Dissimilarity-based Representation for Local parts

A. Carli[1], U. Castellani[1], M. Bicego[1,2], V. Murino[1,2]
[1] Department of Computer Science, University of Verona, Strada Le Grazie 15, 37134 Verona (Italy)
[2] Istituto Italiano di Tecnologia (IIT), Via Morego 30, 16163 Genova (Italy)

***Abstract*—In this paper a novel approach for dissimilarity-based representation is presented, which combines local image descriptors with several dissimilarity functions. The basic idea consists of defining the set of prototypes in terms of local descriptors of image parts, namely feature points extracted from the training set. Therefore, according to the dissimilarity-based approach, a new image can be characterized on the basis of its dissimilarity with each of the given prototypes. This leads to a new class of Local Kernels which exploits the use of dissimilarities between image parts. In particular, we show that the classic Bag-of-Feature (BoF) kernel can be revised as a special case of our new formulation, and better performance can be obtained when new dissimilarity functions are employed. Moreover, we observe that any variants of the basic BoF kernel can take advantage from our approach as we show for the case of the Pyramid Match kernel. Promising results are shown for image categorization on the ETH-80 database.**

I. Introduction

In recent years, the dissimilarity-based representation paradigm [1], [2], [3], [4], [5], [6], [7] has aroused a lively interest in the pattern recognition community. This paradigm differs from typical pattern recognition approaches where objects to be classified are represented by feature vectors. In the dissimilarity-based paradigm, objects are described using pairwise (dis)similarities. In this way, objects are not constrained to be explicitly represented in a feature space, and all that is necessary is a way to compute (dis)similarities between pairs of objects. The goal is then to learn a classifier only from these relational data. The general idea is the following: given a set of pairwise dissimilarity values, a new representation space can be built, in which each object is described by these values. In [5], a simple synthetic experiment shows that a complex problem in a 2D space (requiring a quadratic classifier to achieve almost correct separation), becomes a linearly separable problem in a dissimilarity space. In the original versions, a given object was characterized with dissimilarities/similarities from other objects in the data set, the so-called prototypes – obviously the choice of the prototypes represents a crucial issue (see for example [8]). This idea has been then refined and generalized, considering also dissimilarities from class-models [9], [10], cluster models [11], or even components of a model [12].

In this paper, a novel contribution in this direction is proposed, considering and exploiting dissimilarities between *parts* of a given object. The basic idea consists of defining the set of prototypes in terms of relevant object subparts extracted from the training set.

Part-based characterization of patterns has been largely applied in many computer vision applications, e.g., object categorization and image retrieval, where an image is decomposed into parts (or local features). Here, we focus on the object classification problem in which the objects are images described in term of local parts.

The typical approach to extract image parts is to employ an interest region detector [13]. Such parts are then characterized using a proper descriptor (like SIFT [14], shape context [15], or others [16]). The number of parts may vary from image to image and there is no ordering among features in a single image. As a result images are represented as *variable-sized sets of unordered features*, which makes not possible to apply standard vector-based classification algorithms.

In the literature, this problem has been addressed by the family of the so-called Local Kernels [17], [18], [19], [20], [21], [22]. In [17], the authors propose to look for explicit correspondences between image features [17]. However, although results are satisfying, the method leads to a non-Mercer kernel [20]) and therefore, it is not safely usable as a kernel function without special wariness. In [19], the intermediate matching kernel is proposed by introducing an intermediary set of so-called virtual local features to select the pairs of image subparts to be matched. In this fashion, the intermediate matching kernel mimics matching algorithms while being positive definite. In [18], a widely applied technique has been introduced, namely the Bag of Keypoints (BoK). Being inspired by the Bag-of-Words approach for text classification, this method consists in transforming the set of features in a histogram, which counts the number occurrences in that image of a given set of visual words (i.e., prototype features). The Bag-of-Feature approach has been extended in [20], [21] by combining the number of bins in a hierarchical fashion leading to the so called Pyramid-kernel-matching paradigm. Similarly, such hierarchical approach has been successfully exploited also in the spatial domain in [22].

In this paper, we will approach the above mentioned problem using concepts and tools of the dissimilarity-based representation paradigm, showing that some of the Local Kernels (like BoK) are just special cases of our methodology. This may also open the new possibility of directly applying some of the results recently proposed in the general dissimilarity-based classification paradigm to the image classification problem. Some experiments on a scene categorization task (using the ETH-80 dataset [23]) show the effectiveness of the proposed approach, also in comparison to other Local Kernels present in the state of the art.

The rest of the paper is organized as follows. In Section II, we analyze the problem in detail and describe the proposed

978-1-4244-6459-3/10/$26.00 ©2010 IEEE

methodology; in Section III, we present the experimental results, and Section IV concludes the paper with a discussion of future research.

II. The proposed methodology

In this section, the proposed methodology is presented in three main stages. We firstly define the new image representation and, then, we define some possible kernels based on such a representation, taking inspiration from the literature on Local Kernels. Finally, we show the relations between the proposed approach and other Local Kernels [18], [19], [20].

A. Dissimilarity-based representation of images

In the part-based formulation, an image is represented as a *set* of unordered feature vectors $X^j = \left\{x_1^j, \ldots, x_{N_j}^j\right\}$ where $x_i^j \in \mathbb{R}^d \ i = 1, \ldots, N_j$ are local descriptors belonging to a d-dimensional space (e.g., 128-dimensional SIFT [14]). The cardinality $N_j = \left|X^j\right|$ may vary across images. The point sets (images) come from the input space $\mathcal{X}$:

$$\mathcal{X} = \left\{X^j\right\} \tag{1}$$

Using the dissimilarity based representation paradigm [7], each image is then represented as vectors of dissimilarities. As explained in Introduction, the dissimilarity may be computed with respect to different entities, like other images, models or others. Here, we adopt an alternative vision, computing dissimilarities with respect to *local parts*. The underlying idea is the following: first, we define a global set of prototypes $\mathcal{P} = \{P_1, \cdots, P_V\}$, which encodes and describes all different local parts of the images of the considered problem. Second, each image is represented as the set of local dissimilarities between the set of image features X^j and each prototype P_k. Such dissimilarities are computed by a certain dissimilarity function $d(X^j, P_k)$, thus resulting in a feature vector of fixed size. The idea is that such vector reflects how many (and which) local features are present in the image, thus representing a significant signature. This is very similar to the basic assumption under the Bag-of-Keypoints approach, which nevertheless, just "counts" the presence of the features, without taking into account the *degree* of the presence (which is actually encoded by the dissimilarity measure).

More formally, we define a mapping

$$\phi\left(\cdot, \mathcal{P}\right) : \mathcal{X} \rightarrow \mathbb{R}^V \tag{2}$$

$$\phi\left(X^j, \mathcal{P}\right) = \left[d\left(X^j, P_1\right), \ldots, d\left(X^j, P_V\right)\right] \tag{3}$$

which defines a V-dimensional vector space, on which a standard vector-based classifier (like a Support Vector Machine) may be used.

As in every dissimilarity-based approach, there are two main problems to be solved:

1) The choice of the prototypes $\mathcal{P}$.
2) The choice of the local dissimilarity function $d(X^j, P_k)$.

The literature on dissimilarity-based representation contains several proposals and suggestions to tackle these tasks. Here, we present some possible options, derived from the specific task we are addressing (image characterization for classification).

1) The choice of the prototypes: We adopt the scheme used in the BoK approach and its variants ([18], [19], [21]), i.e., we use a clustering technique to find out a relevant visual vocabulary. In particular, we define $\mathcal{P} = \mathcal{C}$, where $\mathcal{C} = \{C_1, \ldots, C_V\}$ – C_k, $k = 1, \ldots, V$ – is a cluster of feature vectors. The clustering is computed on the whole set of feature vectors extracted from all the images of the training set. We denote cluster centroids as c_k, $k = 1, \ldots, V$.

2) The choice of the dissimilarity: Given our choice of prototypes, the problem of computing the mapping $\phi\left(X^j, \mathcal{P}\right)$ relies on the computation of the distance $d\left(X^j, C_k\right)$, that is, a distance measure between the two sets. If we represent each cluster C_k with the centroid c_k, we have a simple distance "point to set".

$$d\left(X^j, C_k\right) = d\left(X^j, c_k\right) \tag{4}$$

There are different choices for the distance in (4) (min, mean, median, over all distances). After a preliminary evaluation, in our experiments it turned out that the optimal was the following

$$d_{avr}^{centr}\left(X^j, C_k\right) = d_{avr}^{centr}\left(X^j, c_k\right) = \frac{1}{N_j}\sum_{h=1}^{N_j}\left\|x_h^j - c_k\right\| \tag{5}$$

If we do not assume that the cluster is represented with its centroid, then the distance is a general "set to set" distance. In [24], the authors presented some modification of the Hausdorff distance for object matching. After a preliminary experimental evaluation, we decided to adopt the following one (consistently with the previous case): given two sets A and B,

$$d_{avr}\left(A, B\right) = \frac{1}{N_a}\sum_{a \in A} d\left(a, B\right) \tag{6}$$

where the distance between a point $a \in A$ and the point set B is commonly defined as $d\left(a, B\right) = \min_{b \in B}\left\|a - b\right\|$ and $\left\|a - b\right\|$ is the distance between two points that we assume to be defined as the Euclidean distance [24].

In our case, A and B represent the image X^j and the cluster $C_k \in \mathcal{C}$, respectively. The distances defined in [24] are asymmetric: we can use them as they are or we can try to symmetrize them by applying the following f_{min}, f_{max}, f_{avr}, f_{wavr} operators:

$$f_{min}\left(d\left(A, B\right), d\left(B, A\right)\right) = \min\left(d\left(A, B\right), d\left(B, A\right)\right) \tag{7}$$

$$f_{max}\left(d\left(A, B\right), d\left(B, A\right)\right) = \max\left(d\left(A, B\right), d\left(B, A\right)\right) \tag{8}$$

$$f_{avr}\left(d\left(A, B\right), d\left(B, A\right)\right) = \frac{d\left(A, B\right) + d\left(B, A\right)}{2} \tag{9}$$

$$f_{wavr}\left(d\left(A, B\right), d\left(B, A\right)\right) = \frac{N_a d\left(A, B\right) + N_b d\left(B, A\right)}{N_a + N_b} \tag{10}$$

B. Kernels on dissimilarity-based representations

To solve the classification task, different classifiers may be used in the dissimilarity space. In this paper, in order to contextualize and compare the proposed approach in the Local Kernels domain, we chose Support Vector Machines (SVM) [25]. Different kernels may be defined, and the most popular are:

1) Gaussian Radial Basis Function (RBF) kernel [25]

$$K_{RBF}(X^i, X^j) = k_{RBF}\left(\phi\left(X^i, \mathcal{C}\right), \phi\left(X^j, \mathcal{C}\right)\right) = \\ = e^{-\frac{\left\|\phi\left(X^i,\mathcal{C}\right)-\phi\left(X^j,\mathcal{C}\right)\right\|^2}{2\sigma^2}} \tag{11}$$

2) Histogram Intersection kernel [26]

$$K_{HI}(X^i, X^j) = k_{HI}\left(\phi\left(X^i, \mathcal{C}\right), \phi\left(X^j, \mathcal{C}\right)\right) = \\ = \sum_{k=1}^{V} \min\left(d\left(X^i, C_k\right), d\left(X^j, C_k\right)\right) \tag{12}$$

Other kernels may be used, taking into account the fact that the dissimilarity-based representation we employed is related to distances between sets, but they will not be explored in this work.

C. Relation to other Local Kernels

The BoK [18] approach, for both RBF kernel and Histogram Intersection kernel, defines the matching between two images in terms of histograms of local features. It is interesting to note that the histogram representation may be considered just a special case of our general dissimilarity representation. More in detail, the choice of the prototypes is the same, namely, the visual vocabulary itself. Moreover, the local dissimilarity function can be defined as:

$$d_{BoK}\left(X^j, C_k\right) = \sum_{x_h^j \in X^j} d\left(x_h^j, C_k\right) \tag{13}$$

and

$$d\left(x_h^j, C_k\right) = \begin{cases} 1 & \text{if} \quad x_h^j \in C_k \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where $x_h^j \in C_k$ if $\left\|x_h^j - c_k\right\| \leq \left\|x_h^j - c_v\right\| \ \forall v = 1, \ldots, V$, $v \neq k$.

Moreover, it is worth to note that some strategies appeared in the literature to improve the basic BoK method can easily be extended also in our dissimilarity-based representation. For instance, in the context of the Local Kernels, a recent and quite performing extension has been proposed with the Pyramid Match Kernels [20], [21], which extend the visual vocabulary of the basic BoK by considering a hierarchy of clusterings.

It is evident that also our methodology can be extended similarly in this way. Actually, we can define a novel kernel considering a hierarchical clustering as a set of L levels of prototypes. Then, the following kernel can be defined:

$$K_{PMK}\left(X^i, X^j\right) = \sum_{l=0}^{L-1} w_l k_{HI}\left(\phi\left(X^i, \mathcal{C}^l\right), \left(X^j, \mathcal{C}^l\right)\right) \tag{15}$$

where $\mathcal{C}^l$ is the set of prototypes at level l with V^l elements and w_l are weights (see [20], [21] for more details). As in [20], [21], the Histogram Intersection kernel has been chosen as basic kernel. Again, when Equation 13 is employed as a dissimilarity function the proposed kernel becomes very close to the standard Pyramid Match Kernel. In particular, we considered a variation of Pyramid Match Kernel because in Equation 15 we counted all matchings at every level of the clustering hierarchy, and not just the new matchings as in [20], [21]. Moreover, in the so called vocabulary-guided version of Pyramid Match Kernel [21] a further level of weights is introduced for each bin of the various histograms. Here, we exploit the dissimilarity representation by adopting the proposed dissimilarity function as defined in Equation 6.

III. Results

We report some experimental results on a standard object categorization task using objects from the ETH-80 data set [1] [23]. This data set is made up of eight object classes, with 10 unique objects and 41 views for each. In our experiments, we used five widely separated views of each object (i.e. a subset of the whole ETH-80 data set), for a total of 400 images (see Figure 1).

The standard validation protocol, as described in [27], has been adopted. The Harris affine [13] detector is used to find interest points in each image and SIFT local descriptors are applied to generate the feature set (i.e., x_i^j, $i = 1, \ldots, N_j$, is a SIFT descriptor with 128 dimensions). A one-versus-all SVM classifier is trained, and performance is measured via cross-validation[2], where all five views of an object are held out at once. Since no instances of a test object are ever present in the training set, this is a categorization task (as opposed to recognition of the same specific object).

We compared the proposed kernels with the original BoK [18] and Pyramid Match Kernel (PMK) [20], [21], which represents the state of the art for this dataset. To run experiments, we used the original C++ implementation of PMK [3] and extended it to implement the new proposed kernels based on the dissimilarity representations.

The first set of experiments was devoted to compare the standard BoK rule with its generalization (note that the basic kernel is the same, the only change is in the definition of the feature vector – histogram vs dissimilarities). We used the distance d_{avr} defined in Equation 6 with operators defined in Equations 7–10. For the clustering, we employed K-means, ranging the number of clusters from 100 to 300.

Figure 2 shows the performance of results by using the RBF kernel as basic kernel. It is evident that the proposed generalization[4] outperforms the classic BoK approach.

As can be seen from Fig. 2 the best accuracy is reached with d_{avr} used in conjunction with f_{avr} operator for almost

[1] http://www.mis.informatik.tu-darmstadt.de/Research/Projects/

[2] Also parameter C of the SVM is estimated by cross-validation.

[3] LIBPMK: A Pyramid Match Toolkit – http://people.csail.mit.edu/jjl/libpmk/

[4] We excluded the rule of Equation 7 since it produces very bad results.

Fig. 1. Example images from the ETH-80 objects database. Five images from each of the eight object classes (apple, cow, dog, pear, car, cup, horse, and tomato) are shown here.

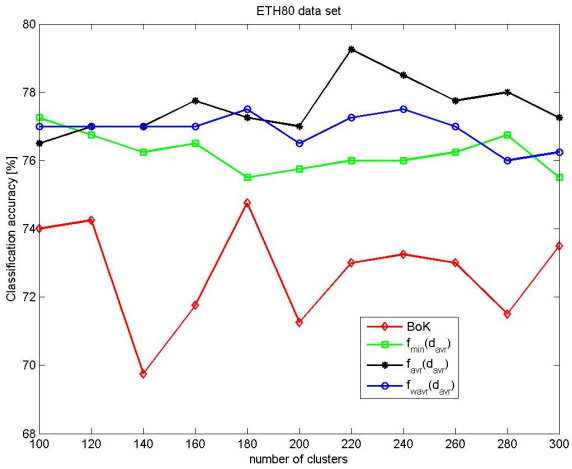

Fig. 2. Classification accuracy for the ETH-80 data set: comparison between the BoK and the proposed dissimilarity representation using distance d_{avr} with operators f_{min}, f_{avr}, f_{wavr} [24], for number of clusters ranging from 100 to 300. The RBF kernel is used as basic kernel.

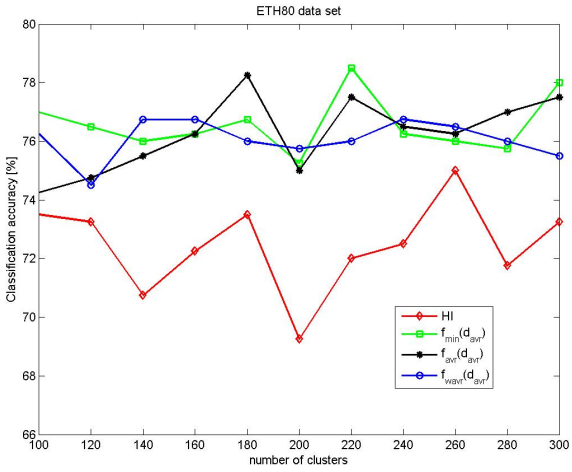

Fig. 3. Classification accuracy for the ETH-80 data set: comparison between the BoK and the proposed dissimilarity representation using distance d_{avr} with operators f_{min}, f_{avr}, f_{wavr}, for number of clusters ranging from 100 to 300. The histogram intersection kernel is used as basic kernel.

all clusters.

As a second set of tests, we repeated the previous experiments but considering the Histogram Intersection kernel as basic kernel. The obtained results are shown in Fig. 3. Also in this case, it is evident the beneficial effect of employing the proposed representation.

In summary, from the proposed results we noted that, independently from the basic kernel (RBF or HI), d_{avr} with operators f_{min}, f_{avr}, f_{wavr} leads to comparable performances resulting in an improvement over classical binary representation (histograms). Note that the computational cost of the proposed methods depends by the chosen similarity measure. Of course, when a set-to-set distance is employed the computational efforts increase.

A final test has been made in order to compare the Pyramid Match Kernel with our hierarchical version, defined in Equation 15. It is worth noting that we did not define a weighting scheme as in [20], [21], i.e. we considered $w_l = 1$ $\forall l = 1, \ldots, L$. Note that this is coherent with the dissimilarity variation introduced by our method. Actually, in the original formulation the weights of the PMK are necessary in order to assign a higher score to the matching at the finest levels and viceversa. Conversely, in our approach this principle is not applicable since the score is accumulated by the dissimilarity functions and not by binary values. A more sophisticated approach can be applied by exploiting recent methods on multiple kernel learning [28].

Using the original C++ implementation of PMK[5] and our extension of it to implement the proposed kernels, the experimental protocol of [27] and 149 SIFT per image on average, our method reached a classification accuracy of 81.5%, whereas the vocabulary-guided (i.e., all the weights are properly estimated) PMK reached a classification rate of 77%.

IV. Conclusions

In this paper, a new approach for learning from pairwise relationships is proposed by showing the effectiveness of the dissimilarity-based approach to encode images in terms of local parts. Our approach leads to a new class of dissimilarity-based Local Kernels which are able to compare images represented by unordered set of feature points with variable size. Our approach can be considered as a generalization of the BoK approach, and its extension for which, instead of simply notifying the presence of the features in a set, we consider also the degree of the presence. To this aim, we tested several variations of Hausdorff distance between feature sets as dissimilarity function. Pomising results have been achieved on a standard benchmark data set for image categorization, namely the ETH-80. The proposed framework outperforms the Bag-of-Feature approach for all the presented dissimilarity representations. Moreover, our method outperforms also the state-of-the art Local Kernels, i.e., the Pyramid Match Kernel. Future work will be devoted to the application of other dissimilarity-based classification strategies on the Local Kernels domain for the image classification problem. In particular, new choices of the prototypes will be investigated in order to exploit the dissimilarity between models (i.e., generative models) estimated from local parts, or the model components.

Acknowledgments

We acknowledge financial support from the FET programme within the EU-FP7, under the SIMBAD project (contract 213250)

References

[1] A. Jain and D. Zongker, "Representation and recognition of handwritten digits using deformable templates," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 12, pp. 1386–1391, 1997.

[2] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, and K. Obermayer, "Classification on pairwise proximity data," in *Advances in Neural Information Processing*, D. C. M. Kearns, S. Solla, Ed., vol. 11. MIT Press, 1999.

[3] D. Jacobs and D. Weinshall, "Classification with nonmetric distances: Image retrieval and class representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 583–600, 2000.

[4] E. Pekalska and R. Duin, "Automatic pattern recognition by similarity representations," *Electronics Letters*, vol. 37, no. 3, pp. 159–160, 2001.

[5] E. Pekalska, P. Paclik, and R. Duin, "A generalized kernel approach to dissimilarity-based classification," *Journal of Machine Learning Research*, vol. 2, no. 2, pp. 175–211, 2002.

[6] E. Pekalska and R. Duin, "Dissimilarity representations allow for building good classifiers," *Pattern Recognition Letters*, vol. 23, no. 8, pp. 943–956, 2002.

[7] E. Pekalska and R. P. Duin, *The dissimilarity representation for Pattern Recongition - Foundations and Applications*. World Scientific, 2005.

[8] E. Pekalska, R. Duin, and P. Paclik, "Prototype selection for dissimilarity-based classifiers," *Pattern Recognition*, vol. 39, no. 2, pp. 189–208, 2006.

[9] C. Lai, D. Tax, R. Duin, E. Pekalska, and P. Paclik, "A study on combining image representations for image classification and retrieval," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, no. 5, pp. 867–890, 2004.

[10] M. Bicego, V. Murino, and M. Figueiredo, "Similarity-based classification of sequences using hidden markov models," *Pattern Recognition*, vol. 37, no. 12, pp. 2281–2291, 2004.

[11] M. Bicego, E. Pekalska, and R. Duin, "Group-induced vector spaces," in *Multiple Classifier Systems*, M. Haindl, J. Kittler, and F. Roli, Eds. Springer, 2007, vol. LNCS 4472, pp. 190–199.

[12] M. Bicego, E. Pekalska, D. Tax, and R. Duin, "Component-based discriminative classification for hidden markov models," *Pattern Recognition*, vol. in press, 2009.

[13] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A Comparison of Affine Region Detectors," *International Journal on Computer Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[15] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[16] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[17] C. Wallraven, B. Caputo, and A. Graf, "Recognition with local features: the kernel recipe," in *Proceedings of International Conference on Computer Vision (ICCV03)*, 2003.

[18] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

[19] S. Boughorbel, J.-P. Tarel, and N. Boujemaa, "The intermediate matching kernel for image local features," in *Proceedings of International Joint Conference on Neural Networks (IJCNN'05)*, Montréal, Canada, 2005, pp. 889–894.

[20] K. Grauman and T. Darrell, "The pyramid match kernel: Efficient learning with sets of features," *Journal of Machine Learning Research*, vol. 8, pp. 725–760, 2007.

[21] ——, "Approximate correspondences in high dimensions," in *Advances in Neural Information Processing Systems*, 2007.

[22] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. II, New York, 2006, pp. 2169–2178.

[23] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *Proceedings of International Conference on Computer Vision and Pattern Recognition 2003 (CVPR03)*, 2003.

[24] M. Dubuisson and A. Jain, "Modified hausdorff distance for object matching," in *Proceedings of the International Conference on Pattern recognition*, vol. 1, 1994, pp. 566–568.

[25] B. Schölkopf and A. J. Smola, *Learning with Kernels*. MIT Press, 2002.

[26] F. Odone, A. Barla, and A. Verri, "Building kernels from binary strings for image matching," *Transactions on Image Processing*, vol. 14, no. 2, pp. 169–180, 2005.

[27] J. Eichhorn and O. Chapelle, "Object categorization with SVM: kernels for local features," Max Planck Institute for Biological Cybernetics, Tech. Rep., 2004.

[28] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *Proceedings of the International Conference on Machine Learning*, June 2009, pp. 1065–1072.

[5]The hierarchical-K-means is applied to obtain hierarchical clustering.