection of images of four objects of the collected dataset. For each row, the last image was used as probe, the remaining to construct the object model. First row: Cappella degli Scrovegni. Second row: Pisa Cathedral. Third row: Statue of Moses by Leonardo. Fourth row: Lateen Sail.

same number of images. To see how the number of images is distributed across the objects, in figure 2 is the histogram of the number of images available for each of object. We can see that for 6 objects, the model was generated from a single image, while the majority of models was generated from 6 images. The tail of the distribution is relative to the objects acquired on the rotating plate.

4.2. Experimental protocol

The first experiment was carried out on the train set to establish the thresholds θ_1 and θ_2. First, the models were generated using the images in the gallery set (6 per object). Two values were selected for the threshold θ_1, 10 and 20, while for each of these two values, θ_2 was set to 50, 100, 150 and 200. The reason behind the choice of the values of θ_1 to test was that when there are less than 10 SIFT points in the model with persistence $\rho \geq 1$, it is reasonable to consider all SIFT points, while values of θ_1 greater than 20 led to a small interval between θ_1 and θ_2. θ_2 was tested up to 200 to limit the number of total SIFT points in the final model. For each couple of values (θ_1, θ_2), with $\theta_1 = 10, 20$ and $\theta_2 = 50, 100, 150, 200$, a test was performed in the following way. From each image of the probe set the SIFT were extracted and compared (see section 3.3) to all models in the gallery, refined using the thresholds (θ_1, θ_2). The pair of thresholds that maximised the recognition rate was $(\theta_1 = 20, \theta_2 = 100)$. Using the established thresholds, the proposed methodology was validated on the test set. First, the models of the 55 objects were generated using the images in the gallery set, and refined using the thresholds $(\theta_1 = 20, \theta_2 = 100)$. Then each image of the probe set was compared to all models and the match was chosen to be the model with the maximum number of SIFT correspondences.

4.3. Experimental results

The SIFT persistence of a model is bound by the number of images that generated it, more precisely, the maximum possible persistence of a model generated from n images is $n - 1$. In general, the persistence of SIFT points depends on a variety

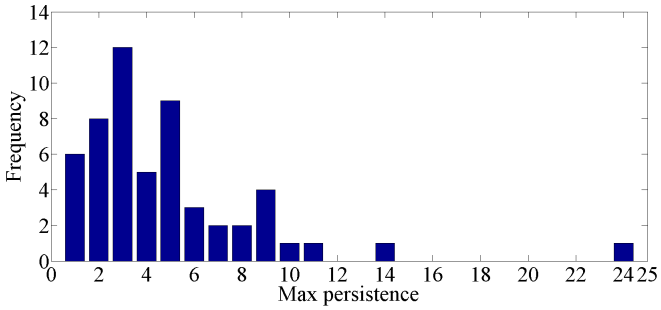

Fig. 3. Histogram of SIFT persistence across object models

of factors, such as how similar the images are, e.g. in terms of viewpoint acquisition, resolution, object appearance etc. In figure 3 we can see a histogram of the maximum persistence's in the models built from the gallery set. Most models have maximum persistence equal to 3 while the model with maximum persistence equal to 24 is due to the fixed background in the images acquired with the object on the rotating plate. Although the histogram does not take into account the number of SIFT with a given persistence, we can deduce that for at least 6 objects (the ones with maximum SIFT persistence equal to 1), all SIFT points were considered, i.e. the refinement did not change the original model.

To track down the locations of persistent points in images, we generated the models of three objects using 4 images for each model. The results are shown in figure 4 where the three models with their SIFT persistence are represented. The first image that generated each of the models is shown twice in figure: on the left of each pair, all SIFT found at the first step of the model generation (so all SIFT of the first image) are visualised by red crosses, while on the right of each image pair are visualised only the SIFT from the first image with $\rho \geq 1$, with the red crosses representing the SIFT with $\rho = 1$, the green ones the SIFT with $\rho = 2$ and the blue ones the SIFT with $\rho = 3$, which is the maximum possible persistence for a model generated with four images. As it can be seen, in all models, persistent SIFT tend to be located on details that are highly distinguishable by the human eye, and obviously significant in the scale-space representation. Notice, for instance, how most SIFT present in the background sky of the Duchi di Urbino first image are not present in the second one, which means that their persistence in the final model was zero. Furthermore, most SIFT that persist in models of non flat objects (such as the Cathedral and the Alabastron), are located on the object of interest, while almost none of the SIFT located in the background persist. Considering persistent SIFT points of images of non flat objects taken from different viewpoints somewhat induces a segmentation of the object, a desirable feature that has the effect of reducing the size of the model by removing points that are not highly significant. In the recognition experiment, the probe set contained a total 206 images and each one of them was compared to the 55 models in gallery. In total 11330 comparisons took place. The recognition rate at rank one was 86.34%, at rank two 90.73% and at rank five 91.22% (see table 3).

Table 3. Test results

	Rank 1	Rank 2	Rank 5
Recognition rates	86.34%	90.73%	91.22%

4.4. Deep CNN's Baseline Experiment

On the same dataset we collected, we the convolution neural network NasNet (Zoph et al. (2017)) trained on ImageNet (Krizhevsky et al. (2012)). The network could not be fine tuned due to the limited number of images per object. The 1001 entries vectors of each image of the same object in the gallery of the test set (the images we used to construct the models) were averaged to give a descriptor vector for each object. For each probe image, the vector was generated with the network and compared to all object vectors from the gallery. The recognition rate at rank one was 32.84%, so clearly, without fine tuning, a step which would require hundreds of images per object, the network is not able to recognise the objects.

5. Conclusions

We proposed a novel view-based object recognition method, with salient features extracted using SIFT descriptors and characterised by a persistency parameter which is used to refine the model by giving priority to the most relevant points located in more than one view. By taking into account the persistency of the points, the model retains some of the 3D information of the object, which proves to be enough to recognise ojects of different nature, from small artefacts to large buildings, with great variations in textures. The experiments carried out on a collected dataset, prove the validity of the method on images acquired with different systems, at different resolutions, and its robustness to variation of illumination, background and occlusions. Recognition rates reach 86.34% at rank one, while a baseground experiment carried out on the same database using Deep CNN trained on Imagenet had a rank one recognition rate of 32.84%. While the recognition paradigm was defined as a probe image vs a model (generated from one to several images), the process could be generalised to model vs model, when more than one image is present as a probe, and also to video, building the model from a selection of frames. These aspects, together with an improvement of the refinement of the model based on point persistency will be the subject of future investigations.

References

Cadoni, M., Lagorio, A., Grosso, E., 2016. Large scale face identification by combined iconic features and 3d joint invariant signatures. Image Vision Comput. 52, 42–55.

Han, S., Vasconcelos, N., 2014. Object recognition with hierarchical discriminant saliency networks. Frontiers in Computational Neuroscience 8, 109.

Hinton, G.E., Osindero, S., Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural Computation 18, 1527–1554.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y., 2009. What is the best multi-stage architecture for object recognition?, in: 2009 IEEE 12th International Conference on Computer Vision, pp. 2146–2153.

Koenderink, J.J., van Doorn, A.J., 1979. The internal representation of solid shape with respect to vision. Biological cybernetics 32, 211–216.

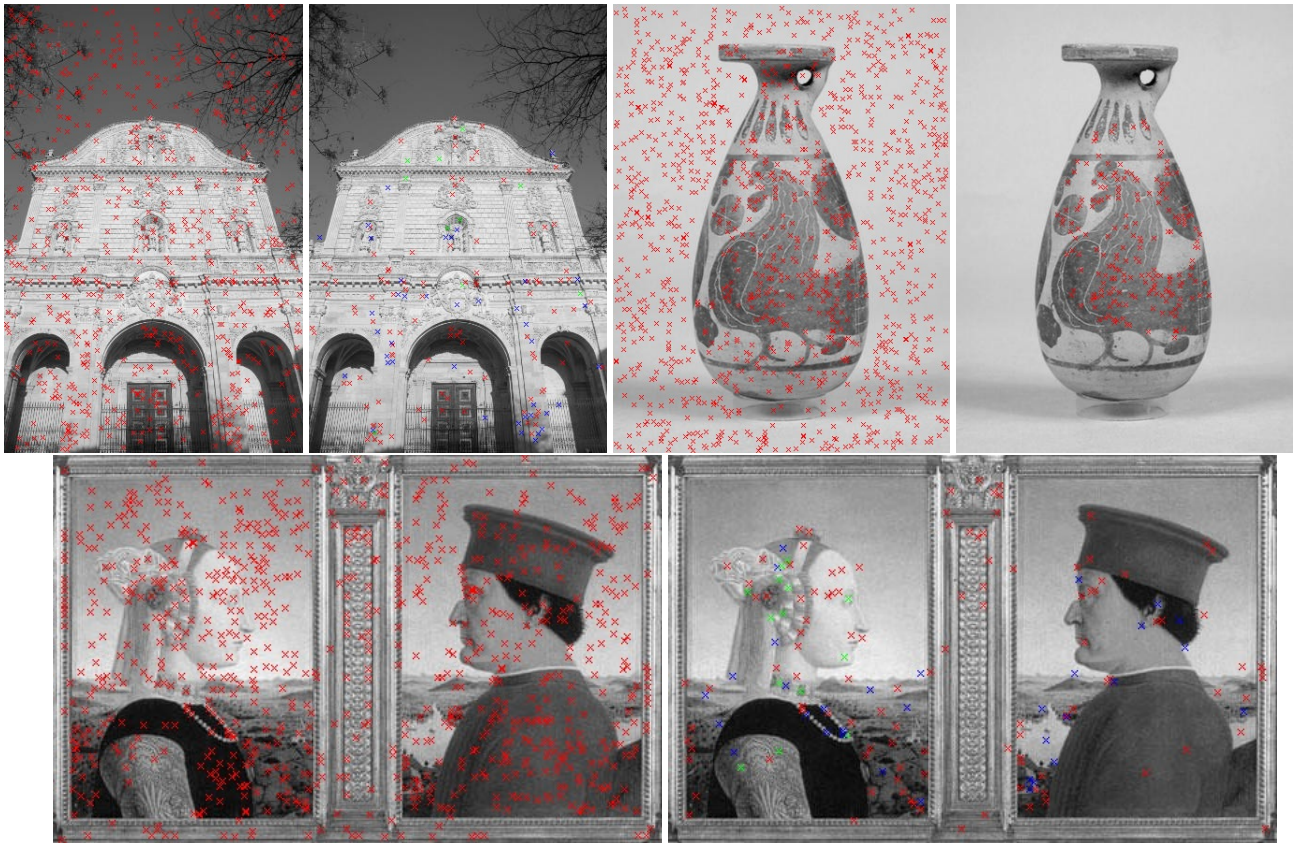

Fig. 4. SIFT persistence in models. Clockwise from top left (by image pair): Sassari Cathedral, Greek Alabastron and Piero della Francesca's Duchi di Urbino.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Lindeberg, T., 1993. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. International Journal of Computer Vision 11, 283–318.

Liu, C., Yuen, J., Torralba, A., 2011. Sift flow: Dense correspondence across scenes and its applications, in: IEEE transactions on pattern analysis and machine intelligence. volume 33, pp. 978–94.

Lowe, D., 2001. Local feature view clustering for 3d object recognition, in: Proc 2001 IEEE Comput Soc Conf Comput Vis Pattern Recogn. volume 1, pp. I–682.

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

Murase, H., Nayar, S.K., 1995. Visual learning and recognition of 3-d objects from appearance. International journal of computer vision 14, 5–24.

Mutch, J., Lowe, D.G., 2008. Object class recognition and localization using sparse features with limited receptive fields. International Journal of Computer Vision 80, 45–57.

van Noord, N., Postma, E., 2017. Learning scale-variant and scale-invariant features for deep image classification. Pattern Recognition 61, 583–592.

Pinto, N., Cox, D.D., DiCarlo, J.J., 2008. Why is real-world visual object recognition hard? PLOS Computational Biology 4, 1–6.

Pope, A.R., Lowe, D.G., 1993. Learning object recognition models from images, in: Computer Vision, 1993. Proceedings., Fourth International Conference on, pp. 296–301.

Riesenhuber, M., Poggio, T., 1999. Hierarchical models of object recognition in cortex. Nature neuroscience 2, 1019.

Sabatta, D.G., 2008. Vision-based topological map building and localisation using persistent features, in: Proceedings of the Robotics and Mechatronics Symposium, Pretoria, South Africa (2008), pp. 1–6.

Serre, T., Oliva, A., Poggio, T., 2007. A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences 104, 6424–6429.

Siddiqi, K., Shokoufandeh, A., Dickinson, S.J., Zucker, S.W., 1999. Shock graphs and shape matching. International Journal of Computer Vision 35, 13–32.

Weber, M., Welling, M., Perona, P., 2000. Unsupervised learning of models for recognition, in: European conference on computer vision, pp. 18–32.

Wiskott, L., Krüger, N., Kuiger, N., Von Der Malsburg, C., 1997. Face recognition by elastic bunch graph matching. IEEE Transactions on pattern analysis and machine intelligence 19, 775–779.

Xu, Y., Xiao, T., Zhang, J., Yang, K., Zhang, Z., 2014. Scale-invariant convolutional neural networks. arXiv preprint arXiv:1411.6369 .

Zhou, H., Yuan, Y., Shi, C., 2009. Object tracking using sift features and mean shift. Computer Vision and Image Understanding 113, 345 – 352.

Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V., 2017. Learning transferable architectures for scalable image recognition. arXiv preprint arXiv:1707.07012 .