Spatially-consistent partial matching for intra- and inter-image prototype selection

S. Kosinov, E. Bruno and S. Marchand-Maillet *

Viper *Group - University of Geneva - Switzerland*

Abstract

This paper describes a method of introducing spatial consistency constraints in the process of matching set-based descriptors extracted from digital images. The proposed matching technique is guided by a rule that can be summarized as follows: a descriptor is important for the match if it is similar to some descriptor from the other image and its spatial neighbors are important. The resulting match is partial in the sense that it deliberately avoids the complexity of searching for one to one correspondences among particular descriptors, but established affinity among groups of descriptors instead.

Formally, the proposed method is expressed as an eigenvalue problem, where the principal eigenvector's components render the importance values of individual descriptors, while the corresponding eigenvalue represents an estimate of the overall strength of affinity between images being matched. These measures of descriptor importance and image affinity are shown to provide a natural basis for intra- and inter-image prototype selection. Several variations of the proposed technique are empirically evaluated on the task of content-based image retrieval, demonstrating encouraging results.

Preprint submitted to Elsevier

1 Introduction

This work is situated in the domain of content-based image retrieval where digital images are represented by a variable number of feature vectors commonly referred to as set-based descriptors. The main contribution of this work is a spatially-consistent partial matching method designed to establish correspondence between groups of descriptors rather than between individual descriptors themselves. The described method is empirically evaluated as an intra- and inter-image prototype selection approach. In the former case, the method allows to find descriptors that contribute best to the match, whereas in the latter case, the technique allows to select descriptors that are representative of a group of images belonging to a certain class.

Our choice of the set-based descriptors is motivated by their superior performance [1] and applicability in a wide range of content-based image retrieval applications. In general, most of these descriptors characterize local image features extracted at certain interest points within an image. Examples of these descriptors include maximally stable extremal regions (MSER) [2], scale-invariant feature transform (SIFT) [3], speeded up robust features (SURF) [4], as well as their extensions such as PCA-SIFT [5], SIFT with global context [6], gradient location-orientation histogram (GLOH) [1].

Further, our additional motivation in this work is to improve the two following

* Corresponding author.
Email address: marchand@cui.unige.ch (S. Marchand-Maillet).

[1] This work is supported by the Advanced Media Management group at Intel Corp. and the Swiss NCCR Interactive Multi-modal Information Management (IM2).

[2] First author S. Kosinov is now with Google Research Labs, Zurich, CH

characteristics of the popular matching and retrieval techniques for set-based descriptors: spatial consistency and many to many correspondence.

Spatial consistency. The former aspect relates to the fact that spatial configuration of descriptors and their positions relative to each other are seldom considered when a match is calculated. We believe that this kind of approach may be detrimental to the overall performance, since oftentimes images contain a large number of high similarity descriptors that are not localized on the visual objects of interest and thus lead to an erroneous match (see section 3.1 for such an example).

Certainly, there exists earlier work that attempts to add some context to the information content of the local feature descriptors. But these contributions together with their advantages have a number of drawbacks as well. For instance, in [6] the authors add only the shape context information to that of the SIFT descriptor, while the semi-local constraints proposed in [7] require threshold tuning for neighbor match percentage and impose complicated restrictions on admissible angles between neighboring descriptors being matched. The proposed approach described below strives to take into account both spatial proximity and feature-based similarity information during the matching process, while at the same time trying to avoid the above mentioned pitfalls.

Many to many correspondence. The latter aspect concerns the way the correspondence is established between matching descriptors in two images being compared. In the proposed method, we focus on deriving a common quality estimate (importance) for each descriptor in an image, and use this value to decide whether a given descriptor belongs to a group that constitutes a match. Thus, a many to many correspondence is found between groups of image descriptors, without the need to resort to model- (e.g., RANSAC [8]) and graph-based (e.g., bipartite graph matching, as-

signment problem [9]) techniques that can be quite costly from the computational complexity point of view.

In the section that follows we will introduce the method of spatially-consistent descriptor matching. In so doing, we will provide a detailed description of how it calculates descriptor importance and image affinity, and show how these two measures may serve as a basis for various intra- and inter-image prototype selection techniques. Then, in section 3, we will present the experimental results which include exploratory data analysis and an empirical evaluation of the proposed method on the task of content-based image retrieval. The article will conclude with a summary of the developed techniques and their important properties.

2 Spatially-consistent descriptor matching

This section presents the method of spatially-consistent partial matching of local image descriptors. Here, we detail the problem formulation of the proposed method and provide an illustrative example of its usage. We also show how descriptor importance and image affinity measures computed by the proposed approach are applied to the problem of intra- and inter-image prototype selection problem.

2.1 Descriptor importance

One way of introducing the spatial consistency constraints in a given matching procedure is to make sure that the quality of a match of a given descriptor depends on both descriptor itself and its neighbors. In order to substantiate this idea, we introduce the concept of *descriptor importance* and the following rule to define it:

A descriptor is important if it matches well some descriptor from the other image and its spatial neighbors are important.

The above rule is fairly straightforward to cast as a mathematical expression:

$$\alpha_i^{(A)} = \sum_{t=1}^{n^{(A)}} \alpha_t^{(A)} \cdot p\left(D_t^{(A)}, D_i^{(A)}\right) \cdot \mathrm{sim}\left(D_i^{(A)}, I^{(B)}\right), \tag{1}$$

where bracketed superscripts refer to the image index that the corresponding parameter pertains to. In this notation, $n^{(A)}$ is the number of descriptors in image A. Similarly, $\alpha_i^{(A)}$ (for $i \in 1 \dots n^{(A)}$) is the importance of i-th descriptor in image A. The value $p\left(D_t^{(A)}, D_i^{(A)}\right) \in (0, 1]$ is the normalized proximity between descriptors t and i within image A, and $\mathrm{sim}\left(D_i^{(A)}, I^{(B)}\right) \in (0, 1]$ is the feature-based similarity between the i-th descriptor from image A and some descriptor from image B. In order to simplify Equation (1), we may omit the superscripts and indices to define importance vector:

$$\boldsymbol{\alpha} = \left[\alpha_1^{(A)}, \alpha_2^{(A)}, \ \dots \ , \alpha_{n^{(A)}}^{(A)}\right]^T, \tag{2}$$

proximity matrix P:

$$P = \begin{bmatrix} & & \\ 1 & p_{ij} & \cdots \cdots \\ & & \\ \cdots & 1 & \cdots \cdots \\ & & \\ \cdots \cdots \cdots \cdots & & \\ & & \\ \cdots \cdots \cdots & 1 & \end{bmatrix}, \tag{3}$$

where $p_{ij} = p\left(D_i^{(A)}, D_j^{(A)}\right)$ for $i, j \in 1 \dots n^{(A)}$, and the diagonal match quality matrix S:

$$S = \begin{bmatrix} s_{11} & 0 & \cdots & 0 \\ 0 & s_{22} & \cdots & 0 \\ \cdots & \cdots & s_{ii} & \cdots \\ 0 & 0 & \cdots & s_{nn} \end{bmatrix}, \tag{4}$$

where $s_{ii} = \text{sim}\left(D_i^{(A)}, I^{(B)}\right)$ for $i \in 1 \dots n^{(A)}$.

Taking advantage of the simplified notation of Equations (2)-(4), Equation (1) may finally be written down as:

$$\boldsymbol{\alpha} = SP\boldsymbol{\alpha}, \tag{5}$$

which is a standard eigenvalue problem. The importance vector $\boldsymbol{\alpha}$, solution of (1), has to be a non-negative eigenvector such as α fullfill the relation $\lambda\boldsymbol{\alpha} = \lambda SP\boldsymbol{\alpha}$ (where λ is an eigenvalue) and $\alpha_i \geq 0, \forall k$. As SP is a non-negative matrix, the Perron-Frobenius theorem [10] guarantees the non-negativity of the *principal* eigenvector (corresponding to the largest eigenvalue). We thus obtain $\boldsymbol{\alpha}$ as the SP's principal eigenvector whose components are importance estimates of the respective image descriptors. The larger the magnitude of the principal eigenvector's component, the more important the corresponding descriptor is for the match between two given images. A fast estimation of $\boldsymbol{\alpha}$ may be obtained with the power method [11], an iterative procedure used to approximate the principal eigenvector of a matrix

$$\boldsymbol{\alpha}^{k+1} = \frac{SP\boldsymbol{\alpha}^k}{||SP\boldsymbol{\alpha}^k||}. \tag{6}$$

Finally, by carrying out this procedure for both images and selecting their most important descriptors, we obtain a spatially-consistent partial matching, which consti-

tutes the essence of the proposed method.

Note that the derived matching is partial in the sense that it deliberately avoids the complexity of searching for one to one correspondences among particular descriptors, but established correspondence among groups of descriptors instead. In the section that follows, we consider an illustrative example of the method outlined above.

2.2 Illustrative example

Consider a simplified example where the two images to be matched are 8 by 8 grids with some colored tiles, as shown in Figure 1.

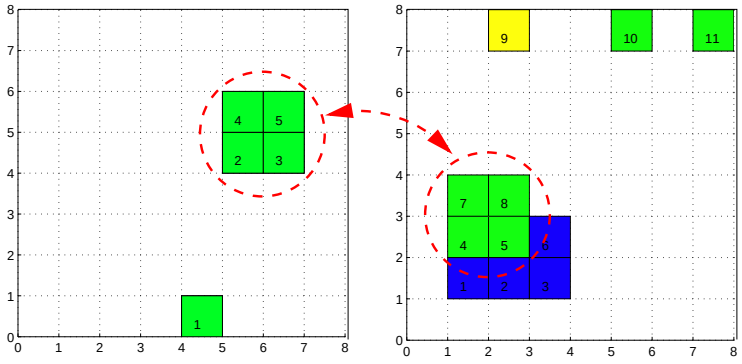


Fig. 1. Example of spatially consistent partial matching: The dashed line shows the matching groups of tiles between which the proposed method established correspondence, see text. (This is image is best viewed in color)

These tiles in general represent image regions, SIFT points or any other kind of set-based image descriptors. The first image, shown on the left of Figure 1, contains 5 tiles, while the other one has 11 tiles. Euclidean distances between tiles on the grid within an image, $d^{(A)}_{\text{grid}}(i,j)$ for $i,j \in 1 \ldots n^{(A)}$, are transformed into normalized

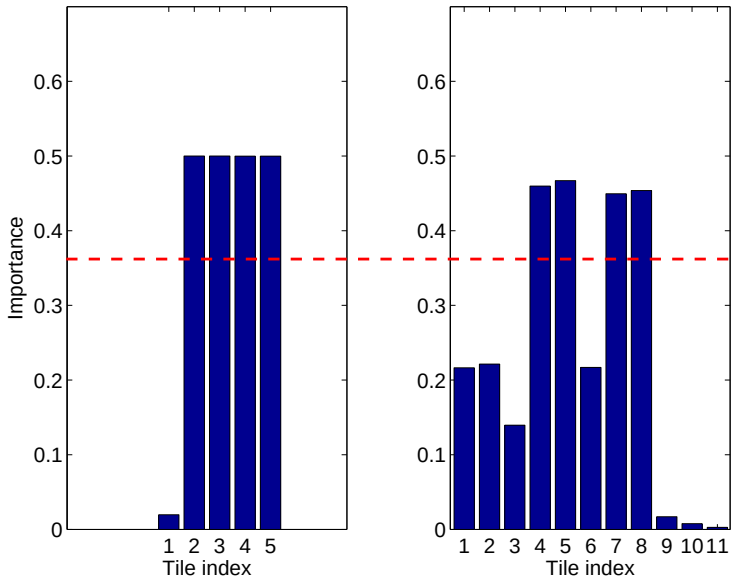


Fig. 2. Example of spatially consistent partial matching: The bar graph of descriptor importance values computed for each of the 5 tiles of the first image (left), and 11 tiles of the second image (right). The dashed line marks the threshold separating the tiles important enough to belong to a match from the rest.

proximities, as follows:

$$p\left(D_i^{(A)}, D_j^{(A)}\right) = \exp\left(-\frac{d_{\text{grid}}^{(A)}(i,j)}{\sigma_{\text{grid}}^2}\right), \tag{7}$$

where σ_{grid} is the Gaussian kernel width parameter that controls what range of distance values gets mapped into the $(0, 1]$ interval. Ideally, a judicious choice of d_{grid} and an optimal setting of σ_{grid} (using automatic scale selection [12] for instance) would ensure the method to be scale invariant. However this issue has not been formally addressed in this study.

Analogously, Euclidean distances between 3-dimensional RGB vectors representing the color content of each tile, $d_{\text{RGB}}^{(AB)}(i,j)$ for $i \in 1 \ldots n^{(A)}$ and $j \in 1 \ldots n^{(B)}$, are transformed into normalized similarities:

$$\text{sim}\left(D_i^{(A)}, D_j^{(B)}\right) = \exp\left(-\frac{d_{\text{RGB}}^{(AB)}(i,j)}{\sigma_{\text{RGB}}^2}\right), \tag{8}$$

with a similarly defined σ_{RGB} parameter. The match quality matrix elements are then derived as the best possible similarity between a given descriptor and any descriptor in the other image:

$$\mathrm{sim}\left(D_i^{(A)}, I^{(B)}\right) = \max_{j \in 1 \ldots n^{(B)}} \mathrm{sim}\left(D_i^{(A)}, D_j^{(B)}\right). \tag{9}$$

Finally, using the above proximities (Equation (7)) and feature-based similarities (Equation (9)), we solve the eigenvalue problem (Equation (5)) for each of the two images and thus obtain importance estimates for every tile in both images, as shown plotted in Figure 2. Then, we select the tiles belonging to a match by simple thresholding of the tile importance at around 0.4, as shown in Figure 2 with a dashed line. The matching tiles in the first image (left) are 2, 3, 4, 5, and 4, 5, 7, 8 in the second image (right). As can be easily confirmed from Figure 1, the matching tile indices correspond to the green cube object found in both images. The spatial consistency of the resulting match is also apparent from Figure 1. Indeed, the tiles constituting the match have similar feature-based content, i.e. green color, and are located close to each other within their respective images. In contrast, the other tiles that might have matched if judged solely by their either feature-based content (e.g., tile 1 in image 1, tiles 10, 11 in image 2) or close proximity (e.g. tiles 1, 2, 3, 6 in image 2), but not both, are assigned low importance and hence left out. In addition to that, this example demonstrates that the match found by the proposed technique is partial, since there is no extra effort spent on trying to establish one to one correspondences between individual matching tiles from the two images. Instead, only the groups of corresponding tiles are discovered.

As one would naturally expect, the solution of Equation (5) is typically found in terms of eigenvectors together with their respective eigenvalues. In order to provide a meaningful interpretation to the eigenvalue that corresponds to the principal eigenvector of SP, consider the following ideal case:

- all descriptors are close to each other, $p_{ij} = 1$,
- all descriptors match very well, $s_{ii} = 1$.

In this ideal case, every element of the matrix product SP is equal to one, the principal eigenvector's components are all equal to a constant indicating that all of the descriptors are important, while the corresponding eigenvalue is equal to the number of descriptors:

$$\boldsymbol{\lambda}^{(A)} = n^{(A)}, \tag{10}$$

where A is a given image. Thus, when divided by the number of descriptors, this eigenvalue gives a summary estimate of the overall image match quality normalized within the $(0, 1]$ interval. By extending this observation to the case when both image A is matched to image B, and image B is matched to image A by solving Equation (5), we define the *image affinity* value between images A and B, as follows:

$$\boldsymbol{a}^{(AB)} = \frac{\boldsymbol{\lambda}^{(A)}}{n^{(A)}} \cdot \frac{\boldsymbol{\lambda}^{(B)}}{n^{(B)}}. \tag{11}$$

While alternative formulations are certainly possible, Equation (11) appears to be a reasonable choice since $\boldsymbol{a}^{(AB)}$ has the advantageous properties of being normalized, symmetric and close to one only when both images match well.

In the above sections we have discussed two measures of descriptor importance and image affinity that may be considered in a more general context of the *prototype selection* problem. According to [13], prototype selection is the process of storing a well-chosen, proper subset of available training data instances. These instances thereby selected are referred to as *prototypes* and used with instance-based classifiers, such as nearest neighbor [14], that predict a class of an unseen data instance by comparing it to a set of prototypes.

In the case of content-based image retrieval where images are represented by set-based descriptors, the problem of prototype selection appears to be quite important because of the vast number of instances that a classifier must deal with in a realistic usage scenario. For instance, a moderate size group of one thousand images with one thousand descriptors per image creates a million descriptors for this group representing a certain class, of which there could be many.

In this context, the descriptor importance measure discussed in section 2.1 may be treated as an intra-image prototype selection technique, while image affinity covered in section 2.3 can be considered an inter-image prototype selection method. Indeed, the former quantity allows one to select the important descriptors that contribute to the match while disregarding those that do not, and the latter may help one find those images that match very well to others within their group and thus are worth being given priority when choosing what image to select descriptors from. Naturally, various other ways of prototype selection approaches are also possible based on combinations of descriptor importance and image affinity, such as those listed in the examples below.

- Affinity-weighted importance q_1 of descriptor $D_i^{(A)}$ from image A within a group of images G representing a certain semantic class:

$$q_1\left(D_i^{(A)}\right) = \sum_{B \in G, B \neq A} \boldsymbol{a}^{(AB)} \alpha_i^{(A)} \tag{12}$$

  combines the overall match quality expressed by image affinity together with the importance of a particular descriptor to derive its prototype quality score, q_1;

- Affinity-weighted importance rank q_2:

$$q_2\left(D_i^{(A)}\right) = \sum_{B \in G, B \neq A} \boldsymbol{a}^{(AB)} \mathrm{rank}\left(\alpha_i^{(A)}\right) \tag{13}$$

  computes prototype quality score q_2 by combining the overall match quality expressed by image affinity together with the rank of a particular descriptor as found in the list of descriptors sorted by their importance, emulating the voting mechanism whereby a descriptor receives one vote whenever it is more important than some other descriptor within an image;

- Importance rank q_3:

$$q_3\left(D_i^{(A)}\right) = \sum_{B \in G, B \neq A} \mathrm{rank}\left(\alpha_i^{(A)}\right) \tag{14}$$

  only emulates the voting mechanism (see above) disregarding the overall image match quality conveyed by the image affinity.

In the following section we are going to evaluate the above prototype selection techniques based on the described quality scores q_1, q_2 and q_3, and also compare their performance alongside a baseline method that makes no use of descriptor spatial information.

3 Experimental results

Here we present the details of the experimental results obtained while evaluating the proposed technique. Throughout all of the experiments, we chose to use SIFT [3] descriptors to provide set-based representation of image contents. Proximity and similarity functions are respectively the same to those defined in equations (7) and (9), but d_{grid} is the Euclidean distance measured in pixels normalized by the size of the image, and d_{RGB} becomes d_{SIFT}, the Euclidean distance between SIFT descriptors. The corresponding scale parameters σ_{grid} and σ_{SIFT} are empirically set to respectively 70 and 20 for all experiments.

Unless stated otherwise, any reference to a baseline method in the discussion that follows refers to a prototype selection approach based exclusively on the feature-based portion of SIFT descriptors' data with no regard to their spatial properties.

3.1 Exploratory analysis

While the toy example reviewed in section 2.2 is illustrative in demonstrating the essential properties of the proposed method, it is helpful to examine how the approach works with real-world examples before doing a full-scale evaluation. To this end, we perform a preliminary exploratory analysis of the results of computing descriptor importance and image affinity, as well as performing prototype selection, as described below.

First, we examine how the intra-image prototype selection results obtained by computing descriptor importance via spatially consistent matching are different from those of the baseline method, as depicted in Figure 3. As is apparent from the

figure, the 15 most important SIFT descriptors derived from spatially consistent-matching seem to localize around visual objects of interest found in both pictures. These results also indicates that in some cases background descriptors may also be selected as important while using the proposed method, which suggests that some voting procedure, such as one mentioned in section 2.4, over a large number of images may be beneficial.

After this, we inspect image affinity values computed for every pair of pictures over a subsample of the Caltech data set representing three semantic classes. In the examined subsample we consider 30 images per class, where the image classes are: cars, faces, motorbikes.

The obtained affinity values are shown in Figure 4 together with a baseline calculated as median SIFT similarity.

One may notice that the class structure, i.e. square blocks of brighter regions that correspond to each of the three classes, is more pronounced in the case of the proposed method, as can be seen in Figure 4(a). This confirms our earlier suggestion that image affinity measure described in section 2.3 may serve as a suitable instrument for inter-image prototype selection. In contrast, the baseline technique relying on the median of all SIFT descriptor similarities across all pairs of images does not show as much class discriminatory power, as can depicted in Figure 4(b).

Finally, we examine how the various prototype selection approaches suggested in section 2.4 manage to select relevant descriptors from the groups of images sampled from the three classes of the Caltech data set [15], as shown in Figures 5 to 8. Here, those SIFT descriptors that are selected as class prototypes are shown as yellow colored circles. As one may notice from the shown example images, the prototype selection method that relies on affinity-weighted importance appears to select SIFT

descriptors that all belong to the visual objects of interest, albeit in a somewhat greedy fashion (see Figure 5). Also, both prototype selection techniques that use importance ranks (see Figures 6 and 7) tend to choose SIFT descriptors over a larger number of images and pick up some unrelated background elements. The latter shortcoming is more apparent with the importance rank selection method that disregards the image affinity information altogether, as can be seen in Figure 7.

While these examples may be helpful in probing and analyzing the traits and properties of the proposed prototype selection methods, they cannot replace an overall evaluation that must be carried out on all of the data set, which is the main focus of the section to follow.

3.2 Evaluation

In our experimental evaluation we use the following experimental setup. The experimental data is a subset of the Caltech data set that consists of 1155 images of cars, 450 images of faces and 826 images of motorbikes. This data is subdivided into the training and testing portions. A hundred images per class are used for training, while all of the remaining 2131 images constitute the testing data. Then, within each group of training images belonging a certain class, we select a predefined number of SIFT descriptors across all images as representative class prototypes via the techniques described in section 2.4. Having selected these SIFT descriptors, we consider them as a pseudo-image, against which test images may be compared. Finally, we rank the test images according to their image affinity (Equation (11)) with the class representative pseudo-images, and compute the non-interpolated average precision [16] for each of the three image classes.

It is important here to recall that the spatial information of the descriptors comprising the pseudo-image need not be known when matching a given test image to a pseudo-image using the proposed method of spatially-consistent descriptor matching. This explains why the descriptors of the pseudo-image may or may not necessarily come from the same training image.

In every experiment, we evaluate the proposed techniques alongside the baseline prototype selection method that discards the spatial information found in SIFT descriptors. The results of these experiments are summarized in Tables 1 and 2.

Table 1

Non-interpolated average precision (%) vs. prototype set cardinality

Prototype set cardinality	**Image affinity**	**Baseline**
25	49.83	37.70
50	49.60	43.54
100	48.67	40.09
150	48.04	40.18

The first table shows that even a relatively small number of prototype descriptors selected as a representative pseudo-image for a given class is enough to achieve reasonable results that compare favorably with the baseline technique.

As for the second table, one may observe that, across all of the evaluated prototype selection techniques, the ranking according to image affinity provides better results. It can also be seen from Table 2 that there still room for improvement for the individual prototype selection methods.

Table 2

Non-interpolated average precision (%) vs. prototype selection method

Prototype selection method	**Image affinity**	**Baseline**
Affinity-weighted importance, q_1	39.52	35.17
Affinity-weighted importance rank, q_2	47.09	36.66
Importance rank, q_3	38.00	35.24
SIFT feature similarity	49.83	37.70

4 Conclusion

We have introduced a spatially-consistent descriptor matching method and demonstrated its possible application in the domain of content-based image retrieval. The developed approach incorporates descriptor proximity data when the matching is computed to make sure that the quality of a match of a given descriptor depends on both descriptor itself and its neighbors.

The proposed matching method has been formulated and shown to be equivalent to a standard eigenvalue problem, where the principal eigenvector's components render the importance values of individual descriptors, while the corresponding eigenvalue represents an estimate of the overall strength of affinity between images being matched. These measures of descriptor importance and image affinity have been used as a natural basis for some new intra- and inter-image prototype selection techniques, several variations of which have been empirically evaluated on the task of content-based image retrieval, demonstrating encouraging results.

References

[1] K. Mikolajczyk, C. Schmid, A Performance Evaluation of Local Descriptors, in: International Conference on Computer Vision and Pattern Recognition, Vol. 2, 2003, pp. 257–263.

[2] J. Matas, O. Chum, U. Martin, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: P. L. Rosin, D. Marshall (Eds.), Proceedings of the British Machine Vision Conference, Vol. 1, BMVA, London, UK, 2002, pp. 384–393.

[3] D. G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[4] H. Bay, T. Tuytelaars, L. J. V. Gool, SURF: Speeded up robust features., in: ECCV (1), 2006, pp. 404–417.

[5] Y. Ke, R. Sukthankar, PCA-SIFT: A more distinctive representation for local image descriptors, cvpr 02 (2004) 506–513.

[6] E. N. Mortensen, H. Deng, L. Shapiro, A sift descriptor with global context, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 1, 2005, pp. 184–190.

[7] C. Schmid, R. Mohr, Local grayvalue invariants for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 19 (5) (1997) 530–535.

[8] M. A. Fischler, R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, in: Readings in computer vision: issues, problems, principles, and paradigms, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987, pp. 726–740.

[9] D. West, Introduction to Graph Theory, Prentice Hall, Upper Saddle River, NJ, 1996.

[10] R. A. Horn, C. R. Johnson, Matrix analysis, Cambridge University Press, New York, NY, USA, 1986.

[11] G. H. Golub, C. F. V. Loan, Matrix Computations, 2nd Edition, Baltimore, MD, USA, 1989.

[12] T. Lindeberg, Feature detection with automatic scale selection, International Journal of Computer Vision 30 (2) (1998) 77–116.
URL citeseer.ist.psu.edu/lindeberg98feature.html

[13] D. B. Skalak, Prototype selection for composite nearest neighbor classifiers, Ph.D. thesis, University of Massachusetts, Amherst, MA, USA (1997).

[14] E. Fix, J. Hodges, Discriminatory analysis: Nonparametric discrimination: Consistency properties, Tech. Rep. 4, USAF School of Aviation Medicine (February 1951).

[15] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsupervised scale-invariant learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, Madison, Wisconsin, 2003, pp. 264–271.

[16] A. F. Smeaton, P. Over, W. Kraaij, Evaluation campaigns and trecvid, in: MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval, ACM Press, New York, NY, USA, 2006, pp. 321–330.

Detailed list of change following the reviewer comments

4.1 R1

Experiments with more classes and more images per class ought to be conducted.
Unfortunately, the new results we wish to obtain with the PASCAL Video Object Challenge (VOC) are not ready yet, and thus cannot be incorporated in the final version of the article. In place we propose an evaluation made on a large subset (2500 images) of the well know Caltech image data set. This corpus is widely used in computer vision and permits an easy comparison with the state of the art.

4.2 R2

I continue to ask that in Eqs. (1) and (5) an eigen-value symbol (e.g. λ) should appear. Concerning the "positiveness" of the importance vector I suggest referring to the Perron-Frobenius theorem on the principal eigenvector of a positive matrix.
We modified section 2.1 according to the reviewer comment.

I think that their spatial distribution is restricted. These features could be redundant, while semantically important features might be omitted. A concluding comment on this aspect would be useful for the readers.
This aspect is treated in the initial assumption on the matching rule, *eg* "a descriptor is important for the match if it is similar to some descriptor from the other image and its spatial neigh- bors are important". The spatial consistency rule is a reasonable way to extract semantic feature as explain in the introduction. Making other assumptions to extract more important semantic features makes sense, but should be considered as an extension of the present work.

4.3 R3

The only suggestion I would have is to make the reference to the definition of the average precision - I was confused in my first review by this term.”

Trecvid guidelines are provided as reference on AP.

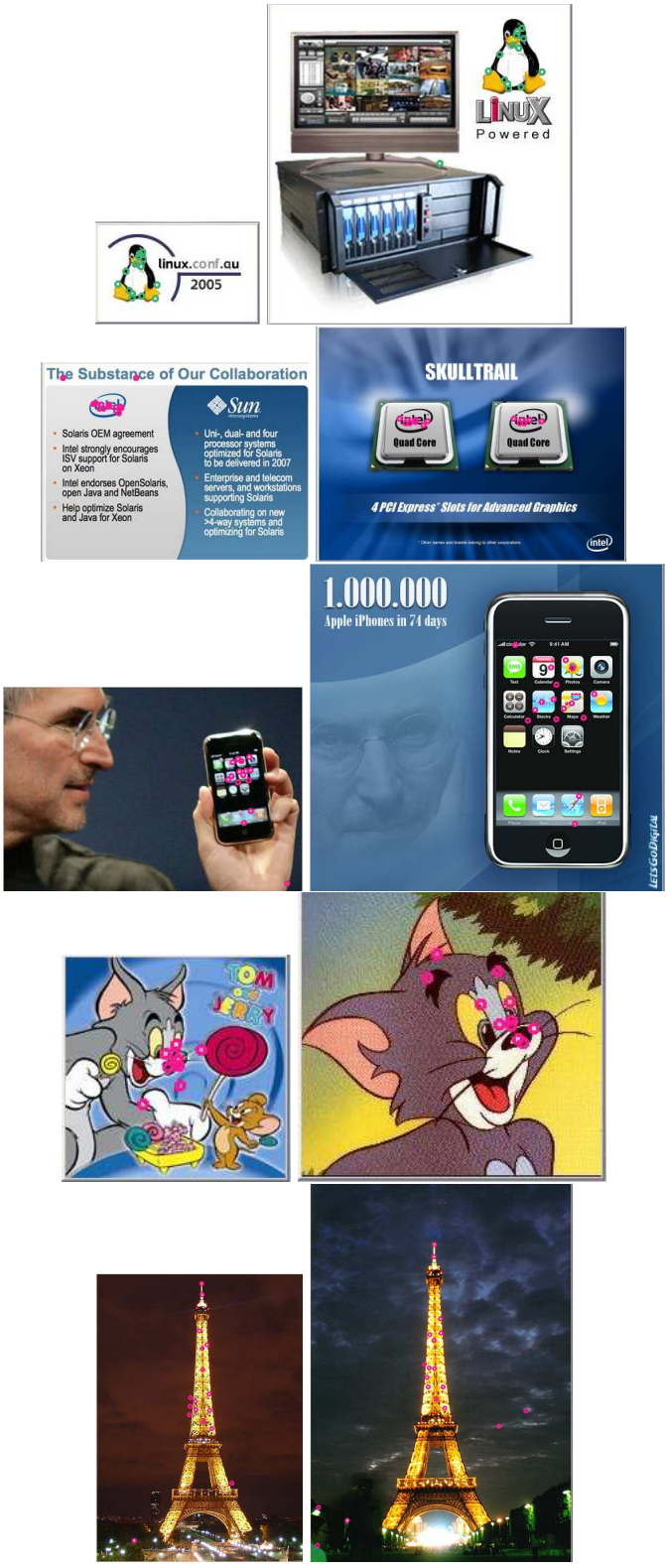
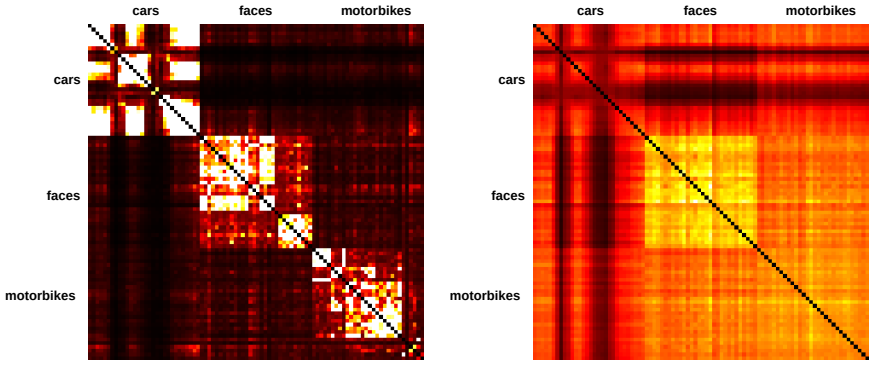


Fig. 3. Example of spatially consistent partial matching. Selected SIFT descriptor locations are shown as solid colored spots. (This is image is best viewed in color).



(A) Image affinity matrix

(B) Median SIFT feature similarity matrix

Fig. 4. Visual comparison of image affinity and SIFT feature similarity values over a subsample of 90 images from the Caltech data set (3 classes, 30 images per class). The brighter regions correspond to larger values. Main diagonal entries are zeroed out, i.e. self-matching excluded.

Fig. 5. Examples of inter-image prototype selection by affinity-weighted importance

Fig. 6. Examples of inter-image prototype selection by affinity-weighted importance rank

Fig. 7. Examples of inter-image prototype selection by importance rank

Fig. 8. Examples of inter-image prototype selection by SIFT feature similarity only (baseline)