Unsupervised classemes

Claudio Cusano[1], Riccardo Satta[2], and Simone Santini[3⋆]

[1] Department of Informatics, Systems and Communication (DISCo), Università degli Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy.
[2] Department of Electrical and Electronic Engineering, Università di cagliari, Italy
[3] Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain

Abstract. In this paper we present a new model of *semantic* features that, unlike previously presented methods, does not rely on the presence of a labeled training data base, as the creation of the feature extraction function is done in an unsupervised manner.
We test these features on an unsupervised classification (clustering) task, and show that they outperform primitive (low-level) features, and that have performance comparable to that of supervised semantic features, which are much more expensive to determine relying on the presence of a labeled training set to train the feature extraction function.

1 Introduction

Several authors have proposed, in the last few years, the use of class information in the definition of features or, to say it in a nother way, the definition of features based not only on the image data but also on certain, limited, semantic information. The general schema of these features is as follows. Consider a reference database of images $D = \{x_1, \ldots, x_n\}$, and a reference partition of D into classes (according to some semantically meaningful criterion), $\mathcal{D}$, with $\mathcal{D} = \{D_1, \ldots, D_q\}$, $\bigcup_i D_i = D$, and for all $i \neq j$, $D_i \cap D_j = \emptyset$. The subsets D_i may or may not have associated labels. Also, let $\mathcal{I}$ be a space of images, and $\phi : \mathcal{I} \to X$ a feature extractor function from images to a suitable feature space X.

These features and the reference data base are used to train, in a supervised way, q classifiers $C_1, \ldots C_q$ with $C_i : X \to [0, 1]$ determining the degree to which the feature vector X belongs to the category i. The outputs of these classifiers are then used, in combination with the low level feature extractor ϕ, to create a semantic feature vector for an unknown image x. That is, an image is represented by the q-dimensional vector

$$F(x) = [C_1(\phi(x)), \ldots, C_q(\phi(x))] \tag{1}$$

we call $F : \mathcal{I} \to [0, 1]^q$ a *supervised semantic* feature extraction function. This general scheme is sometimes implemented with a slight modification, which has

⋆ The authos was supported in part by the *Ministerio de Educación y Ciencia* under the grant N. TIN2011-28538-C02, *Novelty, diversity, context and time: new dimensions in next-generation information retrieval and recommender systems.*

been shown to improve perfoamance in certain cases. The feature space X is partitioned into p parts (the nature of this partition depends on the specific method used), and q classifiers corresponding to the q categories are trained separately on these p partitions. In this case, the feature vector will have dimension pq. The recent literature proposes several instantiations of this general scheme. Vogel and Schiele presented an image representation formed by local semantic descriptions [18]. They classify local image regions into semantic concept classes such as water, rocks, or foliage. Images are represented through the frequency of occurrence of these local concepts. Li *et al.* [12] in their *object bank* system use 177 object recognizers at 12 different scales, over 21 image regions, obtaining a vector of $177 \times 12 \times 21 = 44,604$ components.

Ciocca *et al.* presented an image descriptor, that they called "prosemantic features", based on the output of a number of image classifiers [3]. The feature vector is created, by concatenating the output of 56 different soft classifiers trained to identify 14 different classes on the basis of four different low-level features.

Torresani *et al.* [17] presented a descriptor composed of the output of 2,659 classifiers trained to recognized scene-level visual concepts (called "classemes") taken from the LSCOM ontology [13].

In [4] we tested these *supervised* features against a representative sample of *primitive* features (low-level features extracted directly fom the images) on an unsupervised classification task. We wanted to test whether the use of supervised learning in the feature definition phase would help in the unsupervised classification of hitherto unseen classes. We performed tests on three different data bases: the Simplicity data set [19] (a subset of the Corel data set), the GIST scene data set [14], and Li and Fei-Fei's event data set [11]. The results were that supervised features outperform primitive ones in all the unsupervised classification tasks. The relative merits of the three types of features that were tested (classemes, prosemantic, object bank) are more debatable and depend on the data set on which we are testing, but the advantage of using semantic information in order to define the features have been clearly estabished.

The work in this paper begins with an important observation from [17]:

> It is not required or expected that these base categories will provide useful semantic labels [...]. On the contrary, we work on the assumption that modern category recognizers are essentially quite dumb: so a `swimmer` recognizer looks mainly for water texture, and the `bomber␣plane` recognizer contains some tuning for the "C" shapes corresponding to the airplane nose and perhaps the "V" shapes at the wing and tail. [..] The true building blocks are not the classeme labels that we can see, but their underlying dumb components, which we cannot.

If this observation is correct, then the class labels (that is, the *a priori* division into semantically defined classes) adds nothing, in the best scenario, to the classifiers. In the worst scenario, the use of an *a priori* division can actually be counterproductive, as class divisions might cut across the *underlying dumb*

components that are what the classifiers really identify. For example, the existence of two classes, `airplane` and `bird` forces the classifiers to disregard the "C" shape of the nose, which is a feature common to two classes. A classifier not forced to distinguish between planes and birds could make a better use of this important low-level clue.

In this paper we propose to replace the *supervised* classification used to train the classifiers $C_1, \ldots, C_q$ with *unsupervised* classification, in which the only external constraint is the number of classes q, which determines the dimension of the semantic feature space. We see two main advantages in this: on the one hand, it will be possible to derive the feature computing function using an *unlabeled* training set, rather than a labeled one as it is the case for supervised features. On the other hand, we hope that the classifiers will "latch" directly to the significant visual cues that determine the structure of the training set rather than having categories imposed from outside.

2 Unsupervised classeme

Our method is a simple modification of the supervised (semantic) feature extraction. We are given a training data set D and, possibly, a desired dimensionality q. The method operates as follows:

i) define a *low-level* feature extractor $\phi : \mathcal{I} \to X$ for a suitable low-level feature space X;
ii) use a clustering algorithm on the feature space X to divide the data base D into q classes $D_1, \ldots, D_q$;
iii) use the classes $D_1, \ldots, D_q$ as ground truth in order to train q classifiers $C_1, \ldots, C_q$ with outputs in $[0, 1]$ to recognize the classes.
iv) define the *unsupervised classemes* feature vector for an image x as

$$U(x) = [C_1(\phi(x)), \ldots, C_q(\phi(x))] \tag{2}$$

with $U : \mathcal{I} \to [0, 1]^q$.

The value q can either be a design parameter (this is the case of our tests, in which we used a k-means clustering algorithm for the first unsupervised clustering) or be determined automatically by the system if a clustering algorithm is used that autonomously determines the number of classes (such as affinity propagation [7] of hierarchical clustering [20]). This possibility adds flexibility to the use of these features: if the design calls for a feature vector of a specific size, then the designer can use k-means, thus establishing the number of classes and therefore the size of the vector. Otherwise, the designer can use a clustering algorithm without a predefined number of clusers and let the system determine the size of the feature vector based on the characteristics of the training set.

Note that although the classifiers $C_1, \ldots, C_q$ are trained using a supervised algorithm, the method, as a whole, is unsupervised. This gives it an important advantage with respect to supervised semantic features, as the determination of the feature extraction function doesn't need a labeled data base.

2.1 Implementation details

In the tests that we propose in this paper, we used four feature sets: a RGB histogram, the first and second YUV moments on a 9×9 subdivision, an edge direction histograms (EDH) computed on a 8×8 subdivision, and bag of SIFT descriptors. Each one of the four features was considered separately. Moreover, each of the feature spaces was partitioned in four sub-spaces, and classifiers were trained separately for each one of them (e.g. the 512 bin RGB histogram was divided in four sub-histograms of 128 bins each), except the YUV block-histogram, which was partitioned in three sub-spaces. In the end, the feature space was partitioned in $p = 4 + 4 + 4 + 3 = 15$ sub-spaces.

Each sub-space of the partition resulted in q clusters (so that we have a total of $15q$ clusters, which is also the dimension of the final feature vector). The number of clusters was 5 or 10 depending on the test (see below), giving us feature vectors of size 150 or 75.

For each sub-space of the partition, we trained classifiers to recognize the q classes derived from the clusters, using for each cluster a one-vs-all training (samples from the selected cluster are used as positive examples, and samples from all the others are used as negative examples, subsampling the negative classes so as to obtain a balanced set). The classifiers $C_1, \ldots, C_{15q}$ were SVM with Gaussian kernels.

3 The comparison methods

We have compared our unsupervised classemes with ten other features on an unsupervised classification problem, three *supervised* methods, and seven primitive features.

3.1 Classemes

Torresani *et al.* use as feature vector the output of a large number of weakly trained object category classifiers [17].

In terms of our schema, the low level space X was composed of Color GIST [15], Pyramid of Histograms of Oriented Gradients [5], Pyramid self-similarity [1], and bag of SIFT descriptors. The space was not partitioned. A large number of classifiers ($q = 2659$) was trained on categories taken from the LSCOM ontology [13]. Each classifier has been trained one-vs-all on a category with the LP-β multi-kernel algorithm [8].

Classemes has been presented as a descriptor for image retrieval. Torresani *et al.* have shown that classification accuracy on object category recognition is comparable with the state of the art, but with a computational cost orders of magnitude lower.

3.2 Prosemantic Features

Prosemantic features are based on the classification of images into a set of 14 categories ($q = 14$): animals, city, close-up, desert, flowers, forest, indoor, mountain, night, people, rural, sea, street, and sunset. Some classes describe the image

at a scene level (city, close-up, desert, forest, indoor, mountain, night, rural, sea, street, sunset), while other describe the main subject of the picture (animals, flowers, people).

The low-level feature space X is defined by four low-level features. The space is partitioned by features, that is, for each class, four different classifiers are defined, each one based on one of the four low-level categories ($p = 4$).

Each classifier has been independently trained on images downloaded from various image search engines with different parameters. The classifiers' output was normalized by a linear transformation

$$\phi'_{c,p}(x) = a_{c,p}\phi_{c,p}(x) + b_{c,p}, \tag{3}$$

where the parameters $a_{c,p}$ and $b_{c,p}$ are determined by a logistic regression which maps the score of the classifier to an estimate of the posterior probability

$$p(c|x) \simeq (1 + \exp(-\phi'_{c,p}(x)))^{-1}. \tag{4}$$

3.3 Object Bank

Object Bank is an image representation constructed from the responses of many object detectors, which can be viewed as a "generalized object convolution" [12]. Two state-of-the-art detectors are used: the latent SVM object detectors [6] for most of the blobby objects such as tables, cars, humans, etc, and a texture classifier [9] for more texture-based objects such as sky, road, sand, etc.

A large number of object detectors are run across an image at different scales. Each scale and each detector yield an initial response map of the image. The authors used 177 object detectors at 12 detection scales. Each response map is then aggregated according to a spatial pyramid of three levels ($1 + 4 + 16 = 21$ regions). The final descriptor is thus composed of $177 \times 12 \times 21 = 44,604$ components ($q = 44,604$).

The authors evaluated the object bank descriptor in the context of scene categorization. By using linear classifiers, they obtained a significant improvement against low-level representations on a variety of data sets.

3.4 Primitive Features

We also compared our feature vector with seven state of the art primitive features. We considered three descriptors defined by the MPEG-7 standard–namely the Scalable Color Descrptor (SCD), the Color Layout Descriptor (CLD), and the Edge Histogram Descriptor (EHD) [16]–the Color and Edge Directivity Descriptor (CEDD) [2], the Gist features [14], bag of features [21], and the spatial pyramid representation [10].

4 The data sets

We tested these feature vectors for unsupervised classification on three data bases representative of different situations common in the applications. The first

experiment has been conducted on the Simplicity data set [19], which is a subset of the COREL data set, formed by ten categories containing 100 images each. It can be considered an "easy" data set, since the ten categories are clearly distinct, with little or no ambiguity. On the one hand, this restricts the significance of the experimentation, but on the other hand, it makes the results more reliable (since there is only a single, reasonable way of dividing the data in ten meaningful clusters).

The second group of tests was done on the scene recognition data set collected by Oliva and Torralba [14] to evaluate features and methods for scene classification. This data set contains eight outdoor scene categories: coast, mountain, forest, open country, street, inside city, tall buildings and highways, for a total of 2,688 images (260–410 images per class). With respect to the Simplicity data set, there is less inter-class variability, and are therefore the classes are expected to be harder to separate.

The third data set considered contains images of eight different classes of events [11]. This data set has been collected in order to evaluate event classification methods. It is composed of 1,579 images (eight classes, 137–250 images per class) showing people performing various sport activities (rock climbing, rowing, badminton, bocce, croquet, polo, sailing, and snowboarding). Of the three data sets, this is undoubtedly the most challenging, as events can't be classified only at a scene level, but object detection and pose recognition are often required.

5 The tests

For each of the three data bases that we are using, we have a ground truth with a definite number of categories (ten for Simplicity, eight for the Torralba data base and for the Event data set). Since our purpose here was to test the features, we avoided possible instabilities deriving from the performance of the clustering algorithm by using k-means, where k was set to the same number of classes as the ground truth. In the following, to avoid confusion, we shall refer to the classes discovered by the clustering algorithm as *u-classes*, and to the actual (ground truth) classes as *g-classes*.

In order to compute the classification rate, for each one of the u-classes obtained by the clustering algorithm, we determine which one of the g-classes is most represented, and assign that u-class to it. This entails that several u-classes can be taken as representatives of the same g-class. In this case, all images of the g-class classified in either of the u-classes is taken as correct. An example of this can be seen in table 2 (matrix on the bottom-left), which shows the confusion matrix for the classeme features on the Simplicity data set. In both u-classes 6 and 7, the most represented g-class is *Horses*, so both these u-classes are considered as representatives of Horses, and both the 44 images of horses in u-class 6 and the 43 images in u-class 7 are considered as correct classifications. On the other hand, it may happen that a g-class is not the most represented class in any u-class. This is the case, in the same matrix, of the class *Elephants*. In this case, all elephant images are considered misclassified.

The results are summarized in table 1. A first general observation is that

Feature	Simplicity	Scene	Events
Unsupervised classemes	81.7	65.0	57.0
Classemes	65.0	76.4	62.0
Prosemantic	73.7	78.3	64.9
Object Bank	57.8	70.0	43.1
GIST	33.7	57.1	46.2
Bag of SIFT	49.0	39.1	36.6
Spatial Pyramid	47.4	43.0	36.7
CEDD	62.2	38.3	40.4
SCD	42.3	27.1	27.9
CLD	54.1	32.1	32.1
EHD	50.4	59.5	49.6

Table 1. Summary of the classification results for the ten methods on the three test data bases.

unsupervised classemes work better than supervised semantic on the Simplicity data base and slightly worse on the scene and event data base. In any case, they outperform all primitive features. Apart from the better performance of unsupervised classemes on the Simplicity data base, this is in line with what one can intuitively expect: unsupervised classemes use more information than primitive features (they use a training set of images) but less than supervised (they don't need a labeled data set), and their performance is placed accordingly. Nevertheless, in some cases, unsupervised classemes can outperform semantic features, despite the absence of labels.

In order to analyze more closely these differences, we show, in tables 2–4 the confusion matrices for unsupervised and supervised semantic features. If we consider again Torresani's observation reported in the introduction, it is clear that unsupervised classemes *lock-in* to the most salient visual features independently of the presence of labels, and form clusters based on these salient features, while supervised features have to find a compromise between salient visual features and the labels that are given to them. If we compare unsupervised classemes with prosemantic on the Simplicity data base, we notice, for instance, the 68 misclassifications of prosemantics between Africa people and Food. Some pictures of Africa people have a color structure similar to pictures of food, so the labeling would in this case give indications contrary to the visual features, and this probably provokes the creation of weaker classifiers (viz. classifiers with a narrower margin). The same happens in the case of classemes between horses and elephants, to the extent that the g-category *Elephants* disappears, and two u-categories are assigned to *Horse*. On the other hand, in the scene data base, we see that, in spite of the lower overall performances, unsupervised classemes

	Unsupervised classemes										Prosemantic									
Class	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
Africa people	**79**	-	-	7	-	1	6	3	3	1	2	-	6	-	68	4	-	18	-	2
Buses	-	**93**	-	1	-	-	1	5	-	-	**94**	-	-	-	-	-	5	1	-	-
Dinosaurs	-	-	**100**	-	-	-	-	-	-	-	-	**100**	-	-	-	-	-	-	-	-
Elephants	3	-	-	**77**	-	-	19	-	-	1	-	-	**61**	-	3	11	-	9	14	2
Flowers	3	-	-	1	**87**	2	-	1	6	-	-	-	-	**99**	1	-	-	-	-	-
Food	4	-	-	15	2	**75**	-	1	2	1	-	5	1	11	**71**	-	-	12	-	-
Horses	-	-	-	-	-	-	**97**	1	2	-	-	-	-	-	-	**88**	-	12	-	-
Monuments	6	1	-	8	2	-	3	**67**	9	4	9	-	4	1	1	1	**56**	**23**	4	1
Mountains	1	-	-	5	-	-	4	2	**74**	14	-	-	7	-	-	-	3	1	**77**	12
Sea	2	-	-	7	-	-	2	4	17	**68**	1	-	6	-	3	-	5	7	10	**68**

	Classemes										Object Bank									
Class	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
Africa people	**61**	-	-	14	8	10	-	5	1	1	**32**	1	-	9	-	41	8	2	6	1
Buses	-	**92**	-	-	-	-	-	6	2	-	-	**97**	-	1	-	-	-	2	-	-
Dinosaurs	-	-	**98**	-	2	-	-	-	-	-	-	-	**99**	-	-	-	-	-	-	1
Elephants	14	-	-	-	1	42	7	3	33	-	1	-	-	**42**	-	1	18	5	14	19
Flowers	5	-	-	**91**	4	-	-	-	-	-	6	-	-	-	**58**	36	-	-	-	-
Food	15	-	-	6	**65**	4	-	-	10	-	10	-	1	4	2	**56**	-	-	25	2
Horses	4	-	-	-	-	**44**	**43**	4	5	-	2	-	-	5	-	-	**66**	8	18	1
Monuments	4	2	1	-	3	5	-	**64**	20	1	12	8	2	13	-	4	2	**39**	13	7
Mountains	2	-	1	1	3	31	-	1	**47**	14	4	-	-	37	-	1	3	2	**31**	22
Sea	1	-	-	1	5	4	-	3	41	**45**	1	-	-	20	-	2	2	3	14	**58**

Table 2. Confusion matrices for unsupervised classemes, Prosemantic, classemes, and Object Bank features on the **Simplicity** data base.

	Unsupervised classemes								Prosemantic							
Class	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
Coast	**224**	3	68	5	19	37	2	2	**294**	2	14	-	27	20	-	3
Forest	-	**311**	-	-	4	13	-	-	-	**291**	-	1	26	10	-	-
Highway	51	3	**158**	18	5	8	13	4	8	1	**206**	7	4	12	14	8
Inside city	8	2	1	**185**	2	2	38	70	-	-	5	**178**	-	2	84	39
Mountain	10	36	9	3	**260**	53	1	2	9	9	2	-	**332**	22	-	-
Open country	36	96	20	2	54	**197**	3	2	24	32	4	-	54	**294**	2	-
Street	-	-	3	43	12	2	**203**	29	-	-	5	32	2	1	**245**	7
Tall building	3	6	15	72	28	11	12	**209**	2	-	2	53	9	1	25	**264**

	Classemes								Object Bank							
Class	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
Coast	**201**	2	92	-	1	62	-	2	**254**	3	1	2	41	58	1	-
Forest	-	**312**	-	-	14	2	-	-	-	**302**	-	23	-	-	3	-
Highway	21	-	**202**	9	5	15	8	-	129	-	6	5	22	86	12	-
Inside City	1	-	-	**249**	-	-	51	7	3	-	**241**	-	-	3	27	34
Mountain	4	17	6	-	**274**	71	2	-	4	29	-	**253**	31	47	8	2
Open Country	11	19	57	-	46	**276**	-	1	37	13	-	54	**205**	**98**	3	-
Street	-	-	1	30	-	-	**259**	2	1	2	9	4	-	9	**248**	19
Tall Building	-	4	-	26	1	-	45	**280**	-	4	53	6	-	1	12	**280**

Table 3. Confusion matrices for the unsupervised classemes, Prosemantic, classemes, and Object Bank features on the **Scene** data base.

correctly identify all the g-classes, although in some cases there is considerable confusion, such as between the classes. Consider, for example, the classes *Mountain* and *open country*. What brings down the performance of unsupervised classemes is in this cases the presence of g-classes poorly separated from the visual point of view. Many images of *Mountain* do satisfy the visual conditions to be considered *open country* and, in the absence of a label that forces the two g-classes to be separated, they form similar clusters, leading to the rlatively high confusion between the two g-classes. Many cases of high-confusion in unsupervised classemes can be ascribed to the presence of classes poorly separable from the visual point of view, with many images that could belong to one or the other (Forest vs. open country, street vs. inside city vs. tall building, etc.). In this case, the supervised features have the advantage that the presence of the label forces the classifiers to focus on certain visual features rather ignoring

others (e.g. the classifier of *tall building* is forced to concentrate on long vertical lines). In the event data base, most methods work rather poorly, as many classes

	Unsupervised classsemes								Prosemantic							
Class	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
Rock climbing	**171**	2	3	6	5	2	2	3	**170**	1	-	-	5	3	-	15
Rowing	3	**165**	12	6	32	14	8	10	5	**156**	6	2	9	25	42	5
Badminton	3	14	**89**	**67**	5	13	4	5	1	3	**168**	6	2	14	3	3
Bocce	38	1	15	13	50	13	1	6	10	1	15	15	60	28	2	6
Croquet	24	2	65	13	**109**	13	-	10	3	2	12	**109**	**76**	31	-	3
Polo	4	2	8	52	29	**83**	-	4	5	1	19	56	13	**82**	-	6
Sailing	5	23	5	9	7	1	**115**	25	-	16	5	-	4	5	**126**	34
Snowboarding	21	5	7	15	20	8	9	**115**	12	13	1	-	5	11	10	**138**

	Classemes								Object Bank							
Class	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
Rock climbing	**158**	-	-	27	-	3	1	5	**97**	**78**	-	-	-	5	3	11
Rowing	1	**187**	6	19	19	8	4	6	2	1	**106**	**76**	17	5	3	40
Badminton	-	1	**150**	5	28	7	-	9	-	-	14	4	**145**	17	8	12
Bocce	7	2	8	82	15	21	1	1	7	7	4	4	16	17	54	28
Croquet	15	2	1	**88**	**83**	44	-	3	5	9	22	44	15	**73**	38	30
Polo	5	3	14	29	36	**90**	3	2	2	1	11	29	24	49	42	24
Sailing	-	15	2	5	12	5	**146**	5	-	2	28	43	41	43	28	5
Snowboarding	8	10	3	44	13	14	21	**77**	27	5	8	16	15	14	**63**	**42**

Table 4. Confusion matrices for unsupervised classemes, Prosemantic, classemes, and Object Bank features on the **Event** data base.

are not distinguishable based on the general characteristics of the scene, their identification depending on the presence of specific objects (such as the presence of a mallet to distinguish *croquet* from *bocce*).

6 Conclusions

We have presented a new model of *semantic* features that, unlike previous methods, does not rely on the presence of a labeled training data base, as the creation of the feature extraction function is done in an unsupervised way.

We have compared our features with three supervised semantic features and seven primitive ones, showing that the performance of unsupervised classemes is definitely better than that of primitive features and of the same order as that of supervised ones. Given the broad availability of large collections of non labeled images, we believe that the method presented here represents a viable alternative to primitive features and supervised semantic features.

References

1. A. Bosch, A. Zisserman, and X. Munoz. Image classification using rois and multiple kernel learning. *International Journal of Computer Vision*, 2008:1–25, 2008.
2. S. Chatzichristofis and Y. Boutalis. CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *Computer Vision Systems*, volume 5008 of *LNCS*, pages 312–322. 2008.
3. Gianluigi Ciocca, Claudio Cusano, Simone Santini, and Raimondo Schettini. Prosemantic features for content-based image retrieval. In *Adaptive Multimedia Retrieval 2009. Understanding Media and Adapting to the User*, volume 6535 of *Lecture Notes in Computer Science*, pages 87–100, 2011.

4. Gianluigi Ciocca, Claudio Cusano, Simone Santini, and Raimondo Schettini. Supervised features for unsupervised image categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. (submitted).
5. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893, 2005.
6. P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.
7. B.J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.
8. P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 221–228, 2009.
9. D. Hoiem, A.A. Efros, and M. Hebert. Automatic photo pop-up. 24(3):577–584, 2005.
10. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–78, 2006.
11. L.J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 1–8, 2007.
12. L.J. Li, H. Su, E.P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. *Advances in Neural Information Processing Systems*, 2010.
13. M. Naphade, J.R. Smith, J. Tesic, S.F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *Multimedia, IEEE*, 13(3):86–91, 2006.
14. A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int'l J. Computer Vision*, 42(3):145–175, 2001.
15. A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.
16. T. Sikora. The MPEG-7 visual standard for content description-an overview. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):696–702, 2001.
17. L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *Computer Vision–ECCV 2010*, pages 776–789, 2010.
18. J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, 72(2):133–157, 2007.
19. J.Z. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.
20. J.H. Jr. Ward. Hierarchical grouping to optimize an objective function. *J. the Am. Statistical Assoc.*, 58(301):236–244, 1963.
21. J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–38, 2007.