From local to global: composition of SIFT features for image representation

Ignazio Infantino - Filippo Vella {infantino,vella}@pa.icar.cnr.it

Introduction

Image description for image representation and retrieval, has been typically achieved with global features able to characterize visual content as distribution of a density function (e.g. histogram of values). The firstly and most used features are related to the distribution of color, texture and shape in image regions. The characterization of values as statistical distribution of multiple features is based on the assumption that objects and in general category of images are related to distinguishable statistical distribution.

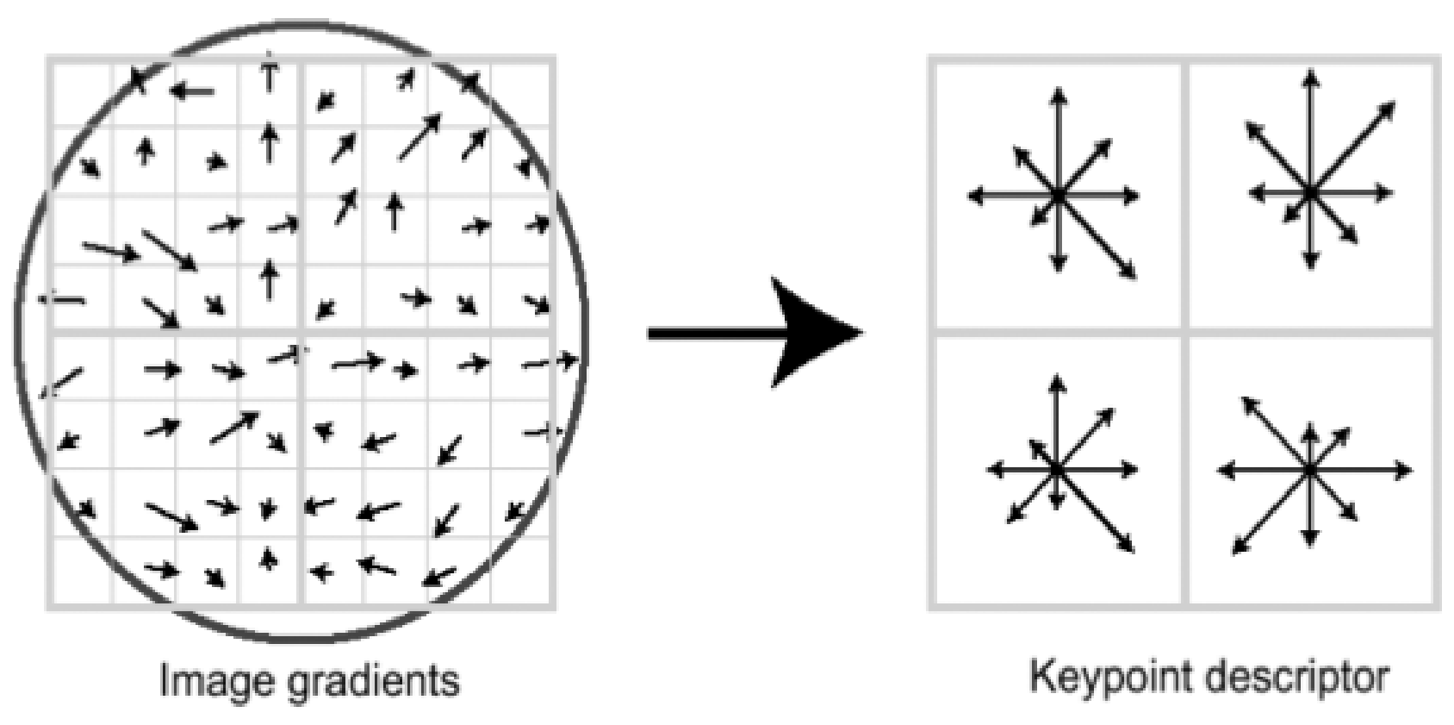
The alternative approach is to describe portion of images with local features correlated with sub parts of images and particular view of objects. In particular we consider to describe and match images with a novel features based on Scale Invariant Feature Transform (SIFT) that have been largely applied and successfully employed for local description of images.

Other works employing local descriptors to attempts to describe image content have been proposed by Csurka, Sivic and Zisserman and Hare. Csurka et al. evaluate all the SIFT features for the dataset, they build a cluster distribution with *k*-means and represent the new images counting how many features fall in the chosen clusters. These new features are used as representation to classify test objects. Authors show promising results for the classification of 7 objects. Similarly Siciv and Zisserman use SIFT descriptor clustered and used as words to apply text retrieval techniques to matching objects in keyframe of video sequences. Hare et al. applied a similar representation to employ cross language latent semantic indexing. Ke and Sukthankar adopt SIFT local descriptors to create a more general descriptor. The proposed descriptor is built considering the Principal Component Analysis to linearly project high-dimensional data, given by the SIFT descriptors, to a low-dimensional data in the principal component space. Each representation deals with a single images key point.

Here we propose a novel representation achieved composing SIFT descriptors to create more abstract descriptors aiming to generate features that are semantically nearer to scene objects and image tags. The new descriptors cover a larger region, that turns to be semantically more relevant, than the original patch of keypoint covered by SIFT descriptors.

SIFT Features

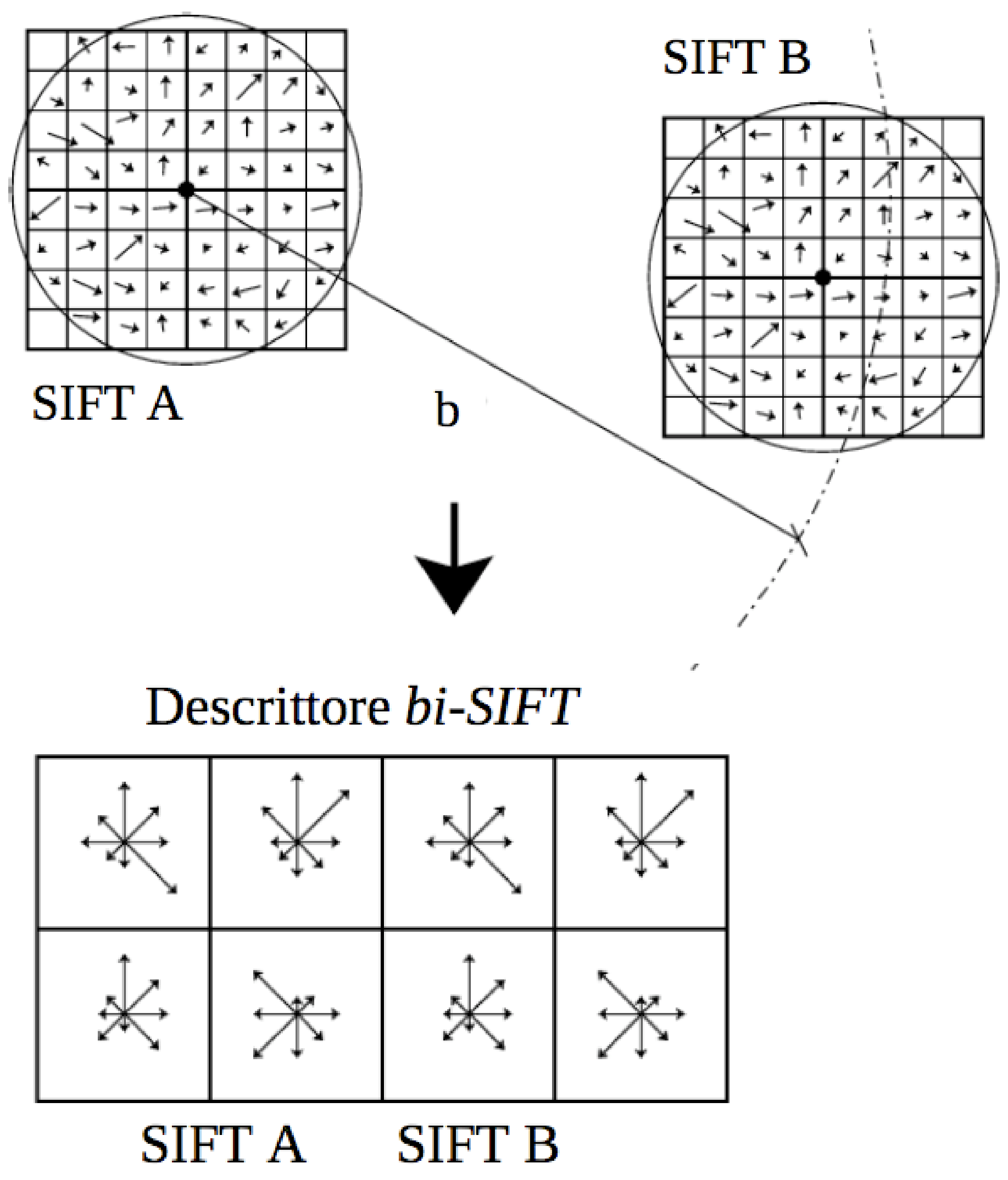
SIFT is an approach presented by Lowe for detecting stable points and extracting descriptors that are invariant against changes in illumination, image scaling, image rotation. The representation is achieved in four steps. At first, interest points invariant to scale and orientation (scale-space extrema detection) are detected spotting points that are a local maximum or a local minimum of the difference-of-Gaussian (doG) originated from input images. For each candidate point the location and orientation are evaluated (*keypoint localization*). The keypoint with low contrast and along edges are removed as they are difficult to distinguish. An orientation is assigned to each point (*orientation assignment*) evaluating the gradient orientation histograms in the neighbors of the keypoints.The modes of the histograms are considered as the dominant orientation for the given keypoints. Since the histograms are populated considering orientation referred to the largest gradient magnitude, the feature is invariant against rotations.



Each histogram contain 8 bins each and a descriptor contains 4 histograms around the keypoint. The SIFT feature is then composed by 4x4x8 = 128 elements. The normalization of histogram allows robustness against illumination changes.

Proposed Feature

SIFT are reliable features and are robust against typical variation in picture viewpoint position. Notwithstanding they are local features and cover local properties of objects and scenes. To create a feature as robust as SIFT and able to describe wider areas of images and scenes we consider the composition of a set of keypoints in a region of image. The new feature is composed taking into account the keypoints falling in a circular region centered in a keypoint and delimited by a fixed radius.



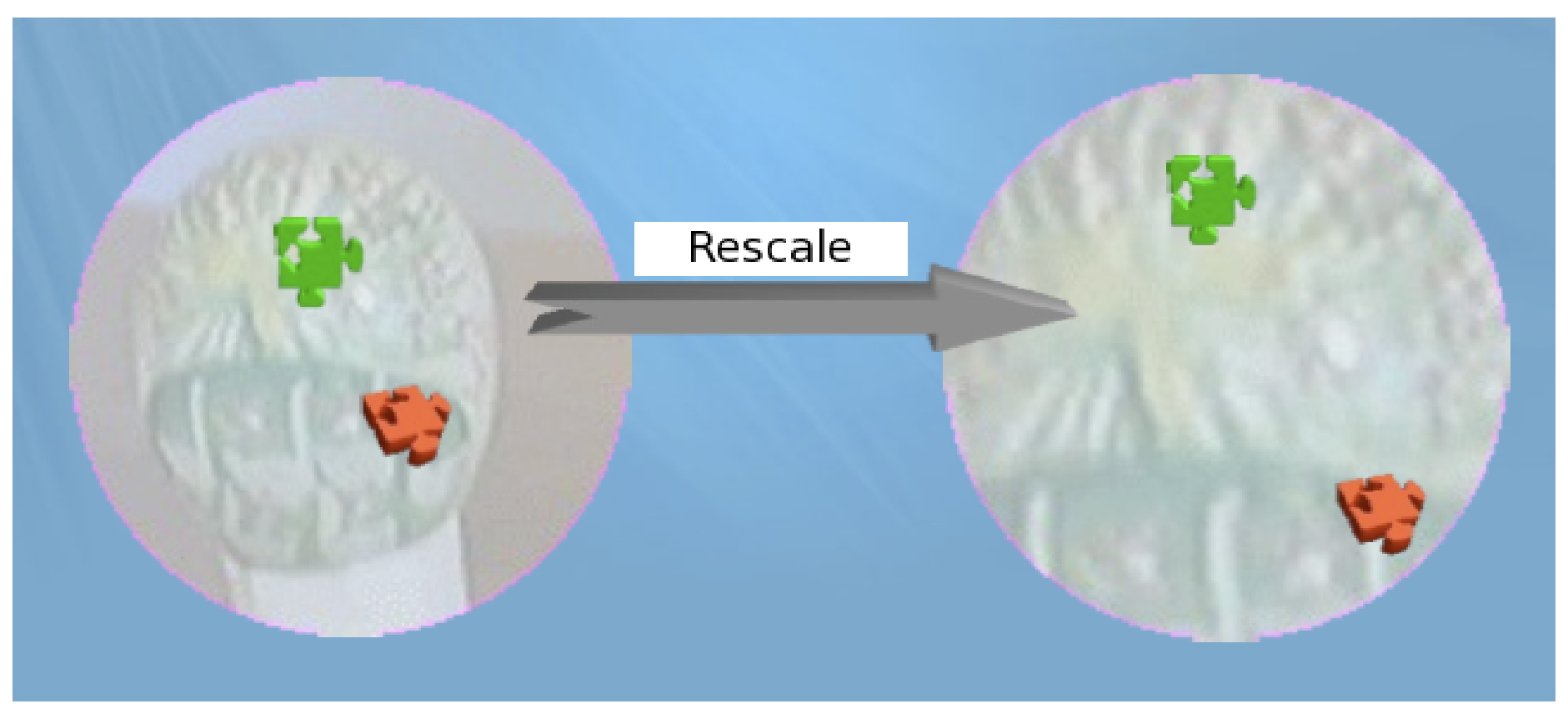
The new feature will represent in a more abstract way a larger piece of image or a complex pattern allowing to capture large portion of objects or scene invariant characteristics. The size of regions described by this novel feature (called *bi-SIFT*) is driven by an empirically fixed parameter. This parameter is called *spatial bandwidth* as it is coherent with a spatial clustering in Mean Shift theory.

The feature is built considering the keypoints falling in a region centered on a keypoint. Inside this image portion one or more keypoints can be found. If only the central point is falling in the region, meaning that the selected part of image captures a region with few relevant points, the *bi-SIFT* feature is not generated. In the other case when more points fall in the region around the keypoint, a composite description of region is considered. Not all the keypoints are taken into account as, in this case, a variable size descriptor would be created. A selection is made instead preserving the most relevant and stable information in the covered areas. The property in SIFT descriptors that most matches with stability and robustness of SIFT features against image transformation is the highest gradient magnitude that can be evaluated as the module of the main orientation in the represented image patch. The SIFT descriptors are then ordered according to their highest gradient magnitude and SIFT descriptors with highest values are retained (in this case we set this value to two). The new feature is formed by the juxtaposition of the SIFT representation of the selected points.

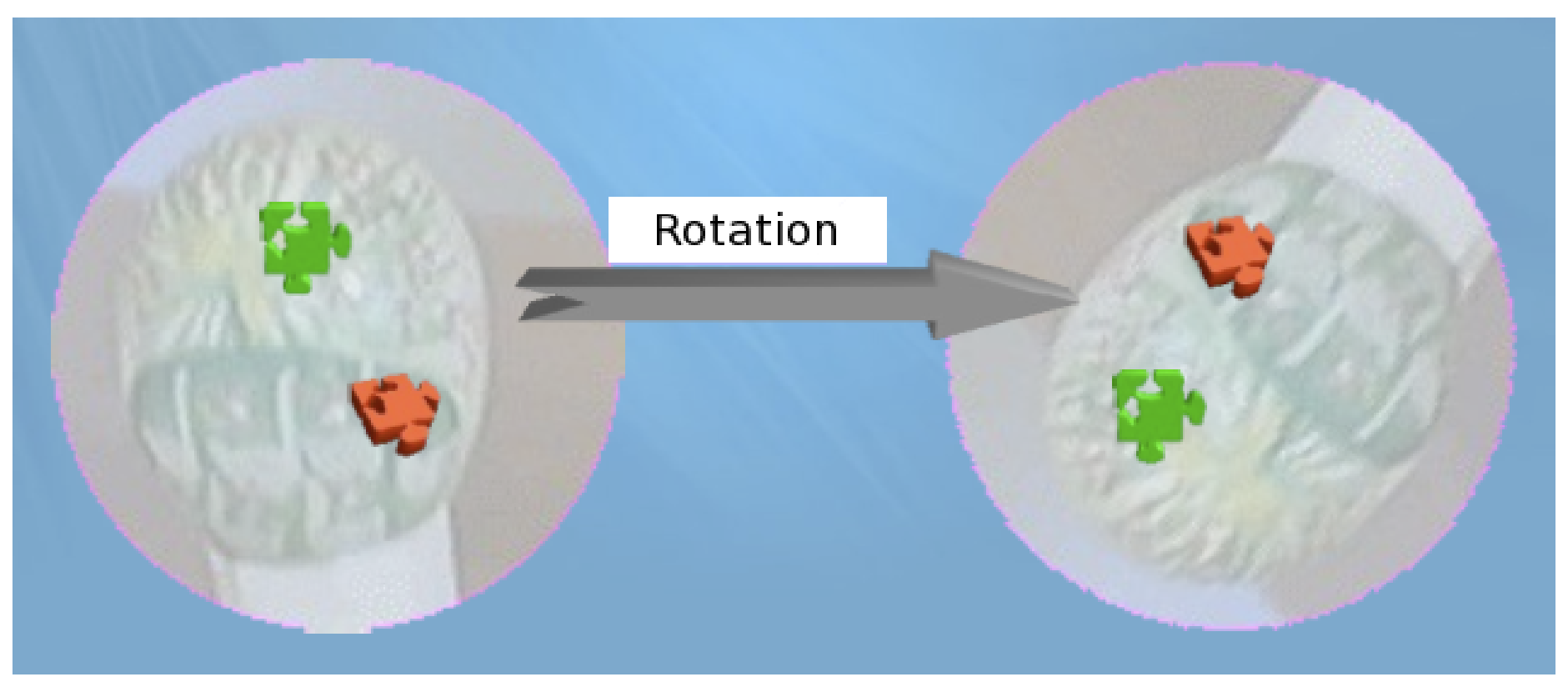
Feature Properties

This feature represents a wider area than SIFT descriptors, maintaining invariance against change of viewpoint.

For these properties, *bi-SIFT* descriptors are reliable in describing portion of objects and relevant patterns in scenes. SIFT descriptors are related to relatively small areas and if couple of points are accidentally near in an images (e.g. a point from object and a point from background) they will be greatly affected by change of viewpoint or will be difficult to retrieved in images depicting the same object in a different scene with low probability. So the recurrence of *bi-SIFT* feature assert the presence of a given object or a characteristic pattern for a given scene. Since gradient magnitude is not affected by scale transformation all change of scale mapping points inside the same spatial bandwidth, are represented with the same *bi-SIFT* feature.



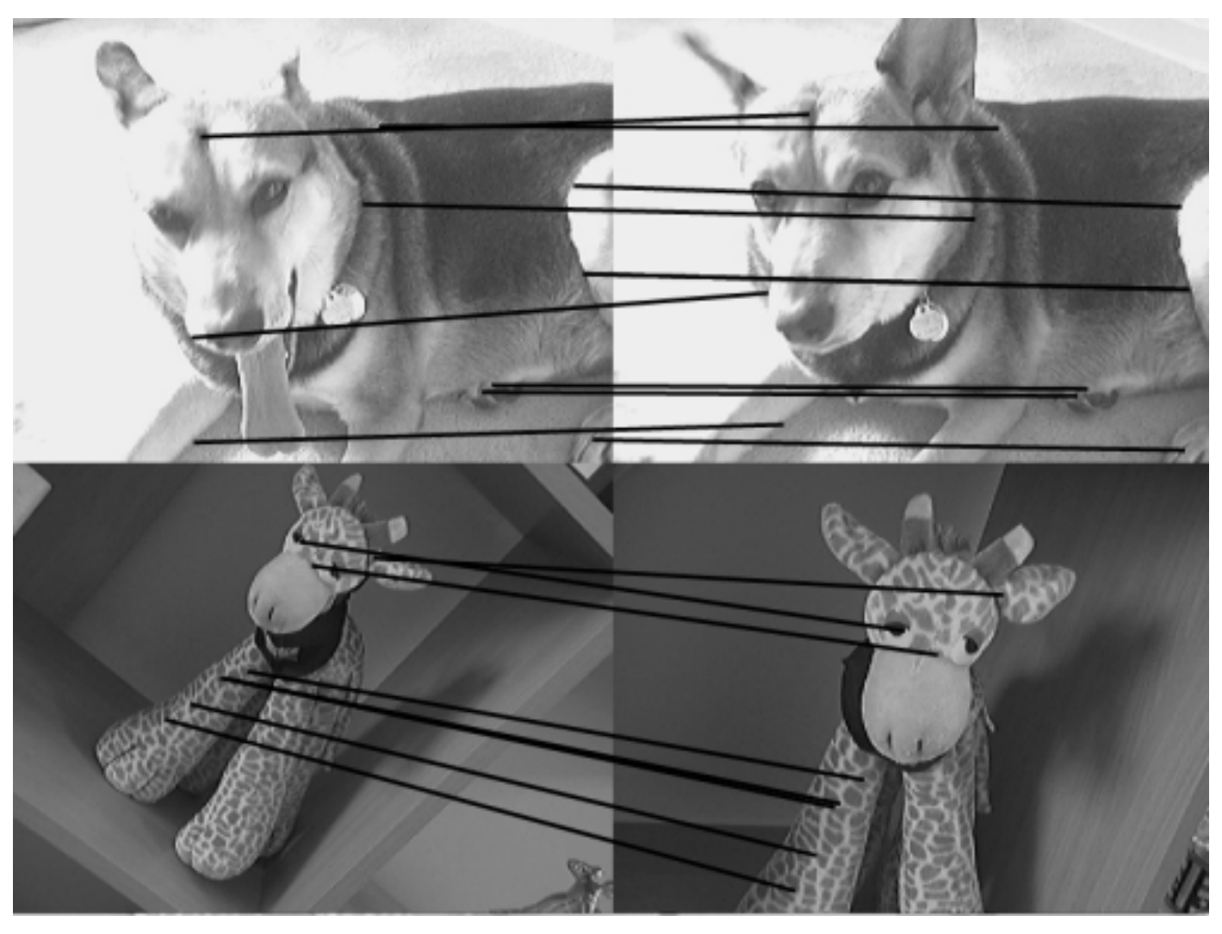
If image is rotated or the viewpoint is changed, a given region will produce an approximation of the original *bi-SIFT* descriptor.



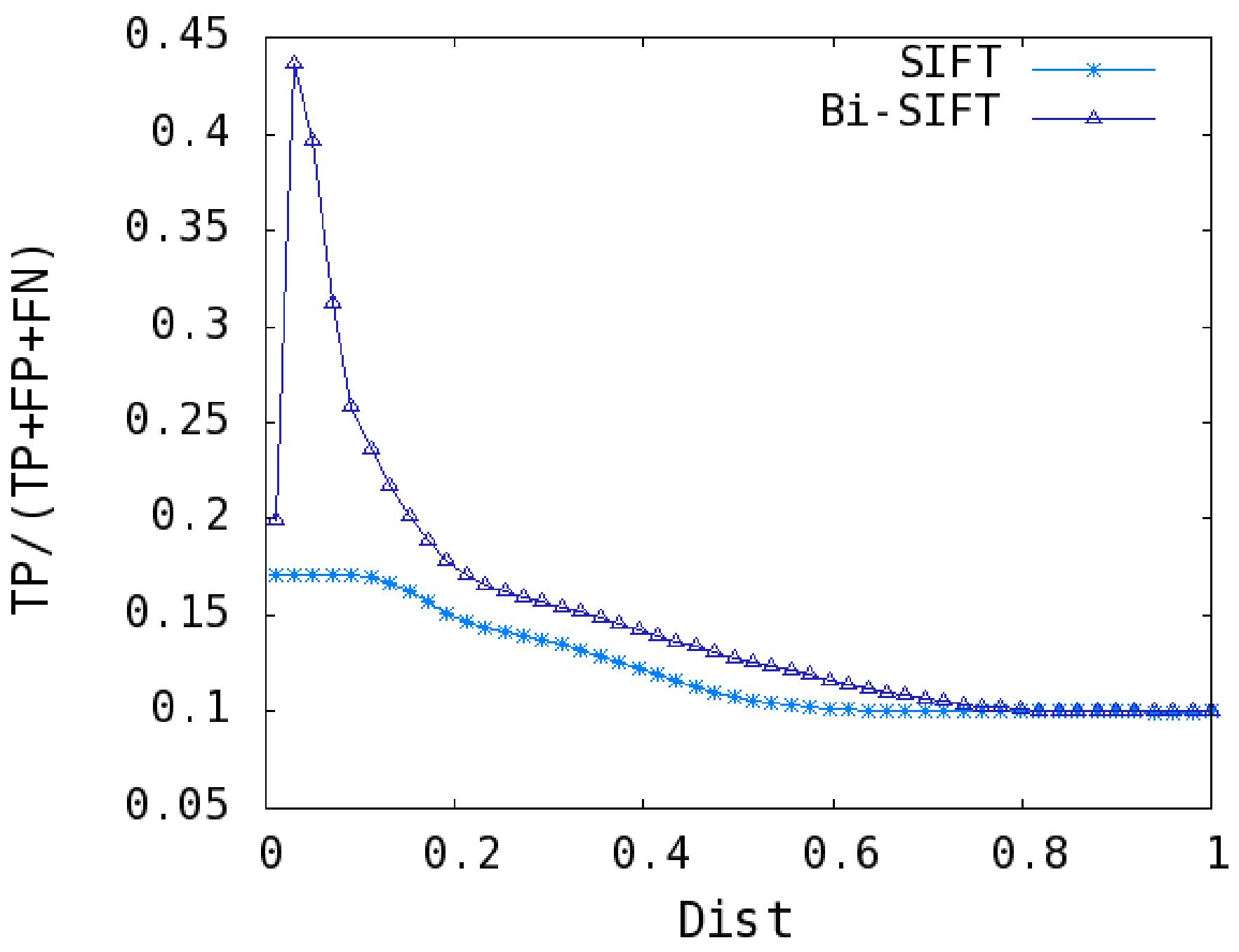
Experimental Results

The representation with *bi-SIFT* features is used to represent and retrieve images from a test data-set with 30 images capturing different objects divided in 10 classes. Classes of dataset are *bunny, camera, dog, drinks, eggs, fire, flower, giraffe, remote, timer*.

Each image depicts objects with different rotation, zooming, viewpoint. The dataset is available on line and has been used for evaluating analog SIFT based representation. We extracted from all the images the SIFT features and we built the corresponding *bi-SIFT* descriptors coupling all the SIFT features in regions around SIFT keypoints with radius of 6 pixels. Starting from a single image, we considered all the matches with the other images in the dataset and we grouped the matches according to the image classes.



The matches among images in the same class and among images in different classes allow to build a confusion matrix counting as true positive the number of matches among images in the same class, false positive and false negative the number of matches among images in different classes. To calculate a single parameter to evaluate the global matching performance, we considered the ratio of true positive upon the sum of true positive, false negative and false positive. This global value has been tested for multiple values of distances and their plot is shown in figure:



A second set of experiments has been considered for testing the matching properties of the *bi-SIFT* when the same scene is captured from different points of view. An example image with the matching got with *bi-SIFT* is shown in in figure. With the images in the dataset the transformation matrices for affine or projective transformation are given as well.

For all the images the position and the representation of SIFT features have been calculated. Starting from them and considering a variable spatial bandwidths the *bi-SIFT* features have been generated. From each keypoint, in the *bi-SIFT* set, the corresponding point in the transformed image is evaluated. The point from the first image can be accurately mapped in the second image re-applying transformation with given parameters. If the matched point falls is in the proximity of the mapped point the matched is validated, while it is not validated in the other case.

Conclusions

We presented a novel local feature, based on SIFT point descriptors, that composing keypoints detected inside a region create a global, more abstract, representation. The feature shows promising results in image matching and retrieval of images according their content. Result of have been shown for retrieval of images in a set 30 images and for matching among images processed with affine transformations. Future works include the test of the proposed feature in larger dataset and the connection of local information represented by proposed feature with image annotation keywords.