bi-SIFT: Towards a semantically relevant local descriptor

Ignazio Infantino, Filippo Vella
Consiglio Nazionale delle Ricerche
ICAR. Viale delle Scienze ed. 11
Palermo, Italy
infantino@pa.icar.cnr.it
vella@pa.icar.cnr.it

Giovanni Spoto, Salvatore Gaglio
Università degli Studi di Palermo
DINFO. Viale delle Scienze ed. 9
Palermo, Italy
spoto@studiospoto.it
gaglio@unipa.it

ABSTRACT: *Local features are widely adopted to describe visual information in tasks for image registration and matching. Nowadays the most used and studied local feature is SIFT (Scale Invariant Feature Transform)[1] since it assures a powerful local description and the invariance when little changes in the viewpoint occur. We propose a feature that is based on SIFT features and tends to capture larger image areas in images and can be used for semantic based task. These features are called bi-SIFT for their resemblance with textual bigrams. We tested the capability of the proposed representation with Corel data-set and publicly available image dataset. In particular we calculated the most representatives features through a clusterization process and used these value according to the visual terms paradigm. Experiments on the representation of sets of images with the proposed representation are shown. Results appear to be encouraging.*

Keywords*:* Image Retrieval, Image analysis, Image representation, SIFT

Received: 2 November 2009, Revised 29 December 2009, Accepted 2 Janurary 2010

1. Introduction

A growing trend in image classification and object recognition has focused attention in finding specific and particular objects instead to address the recognition of classes of objects. For this reason much attention has been raised by the possibility to describe objects and portion of an image with features able to capture the appearance of particular point of interest [2] [3] [4] [5]. Among them SIFT (Scale Invariant Feature Transform) features [1] are widely adopted to represent visual information for tasks involving object identification, image registration [6][7] and image matching. These features have the properties to be invariant to variation of scale, image rotation and affine variation of viewpoint.

Being local descriptors these features take into account the values of the pixels in a region around the key point but lack to capture a global information bound to the higher level semantic of a visual scene.

On the other side, a descriptor aimed at capturing semantic content - related to objects and "large scale" pattern - should represent characteristics that are invariant for a great gamma of transformation and should capture values that are less affected by different views and acquisition parameters. SIFT are a good starting point since they are invariant to rotation and scale variation and these properties tend to preserve the objects characteristics in different scenes.

Changes of luminance are compensated by the SIFT representation as gradient histogram that is invariant to variation in luminance, changes of scale are compensated by the point selection, and changes of rotation are compensated by the orientation normalization. Unfortunately local features, as SIFT, are bound to local regions and do not cover relevant portion of images. We propose a technique to compose SIFT features to create more abstract features related to wider areas in images. A more global information is captured from values related to local features.

The idea is similar to what is done in linguistic computation when couples of single words are composed in bigrams, creating a new symbolic representation that is able to cover a larger language semantics than a single word [8] [9].

In literature other works employ local descriptors to describe image content and have been proposed by Csurka [10], Sivic and Zisserman [11] and Hare[12].

Csurka et al. [10] evaluate all the SIFT features for the data-set, they build a cluster distribution with k-means and represent the new images counting how many features fall in a chosen clusters. These new features are used as representation to classify test objects. Authors show promising results for the classification of 7 objects.

Similarly Sivic and Zisserman[11] use SIFT descriptor to form a set of values clustered and used as words to apply text retrieval techniques to matching objects in keyframe of video sequences. To validate the matching among SIFT features a property of spatial consistency is checked. In particular, a search area formed considering the 15 nearest spatial neighbors is defined. A match between two regions centered on a key point is validated if a key point, among the nearest 15 of the starting point, matches with a key point among the 15 nearest point around the target one. Matches that fall outside this frame are not considered as matching score.

Hare et al. [12] applied a cross language latent semantic indexing technique from computational linguistic when the visual portion of information is described by SIFT features.

Ke and Sukthankar[13] 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.

As further reference in literature, attempts to describe objects with affine invariant descriptors can be found in [14][15]. Lazebnik et al. [14] propose descriptors that show invariance to affine transformation identifying matches among couple of images depicting the same subject (authors concentrated their test on a set of six hundreds butterflies), the initialized descriptors are matched against a larger validation set. Affine regions are spotted with a Laplacian blob detector based on Lindberg descriptors[4].

The found regions are represented with multiple types of image information as spin images[16] and RIFT descriptors[17].

Brown et al. [15] propose a family of features which use groups of interest points to form geometrically invariant descriptors. Interest point in the scenes are located at the extrema of the Laplacian of the image in scale-space and are described with a family of 2D transformation invariant features. In particular groups of interest point which have in the nearest neighbors among 2 and 4. The 2xn parameters transformation to a canonical frame are computed. The descriptor is formed resampling the region local to the interest points in the canonical frame. Hough Transform is used to find a cluster of features in 2D transformation space. RANSAC is adopted to improve the 2D transformation estimate.

We considered that SIFT is a transformation very useful to capture local information generating hundreds or thousands of key points in a generic image and it is particularly suitable for tasks as image registration or image matching. In this work, we propose a novel representation achieved composing, in a suitable way, SIFT descriptors to create a reduced number of points in an image and that, at the same time, conveys a more abstract representation. The new descriptors cover a larger region instead of single point and that turn out to be semantically more relevant than the original patch of keypoint covered by SIFT descriptors.

The representation generates features that tend to be semantically nearer to scene objects and image tags. The new feature maintains the property of keypoints in robustness against variations in illumination and changes of scale and, for the particular way SIFT features are composed and still holds the invariance against rotations.

We show in the experiments that this composite feature, called bi-SIFT, generated from SIFTs allows to have a reliable representation improving solutions based on SIFT. The paper is organized as follows: Section II describes the identification of keypoints and creation of SIFT features, Section III shows how features are computed and how images are represented with the proposed feature. In Section IV the results of experiment setup are shown. Finally in the section V conclusions are drawn.

2. SIFT Features

SIFT features have been proposed by Lowe [1] for detecting points that are invariant against changes in illumination, image scaling, image rotation. These features follow the research of local features such as corner detectors by Harris[2] and Lucas [3], scale space theory by Lindberg [4] and edge density by Shi[5] . In particular, Lowe created local features aiming at identifying features less affected by geometric (as rotation, translation, scale change) and intensity variation (linear variation).

To automatically detect points that are more robust against these deformations, SIFTs are extracted creating a pyramid of Gaussian image transformation at different scales, finding in the pyramids the peaks that are independent from variation in scale and normalizing features according image dimension and rotation. In particular, the Gaussian Pyramid is calculated for the sample image and from each layer of pyramid the difference is calculated to obtain the Difference of Gaussian (DOG) pyramid. To get the local extrema in the images, the neighbor points in the same scale and in different scales are considered and points are retained when they are the greatest values among the neighbor point in the same image and in all the other scales. Similarly are retained points that are the smallest. This step is called keypoint localization.

For each candidate point the location and orientation are evaluated. Considering that points with low contrast are difficult to detect and points along edges are very unstable when noise is present, these points are discarded. Once the point is selected, an orientation histogram formed with the gradient orientations of sample points within a region around the keypoint is formed. The orientation histogram has 36 bins covering the 360 degree range of orientations. Each sample added to the histogram is weighted by its gradient magnitude and by a Gaussian-weighted circular window with a σ that is 1.5 times that of the scale of the keypoint. The modes of the histograms are considered as the dominant orientation for the given keypoints.

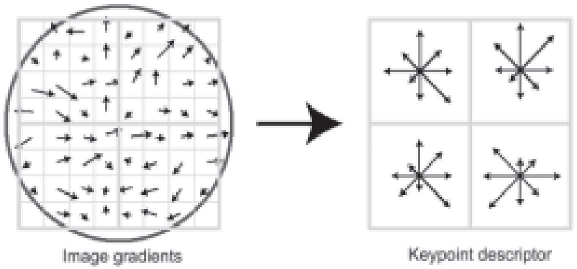


Figure 1. SIFT feature descriptor

All the orientation within the 80% of the maximum values are retained while the other values are discarded.

The detected keypoints are represented forming a descriptor with an orientation histogram on a region formed by 4x4 neighbor pixels. Since the histograms are populated considering orientation referred to the largest gradient magnitude, the feature is invariant against rotations. Each histogram contains 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 a further degree of robustness against illumination changes. A schematic representation of SIFT feature is shown in figure 1.4

3. Proposed Feature and Image Representation

SIFTs 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 a suitable 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 (bi-SIFT) is driven by an empirically fixed parameter. This parameter is called spatial bandwidth b as it is coherent with a spatial clustering in Mean Shift theory[18].

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 a point is falling in the region, meaning that the selected part of image captures a region with a reduced textured information, 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 of all 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 but could be extended).

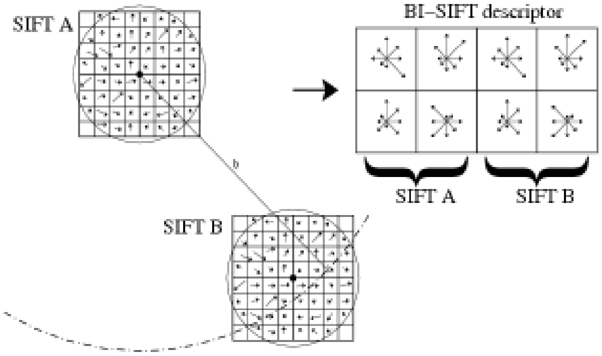


Figure 2. bi-SIFT composition

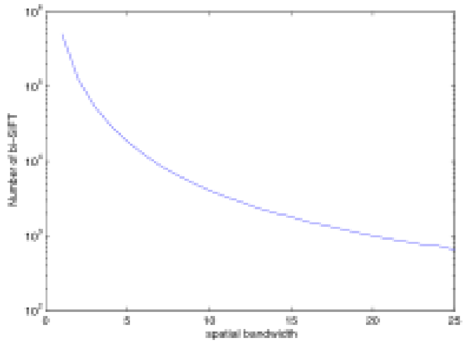


Figure 3. Number of bi-SIFT vs spatial bandwidth

The selected points are the most stable against variation in capture condition and are able to characterize scene invariant values. The new feature is formed by the juxtaposition of the SIFT representation of the selected points. A schematic representation of the new feature is shown in figure 2. This feature represents a wider area than SIFT descriptors, maintaining invariance against change of viewpoint.

The value of the spatial bandwidth affects the stability and the representation capability of the bi-SIFT. If the value of spatial bandwidth is set to zero on the couples of SIFT having the same keypoint would be chosen, if the value of the spatial bandwidth is set to a higher value the covered area is larger but the feature is less stable. In figure 3 it is plotted the number of bi-SIFT for a generic image versus the value of the spatial bandwidth b between 1 and 25. The value chosen for the parameter b is 6 because it is a good trade-off between the number of features and the stability of the descriptors and if not otherwise indicated it is the value that is applied for the experimental setup.

The variation of luminance does not affect SIFT for the normalization adopted and therefore will not affect the bi-SIFT.

Single SIFT features are also invariant against variation in scale. If a change of scale occurs the same region will be described by similar SIFT descriptors in the scaled images. If two keypoints are selected to form a bi-SIFT feature in the original image, given that change of scale maintains the two points inside the spatial bandwidth, the same two keypoints will be selected to form the bi-SIFT feature for the same region in the scaled image, providing invariance of bi-SIFT against change of scale.

If the image is rotated or the viewpoint is changed, a given region will produce an approximation of the original bi-SIFT descriptor. Single points described with SIFT features are invariant, for SIFT properties, to rotation. A couple of keypoints forming a bi-SIFT feature is mapped by rotation in two different position but since the gradient magnitude is not affected by rotation, the sort of feature will produce the same keypoints order inside the given region and the same bi-SIFT will be formed thus assuring invariance against rotations.

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 a couple of points that are accidentally near in an images (e.g. a point from object and a point from background) will be greatly affected by change of viewpoint and they will be retrieved in images depicting the same object in a different scene with low probability. So the recurrence of bi-SIFT feature asserts the presence of a given object or a characteristic pattern for a given scene. Some examples of bi-SIFT are shown in 4 for two images of theof the indoor class of the Corel data-set. The distribution of bi-SIFT in the images shows how these features are typically placed along object edges and corners.

The proposed features, for their properties, can be also profitably used to find reliable points for matching between images. Some experiments about the matching properties of bi-SIFT are shown in figure 5d and figure 5e. The images[1] in figure 5 have been scaled with a factor from 0.75 to 1.25 and rotated. For each row in figure 5d the performance with image 5a, 5b and 5c is shown. The x axis depicts the angle of rotation in degrees and y axis the percentage of matching features.

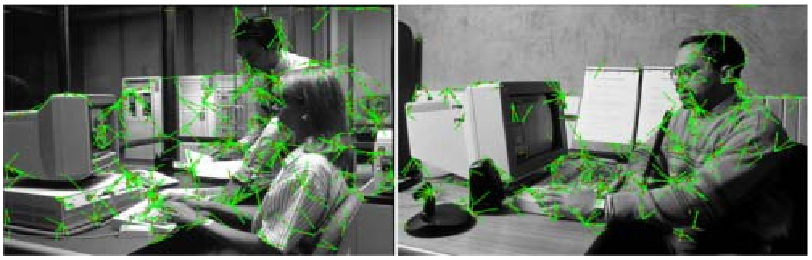


Figure 4. Example of bi-SIFT in indoor scenes

The triangles are referred to bi-SIFT while the circle are referred to SIFT. The number of SIFT is much higher than the number of bi-SIFT and it happens in some case that no match among bi-SIFT is found. In this case the marker is missing. On the other side when a match is detected there is a very high probability that the match is correct.

The image 5e shows the experiment of the same image set when data are aggregated according to the rotation angle. The x axis shows the change of scale between 0.75 and 1.25 and the y axis shows the percentage of correct matches.

This tests have been lead using the SIFT implementation by Vedaldi and Fulkerson[19].

A further example of matching between images of the Graffiti data set is shown in figure 6. Images have been acquired with different points and the quality of matched points shows the good performance of the bi-SIFT also in image registration tasks.[2]

Given two images of the Graffiti data-set the matching among the two set of bi-SIFT are shown in figure. The plots show the values of Recall versus 1-Precision in a similar way to what is done in [13].

The two values are defined accordingly as:

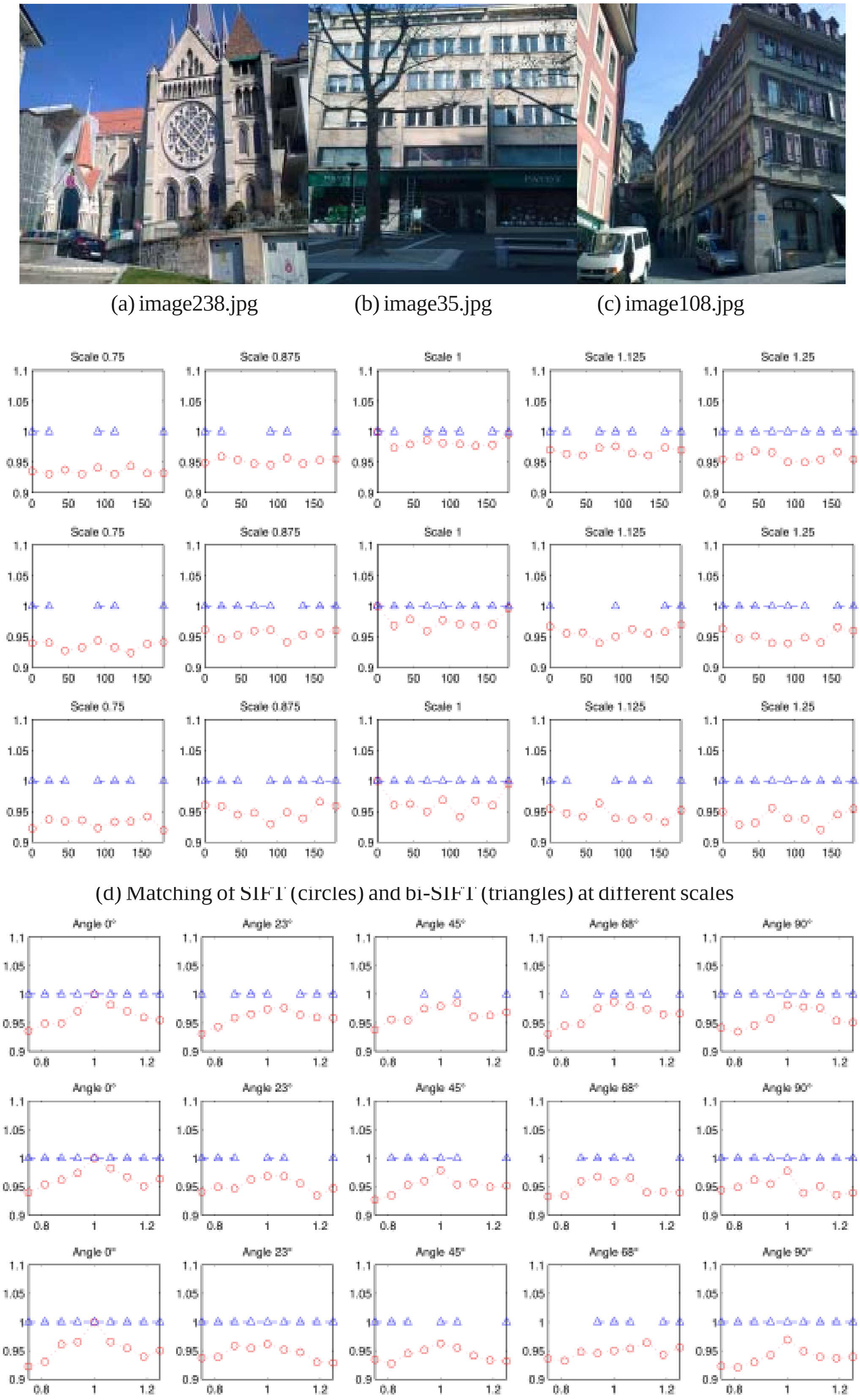
$$recall = \frac{\text{number of correct positive}}{\text{number of total positive}} \tag{1}$$

and

$$1 - precision = \frac{\text{number of false positive}}{\text{total number of matches}} \tag{2}$$

[1] *Images are available at http://http://comminfo.rutgers.edu/conferences/mmchallenge/2010/02/10/nokia-challenge*

[2] *Images are available at http://www.robots.ox.ac.uk/67vgg/data/dataaff.html*



(a) image238.jpg (b) image35.jpg (c) image108.jpg

(d) Matching of SIFT (circles) and bi-SIFT (triangles) at different scales

(e) Matching of SIFT (circles) and bi-SIFT (triangles) with different angles

Figure 5. Comparison of SIFT and bi-SIFT matching performance with variation in scale and rotation

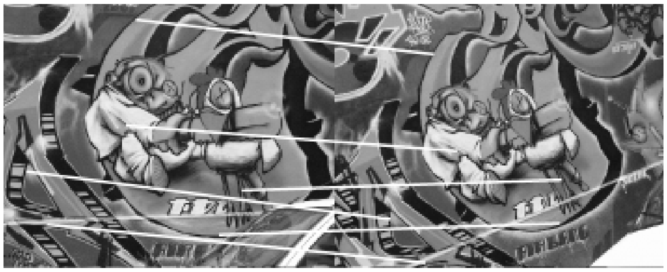


Figure 6. Matching example with bi-SIFT

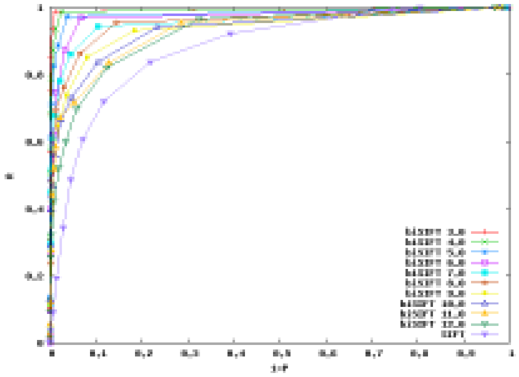


Figure 7. Recall vs 1-Precision in matching task with Graffiti data-set

The plot shows as the bi-SIFT features allow a higher precision with an equal recall compared to SIFT and how the performance is better when the spatial bandwidth, used to build the composition of SIFT decreases. The tested values of b are among 3 and 12.

A. Bag of bi-SIFT

Due to the invariance properties and description capability the proposed feature can be successfully used to represent images in large data set. A typical problem in the description of large set of images can raise when the number of features extracted from each image can make the problem cumbersome. SIFT features are affected by this problem for the high number of features extracted from each image.

Although the number of bi-SIFT is a reduced number if compared with the number of SIFT, we choose to adopt and test the representation based on visual terms [20]. The main advantage of this technique is that to consider the most relevant features and to represent values according to the more frequent and intrinsically more significant points in the representation space.

Here we propose the application of a technique based on visual terms where the feature to cluster is given by the above described bi-SIFT feature. The underlying hypothesis is that couples, or bigrams, of SIFT are suitable to describe areas larger than the single local features expressed by SIFT and these features, collecting information at a level between objects and pixels, are good candidates in the attempt for the reduction of the semantic gap.

The conceived features are used to find reliable points in images for matching of local part of images and furthermore to represent the visual content of images. The set of local features is used to create a dictionary of features for the description of a generic visual content. The set of values used as symbolic descriptor is evaluated clustering the bi-SIFT, represented as vectors of 256 real numbers, and extracting the centroids as fundamental values. For each category the corresponding set of bi-SIFT is considered and the visual terms for each category is added to the global set.

A generic image is therefore represented as a bag of visual terms and different images will be represented by vectors considering the distribution of these values in the image. The process that extracts visual terms as centroids of feature clusters allows to reduce the presence of noise in the features reducing the irrelevant information in representation. Extracted visual terms are collected forming a dictionary used to represent any visual content. Some examples of visual term obtained from the set of the bi-SIFT are shown in figure 8. For each image in the data-set the SIFT features are extracted, the points are coupled to form bigram of SIFT as described above. For each bi-SIFT the nearest feature in the dictionary is found. For each image a vector with a cardinality equal to the size of visual dictionary, is filled with the visual terms found in it.

The visual dictionary is evaluated considering all the bi-SIFT inside a given set of images and clustering values for the bi-SIFT elaborated with the above described approach. In particular, clustering of features is accomplished through a technique based on the feature density estimation. This technique is called Mean Shift since the algorithm iteratively follows the mean vector

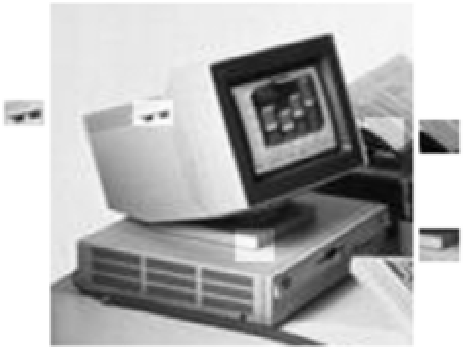


Figure 8: Example of bi-SIFT visual terms for an image depicting a computer

along the gradient of density[18]. The algorithm is based on bandwidth parameter that drives the clusterization process and creates a number of clusters according to the density estimation.

For higher data dimension an improved version of Mean Shift has been proposed by the same authors called Adaptive Mean Shift. According this second algorithm data are partitioned with Local Sensitive Hashing (LSH) to get a reduced dimension representation of data. On the hashed data the Mean Shift algorithm is applied [21]. The clusterization parameters are picked empirically.

The set of all the visual terms got from any category, through the clusterization process, is added to the global dictionary and will be used to describe an image in the data-set and a generic image too. The representation of an image is achieved with a set of values filled with tij values defined below:

$$tij = \frac{nij}{ni} log \frac{N}{Nj} \qquad (3)$$

where n_{ij} is the number of occurrence of *j -th* visual term in the *i - th* image, n_i is the number of terms in the *i-th* image, *N* is the number of images in the data set and N_j is the number of occurrence of the *j -th* visual term. This processing multiplies the term frequency n_{ij}=n_i for the *inverse document frequency log(N=*N_j*)* giving an higher value to terms that are present in shorter visual document and to less frequent visual terms.

An advantage using *bi-SIFT* is the possibility to store stable composition of SIFT keypoints. For example if two SIFT keypoints are near and related to the same object or the same pattern in the scene, when the view changes or the object is slightly moved in the scene, the *bi-SIFT* will not change considerably. On the other side, if the SIFT keypoints forming the bi-SIFT come from different objects or came from an object and background, there is a low probability that they will be near in an other scene. In particular, during the clustering process, if no similar feature is found, feature will not contribute in forming clusters, since they are rare in images representation and definitely will not affect global representation.

4. Experimental Results

The representation with bag of *bi-SIFT* has been tested with images from the Corel data set. It has been widely used as benchmark in image classification and annotation tasks (e.g.[20] [22]). It consists of 5000 images divided in a set of CD containing images with homogenous categories. Some images from category ”computer” and ”tiger” are shown in the figure 9.

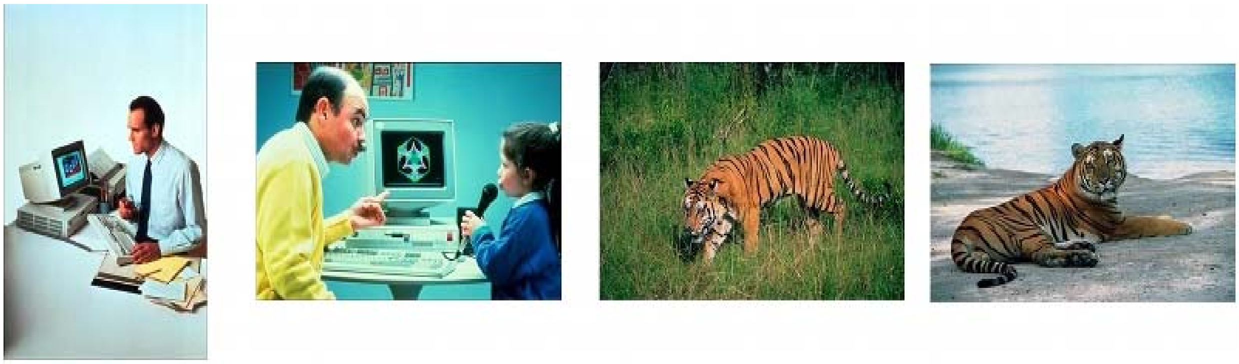


Figure 9. Example of Corel Images from category Computer and category Tiger

To reduce the experiments time and produce a subsampled test, a subset of 765 images has been extracted from the Corel data-set. SIFT features have been calculated and have been coupled to form the corresponding *bi-SIFT* features.

Each image in the data-set has been processed to discard the color information and to extract the SIFT features. Although color channel conveys a relevant portion of information in this case we consider features that involve just luminance information with the possibility to add chroma information as an additional information channel in a second moment. To evaluate the value of spatial bandwidth (see section III), related to the image area covered by a *bi-SIFT* feature, a clusterization process has been tested with different values of spatial bandwidth. Results of experiments are shown in figure 10. The number of centroids versus the value of *kernel bandwidth* adopted with Adaptive Mean Shift is shown. In particular the values are plotted when the *bi-SIFT* spatial bandwidth is among 2 and 14. The graph shows that the larger number of different centroids is produced when the value for the Adaptive Mean Shift *kernel bandwidth* is set to 0.065.

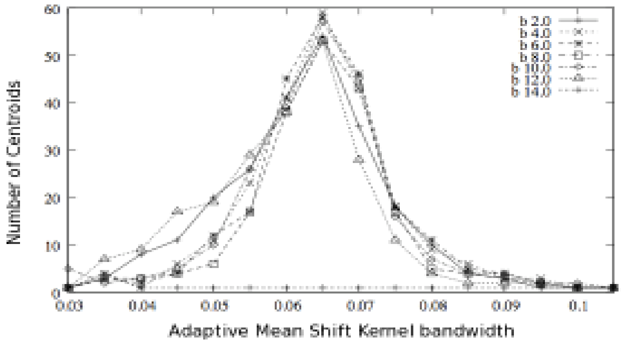


Figure 10. Number of Clusters versus the spatial bandwidth with Adaptive Mean Shift Clustering

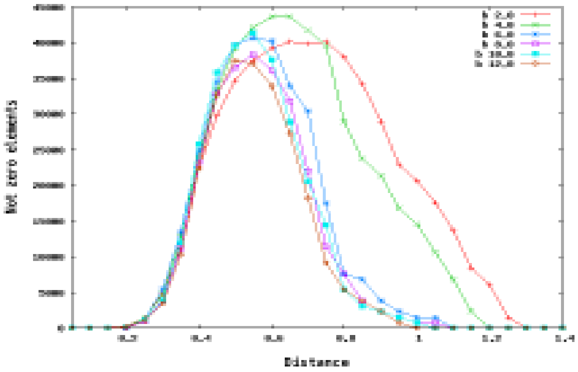


Figure 11. Number of not zero values versus the distance

Features extracted are clustered with Adaptive Mean Shift algorithm using a *kernel bandwidth* of 0.065 and varying the spatial bandwidth in the feature creation. The centroids generated by the clustering process are used as *visual terms* allowing to describe images with a set of symbolic features. The process, as described in section III-A, creates a representation with

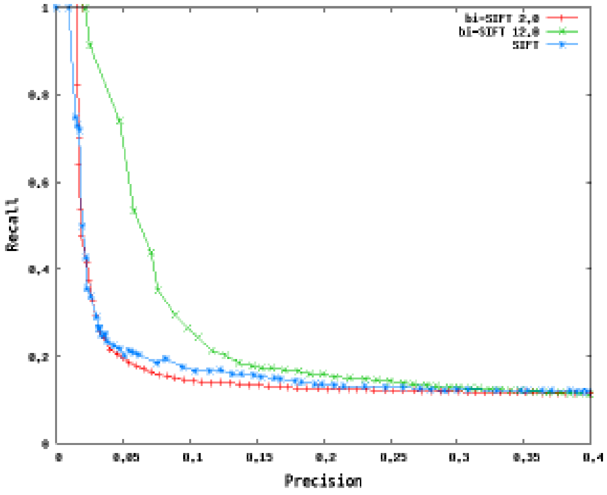


Figure 12. Recall versus Precision

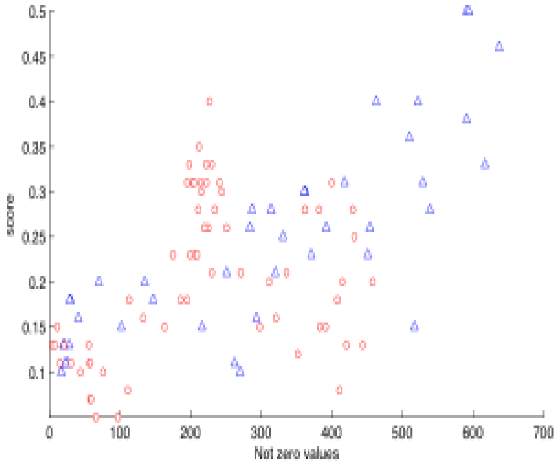


Figure 13. Score matching comparison between SIFT (circles) and bi-SIFT (triangles)

each row corresponding to an image represented according to the extracted set of visual terms. The representation of images with the *tf-idf* is function of the number of visual terms, of the chosen features and of the distance used to match feature with visual terms.

Figure 11 shows how the population of the matrix, having with rows referred to images and columns referred to visual terms, depends on the value of distance threshold used to match a given feature with a visual term. If distance threshold is too low, there are few matches among image *bi-SIFT* and the visual terms; on the other side if threshold is too high many matches are found and all these values are cut by the entropic filtering (see eq. 3). The plot is a gaussian curve as shown in figure 11. For different value of *bi-SIFT* spatial bandwidth the gaussian curve is slightly shifted. When the spatial bandwidth is set to 7, the distance threshold that allows the least number of not zero value is 0.6.

In figure 12 the Recall versus Precision is shown for two values of spatial bandwidth compared with SIFT curve. The *bi-SIFT* curve corresponds to values of spatial bandwidth equal to 2 and 12. The curves for the other value of bandwidth are among these two ones and the SIFT curve is very near to the curve with value equal to 2.

A further experiment, employing visual terms, has been considered using the data set *lausanne*. Selecting images of the set have been created 20 classes containing visually similar images. Each class contains 3 photos that are shot in the same place and with an intersection among the three figures. Images have been described with SIFT and *bi-SIFT* visual terms and a retrieval experiment has been run.

One image has been used as test image and a match with all the other images in the data-set has been calculated. Images matched with the test image have been ordered according the number of found visual terms. A score has been assigned to the test if the images in the same class have been retrieved among the first three images or not. If a single image of the class has been found in the first three positions a score of 1 has been given, if two images have been found in the first three positions, a score of 2 has been given. This experiment is analogous to the experiment of Ke and Sukthankar in [13].

The best average score when images are described with SIFT visual terms is 0.40 while the best average score with *bi-SIFT* is 0.5. For these experiments the visual terms have been generated employing Adaptive Mean Shift clustering algorithm with a *kernel bandwidth* of 0.065. According to the kernel bandwidth and the clustering parameters the population of the matrix of
images vs visual terms can be considerably changed. A plot of how these parameters affect retrieval and matching performance is shown in figure 13.

The best value when images are described with SIFT visual terms (circles in the figure) is 0.40 when the matrix *tf-idf* has 227 not zero elements, while the best score with *bi-SIFT* (triangles in the figure) is 0.5 when the number of not zero elements is 591.

Conclusions

An approach aiming at creating a semantically relevant feature has been presented. The feature called *bi-SIFT* is based on the well-know SIFT feature. To create a more abstract representation features SIFT keypoint descriptors are composed to form a feature that resembles the text bigrams and that tends to capture a larger part of image and can be used to semantic oriented tasks. The feature has the invariance properties of SIFT and Experiments show promising results and future works will include the application in task of scene understanding and automatic image annotation.

References

[1] Lowe, D.G. (2004). Distinctive image features from scaleinvariant keypoints, *International Journal of Computer Vision*, 60, 91–110.

[2] Harris, C., Stephens, M.(1988). A combined corner and edge detection, *In*: Proceedings of The Fourth Alvey Vision Conference. p. 147–151.

[3] Lucas, B. D., Kanade, T. (1981). An iterative image registration technique with an application to stereo vision.

[4] Lindberg, T. (1993). Effective scale: A natural unit for measuring scale-space lifetime, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15 (10) 1068–1074.

[5] Shi, J., Tomasi, C. (1994). Good features to track, *In*: Proceedings of the Conference on Computer Vision and Pattern Recognition, June. p. 593–600.

[6] Cheng, S., Stankovic, V., Stankovic, L (2009). Improved sift-based image registration using belief propagation," Acoustics, Speech, and Signal Processing, IEEE International Conference on, vol. 0, p. 2909–2912.

[7] Fan, Y., Ding, M., Liu, Z., Wang, D.(2007). Novel remote sensing image registration method based on an improved SIFT descriptor, *In*: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Nov. 2007, vol. 6790 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series.

[8] Infantino, I., Spoto, G., Vella, F., Gaglio, S. (2010). Composition of SIFT features for robusts image representation, *In*: Multimedia Content Access: Algorithms and Systems IV, *In*: Proceedings of Society of Photographic Instrumentation Engineers (SPIE) V. 6540B.

[9] Infantino, I., Spoto, G., Vella, F., Gaglio, S. (2009). Image representation with bag-of-biSIFT, *In*: Proceedings of the 5th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS'09),

[10] Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C. (2004). Visual categorization with bags of keypoints, *In*: In ECCV International Workshop on Statistical Learning in Computer Vision.

[11] Sivic, J., Zisserman, A (2003). Video google: A text retrieval approach to object matching in videos, *In*: Proceedings of the International Conference on Computer Vision, V. 2, p. 1470–1477.

[12] Hare, J., Lewis, P., Enser, P., Sandom, C. (2006). A linearalgebraic technique with an application in semantic image retrieval, *In*: Proceedings of the International Conference on Image and Video Retrieval, p. 31–40.

[13] Ke, Y., Sukthankar, R (2004). Pca-sift: A more distinctive representation for local image descriptors, *In*: Proc. of Computer Vision and Pattern Recognition (CVPR) 04, p. 506–513.

[14] Lazebnik, S., Schmid, C., Ponce, J (2004). Semi-local affine parts for object recognition, *In*: Proceedings of BMVC, 2004.

[15] Brown, M., Lowe, D. (2002). Invariant features from interest point groups, *In*: Proceedings of BMVC 2002, 2002.

[16] Johnson, A. (1999). Using spin images for efficient object recognition in cluttered 3d scenes, *IEEE Transaction on Pattern Analysis and Machine Intelligence,* 21 (5) 433–449, .

[17] Lazebnik, S., Schmid, C., Ponce, J. (2004). A sparse texture representation using local affine regions, *In*: Technical Report, CVR-TR-2004-01, Beckam Institute, University of Illinois.

[18] Comaniciu, D., Meer, P (2002). Mean shift: A robust approach toward feature space analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 24. 603–619.

[19] Vedaldi, A., Fulkerson, B. (2008). VLFeat: An open and portable library of computer vision algorithms, http://www.vlfeat.org/

[20] Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., Jordan, M.I. (2003). Matching words and pictures, *Journal of Machine Learning Research,* 3. 1107–1135.

[21] Georgescu, B., Shimshoni, I., Meer, P (2003). Mean shift based clustering in high dimensions: A texture classification example, *In*: Proceedings of IEEE International Conference on Computer Vision.

[22] Vella, F. , Lee, C.-H., Gaglio, S. (2007). Boosting of maximal figure of merit classifiers for automatic image annotation, *In*: Proc. of Internation Conference on Image Processing.