Enhancing Semantic Features with Compositional Analysis for Scene Recognition

Miriam Redi and Bernard Merialdo

EURECOM, Sophia Antipolis
2229 Route de Cretes
Sophia Antipolis
{redi,merialdo}@eurecom.fr

Abstract. Scene recognition systems are generally based on features that represent the image semantics by modeling the content depicted in a given image. In this paper we propose a framework for scene recognition that goes beyond the mere visual content analysis by exploiting a new cue for categorization: the image *composition*, namely its photographic style and layout. We extract information about the image composition by storing the values of affective, aesthetic and artistic features in a *compositional vector*. We verify the discriminative power of our *compositional vector* for scene categorization by using it for the classification of images from various, diverse, large scale scene understanding datasets. We then combine the compositional features with traditional semantic features in a complete scene recognition framework. Results show that, due to the complementarity of compositional and semantic features, scene categorization systems indeed benefit from the incorporation of descriptors representing the image photographic layout (+ 13-15% over semantic-only categorization).

1 Introduction

The automatic recognition of visual scenes is a typical, non-trivial computer vision task. The aim is to automatically identify the place where a given image has been captured, or, for example, the type of environment in which a robot is navigating. The general approach is to build a statistical model that can distinguish between pre-defined image classes given a low-dimensional description of the image input, namely a feature vector (here also *signature* or *descriptor*).

Fig. 1. Similar images share similar compositional attributes: depth of field for monuments, point of view for sports field, contrast for natural scenes, level of details and order for indoor scenes.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 446–455, 2012.

One of the main elements influencing the effectiveness of categorization frameworks is indeed the composition of the descriptors used for categorization, because it represents the visual content of the image, i.e. its *semantics*, and semantic analysis is of crucial importance for the identification of the scene category. In scene recognition literature, semantic features are extracted to analyze the image content using either local analysis, based on local interest point descriptors [1] aggregated into a compact image representation [2], or global analysis [3], where general properties of the image, such as color or texture distribution, are summarized into a single descriptor.

Semantic information is without discussion the primary cue for scene identification. However, there exists another important source of information regarding the image scene, namely its *composition*, that could be helpful to recognize the scene category. It has been indeed extensively studied and verified in photography theory [4] that the composition of an image and the content depicted are closely related. We understand here as image composition a combination of aesthetic, affective and artistic components that concur in creating its photographic style, intent [5] and layout. How is this related to scene identification? For example, intuitively it is more likely than an image with a high level of symmetry depicts a non-natural scene (e.g. a building), or that a picture with high level of detail comes from indoor environments. Moreover, as proved in [6], groups of semantically similar images can share the same compositional attributes (e.g. same point of view and depth of field for buildings or sport fields, same color contrast for natural outdoor scenes, see Figure 1).

Given these observations, in this paper we explore the role of compositional attributes for scene recognition using a computational approach. This work represents one of the first attempts of verifying the discriminative ability of compositional features for scene categorization. We design a categorization system that incorporates affective, aesthetic and artistic features, and combines them with traditional semantic descriptors for scene classification. The fusion of such different, discriminative and complementary sources of information about the scene attributes brings a substantial improvement of the scene categorization performances, compared to systems based on semantic features only.

While in literature [7] compositional attributes are generally related to the simple image layout (**aesthetic** attributes, e.g. rules of thirds), here we extend this definition to include **affective** (emotional) and **artistic** attributes that can help characterizing the "intent" [5] of the photographer when composing a given picture. Arranging pictures is not only about applying objective rules, but it is also about following an artistic, intuitive process and convey intentions, meanings and emotions [5]. In order to properly describe the image composition, we therefore extract a set of features from three closely related domains, namely computational aesthetics [8,9] , affective image analysis [10] and artwork analysis [11], and collect them into a single *compositional descriptor*. Many of the features we extract have been proved to be discriminative in their respective domains, but here, we test their discriminative ability for scene classification. In addition to existing features, e.g. low depth of field indicators [8], or color names

[10], we implement two new compositional features: our own version of "image uniqueness", namely a measure evaluating the novelty of the image content, and our own formula to determine image "symmetry". Moreover, we also extract popular semantic features such as the Saliency Moments [12] and the Bag of Words [2]. Then, for both sources of information (compositional+semantic), we use Support Vector Machines to model the feature space and predict the scene category. We then experiment with different fusion methods (early, late) to combine the semantic and compositional information extracted with such features.

We test the effectiveness of our *compositional descriptor* for scene classification using a variety of challenging datasets [13,3,14], including the SUN [14] dataset, that contains around 400 categories of very diverse scenes. We first use our compositional vector as a stand-alone descriptor and we verify that compositional features carry discriminative power for scene categorization. Moreover, we show that, by summarizing the image layout properties into an image descriptor for classification, we introduce a new, complementary source of information regarding the scene characteristics. Therefore, when we combine our descriptor with traditional semantic features in a complete scene categorization system, we increase the classification accuracy of a semantic feature-only system by 13-15% for both small-scale [13,3] and large-scale [14] scene understanding datasets.

The remainder of this paper is organized as follows: in Sec. 2 we outline the state of the art methods related to compositional scene analysis; we then show in Sec. 3 the details of our scene categorization framework embedding compositional and semantic features; finally, we validate our hypothesis with some experimental results in Section 4.

2 Related Work

Compositional features as we understand them have been used in literature for aesthetic, affective or artistic image analysis. Aesthetic image analysis aim at building systems that automatically define the beauty degree of an image: for example, Datta et al. in [8] extract features that model photography rules using a computational approach to predict subjective aesthetic scores for photographs; such model is improved in [15] by adding saliency information in the aesthetic degree prediction framework. In affective image analysis, the aim is to automatically define the type of emotions that a given image arouses: in [16], specific color-based features are designed for affective analysis and in [10], a pool of features arising from psychology and art, and related to the image composition, is proposed to infer the emotions generated by digital images. In art image analysis, specific computational features (e.g. complexity, shape of segments) are designed to investigate patterns in paintings [11] or to assess artwork quality [17].

The interaction between semantic and compositional information has been studied before to improve the modeling of aesthetic/artistic properties of digital images. For example, in [7] semantic concepts are detected in order to enhance the prediction of image aesthetic and interestingness degrees; another approach that combines computational aesthetics with semantic information is proposed

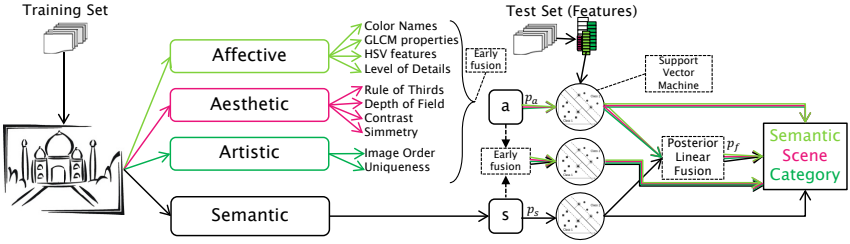

Fig. 2. Combining compositional and semantic attributes for scene recognition

by Obrador et al. [9], that build a set of category-based models for beauty degree prediction; moreover, in [18], painting annotation performances are improved by adding semantic analysis in the artwork understanding framework.

While the relation between semantics and composition has been investigated to improve aesthetic/artistic/emotional analysis, few works have explored the other way around: are compositional features useful for semantic analysis? In this paper, we address this question by combining typical stylistic features with semantic descriptor for scene classification. To our knowledge, the only related work that addresses the same question is the one presented by Van Gemert [6], that generalize the spatial pyramid descriptor aggregator by incorporating photographic style attributes for object recognition. Our work differs from the one in [6] because (1) we focus on a different problem, namely scene categorization rather than object recognition, testing on a variety of challenging databases (2) we test the effectiveness of the actual compositional feature for scene recognition, rather than being inspired from photographic style to modify an existing algorithm.

3 Analyzing Compositional Attributes for Scene Recognition

Scene recognition systems automatically categorize a given image into a predefined set of semantic classes corresponding to different scenery situations. In our approach, we exploit for this purpose the informativeness regarding image composition and photographic style typical of aesthetic, artistic and affective image features. We then combine them with the discriminative traditional semantic features in a complete scene categorization system that predicts an image class based on such diverse sources of information.

Our general framework is basically a traditional image categorization/retrieval framework (see Fig. 2): based on compositional image features, for each category, we learn a model from the training images with Support Vector Machines (SVMs). Similarly, we train a set of SVMs (one for each class) using a set of semantic features. In the test phase, for a new image, given both compositional and semantic features and the models previously computed, we obtain, for each category c, $p_a(c)$ i.e. the category score given compositional features, and $p_s(c)$,

i.e. the category score given semantic features. We retain the prediction from each model to test the discriminative ability of each feature, and we assign the category as $\arg\max_c p_x(c)$, being $x = a, s$. We then combine the prediction scores with weighted linear fusion, namely $p_f(c) = \lambda(p_a(c)) - (1 - \lambda)(p_s(c))$, where λ is a value learnt during training. The final image category is assigned according to the resulting category scores after fusion.

The peculiarity of our system is the choice of particular, discriminative image features that go beyond the traditional semantic descriptors for scene categorization by evaluating not only the content but also the compositional style of the image. In the remainder of this Section, we therefore focus on the analysis of the compositional features we extract from the image, together with some insights about the type of semantic analysis we perform to complete the scene recognition task.

3.1 Compositional Features: Aesthetic, Affective and Artistic Features

Previous works in computational image composition [9,7] understands it as a set of objective rules for constructing the image layout. For example, compositional attributes have been defined for aesthetic scene analysis as *"characteristics related to the layout of an image that indicate how closely the image follows photographic rules of composition"* [7]. Here, we extend this concept to include features describing image emotional and artistic traits. As Freeman states in [5] *"So far we have been concerned with the vocabulary of grammar and composition, but the process usually begins with purpose - a general or specific idea of what kind of image a photographer wants"*. In order to model the photographer's "intent" as defined by Freeman, we summarize the image composition using affective attributes, that describe the emotions that a given image arouses through affective measures, and artistic attributes, that determine, for example, the "uniqueness" of a given image.

In order to effectively describe the image photographic and artistic composition, we therefore design a compositional descriptor of 43 features coming from emotion-based image recognition, computational aesthetics, and painting analysis. For each image/frame, we extract our compositional 43-d feature vector $a = \{a(i)\}_{i=1}^{43}$, as follows:

Color names, a(1-9). Similar to [10] we count the amount of 9 different common colors in the image: different color combinations are used from artists/photographers to arouse different emotions.
GLCM properties, a(10-19). Gray-level co-occurrence matrices [19] are efficient ways to infer the image texture properties, which are of crucial importance to determine the affective content of a given image. Here, similar to [10], we fill our feature vector with the properties of correlation, homogeneity, energy, entropy and dissimilarity inferred from the GLCM matrix of a given image.
HSV features, a(20-25). After transforming the image into HSV space, we take the mean of hue, saturation and brightness, and compute *pleasure, arousal* and *dominance* features according to [10].

Level of detail, a(26). We measure image homogeneity from [10] based on the number of segments resulting after waterfall segmentation.
Rule of thirds, a(27-29). We evaluate how much the image follows the photography rule of thirds by taking the mean of Hue, Saturation and Brightness of the image inner rectangle, as in [8].
Low depth of field, a(30-38). The depth of field measures the ranges of distances from the observer that appear acceptably sharp in a scene. We extract low DoF indicators using wavelet coefficients as described in [8].
Contrast, a(39). As in [20], we extract the contrast Michelson measure [21].
Image Order, a(40,41). According to Birkhoff [22], image beauty can be found in the ratio between order and complexity. Following this theory, image (in particular, arts and painting) order is computed in [11] using an information theory approach. We compute here the image order using Shannon Entropy and Kologomorov Complexity approaches proposed in [11].
Symmetry, a(42). Image Symmetry is a very important element to define the image layout. We define our own symmetry feature: We extract the Edge Histogram Descriptor [23] on both the left half and the right half of the image (but inverting major and minor diagonals in the right half), and retain the difference between the resulting histograms as the amount of symmetry in the image.
Uniqueness, a(43). How much an image represents a novelty compared to known information, how much is an image unique, i.e. it differs from the common image behavior? this variable can tell much about the artistic content of an image. We propose a new solution to address this question. We define the common image behavior according to the "1/f law" [24], saying that the average amplitude spectrum of a set of images obeys a 1/f distribution. We measure the uniqueness by computing the Euclidean distance between the average spectrum of the images in the database and the spectrum of each image.

We finally normalize all the features in the range [0,1] and combine them into our compositional vector.

3.2 Semantic Features

The core of the discriminative power of our scene recognition system is still the set of semantic features for categorization. Here, we select to compute a powerful global feature for scene recognition, namely the Saliency Moments (SM) descriptor. The SM has been proved in [12] to outperform existing features for image categorization and retrieval and it was effectively used in various Trecvid runs (e.g. [25]) due to its complementarity with the state of the art image descriptors. The SM descriptor exploits the informativeness of the saliency distribution in a given image and computes a fast, low dimensional gist of the image through visual attention information summary. First, the spectral saliency map [26] is extracted. Such spectral signal is then sampled using Gabor Filters: the resulting Saliency Components are then decomposed into smaller regions, then mean and higher order statistics are calculated for each region and stored in the final 462-d feature vector $s = \{s(l)\}_{l=1}^{462}$.

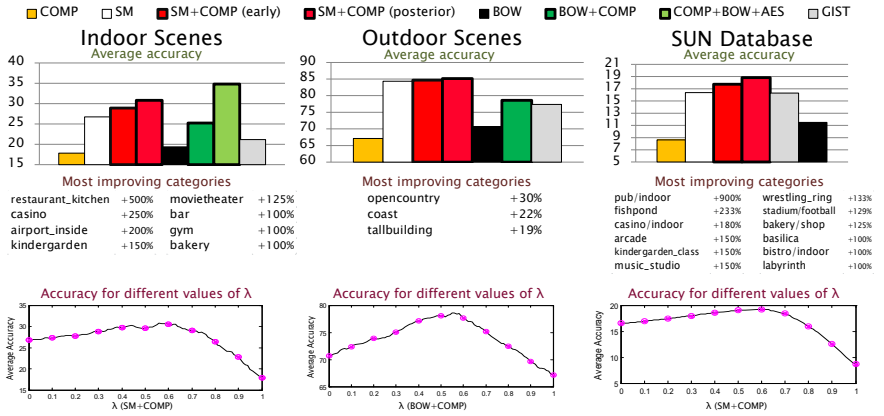

Fig. 3. Results of large scale and small scale scene recognition

Moreover, for indoor and outdoor scene recognition, we extract also a sematic feature based on local image descriptors aggregation, namely the Bag-of-Words (BOW) feature. The BOW model [2] is one of the most used approaches for semantic indexing and image retrieval. In this approach, local descriptors as SIFT [1] are computed to describe the surroundings of salient [27] or densely sampled [28] points. Each image is then mapped into a fixed length signature through a visual codebook computed by clustering the local descriptors in the training set. We chose this feature for its high discriminative ability and its complementarity to global features such as SM and our compositional feature.

4 Experimental Results

In order to test the effectiveness of the proposed approach, and verify the usefulness of aesthetic and affective features for semantic indexing, we use our framework for two scene recognition tasks: small scale categorization and large scale categorization. For the first task, we use two very popular benchmarking datasets for indoor [13] and outdoor [3] scene recognition, while for large scale scene recognition, we test our system on the challenging SUN database [14].

For each database, we first compute the classification accuracy given the model built using each semantic feature (i.e. "BOW" or "SM" in Fig. 3). We then look at the classification performances resulting from using our compositional feature ("COMP") as a stand-alone descriptor. Furthermore, we show the effectiveness of the combination of aesthetic and compositional features by first fusing semantic and aesthetic features in a single, early fused descriptor (e.g. "SM+COMP (early)"). Finally, we combine the predictions of the single-descriptor-based models with posterior linear fusion. We fix the parameter λ for fusion and show the resulting, improved, performances (e.g. "SM+COMP (posterior)" in Fig. 3). For all descriptors and datasets proposed, we learn the feature space through a multiclass SVM with Radial Basis Function Kernel and we evaluate the performances by average multiclass accuracy.

4.1 Small Scale Scene Recognition

Automatic classification of images into scene categories is performed here using the proposed framework over two small scale dataset for indoor and outdoor scene recognition.

Outdoor Scenes
The Outdoor Scenes Dataset was first introduced in [3] to evaluate the performances of a very popular descriptor for scene categorization, namely the Gist descriptor. It is composed of 2600 color images spanning 8 categories of natural outdoor scenes. In order to perform our experiments, we split the outdoor scene dataset into 100 images per class for training and the rest for testing, as proposed in [3]. For this dataset, we compute both the SM and the BOW descriptors, and combine them with the compositional descriptor proposed in this work.
Results show that, by combining aesthetic, affective and artistic features in our compositional descriptor ("COMP") we obtain an effective descriptor (68% of accuracy VS 12.5% of a random classifiers) for outdoor scene recognition. Moreover, we can see that, while its combination with the SM descriptor does not bring much improvement[1], its fusion with the BOW features increases the performances of the BOW-only classification by 11%.

Indoor Scenes
The Indoor Scenes Dataset,was proposed in [13] as a new, unique database for indoor scene recognition, collecting around 15000 images from various sources, and considering 67 different image categories related to indoor environments. For our experiments, we split this datasets as proposed in [13]: for each class, we retain 20 images for testing and the rest for training. Again, for this small-scale database we compute both SM and BOW and we combine it with the aesthetic/artistic/affective feature vector.

Results in this task clearly highlight the effectiveness of compositional features for scene recognition: while the accuracy of the compositional descriptor alone is not as good as semantic features (around 17% vs. 26% of SM), but still more than 10 times better than a random classifier ($\sim$ 1,4%), the scenario changes when we combine it with traditional semantic features. As a matter of fact, both the early (+ 8%) and the posterior (+ 15%) fusion with the Saliency Moment descriptor successfully enhance the final scene recognition performances. Similar, more evident behavior when we combine the compositional features with the BOW descriptor: such fusion brings an improvement of 30 % compared to BOW-only classification. Being BOW and SM complementary, and being both complementary to compositional features, we also tried to combine the predictions resulting from the three stand-alone models using posterior linear fusion. The improvement over the classification based on SM (i.e. the most performing stand-alone descriptor) in this case is more than 20%, suggesting that introducing compositional features in the pool of existing semantic features is a promising cue for indoor scene recognition.

[1] This is because SM is an extremely effective descriptor by itself for outdoor scenes, and because it contains already some compositional information related to saliency

Large Scale Scene Recognition
Finally, we present our results for large scale scene recognition over the challenging SUN database, proposed in [14] as a complete dataset for scene understanding, with a variety of indoor and outdoor scene environments, spanning 899 categories for more than 130,000 images. As in [14], for benchmarking purposes, we select a pool of 397 scenes out of the categories proposed, and we use a subset of the SUN dataset consisting 10 folds that contains, for each category, 50 images for test and 50 for training. Results are obtained by averaging the performances of the descriptors over the 10 partitions considered. In order to test the effectiveness of our approach, we compute here the SM descriptor and combine it with the compositional feature we propose.
Results on this dataset follow the same pattern of the previously analyzed experiments: the combination of the SM with aesthetic/affective features brings an improvement of 8% with early fusion and 13% with late fusion compared to the SM-only classification, thus confirming the discriminative ability and the complementarity of aesthetic and compositional features for scene recognition even on a large scale.

5 Conclusions

This work represents a first attempt of combining semantic, artistic, affective, and emotional image analysis in a unique framework for scene recognition. We showed with our results that categorization systems benefit from the incorporation of compositional features. The current system can be improved by experimenting with different types of fusion or by designing a set of category-specific compositional vectors, which can be constructed based on the discriminative ability of each feature of each class.

References

1. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
2. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, vol. 1, p. 22. Citeseer (2004)
3. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42 (2001)
4. Krages, B.: Photography: the art of composition. Allworth Pr. (2005)
5. Freeman, M.: The photographer's eye: composition and design for better digital photos. Focal Pr. (2007)
6. van Gemert, J.: Exploiting photographic style for category-level image classification by generalizing the spatial pyramid. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, p. 14. ACM (2011)
7. Dhar, S., Ordonez, V., Berg, T.: High level describable attributes for predicting aesthetics and interestingness. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1657–1664. IEEE (2011)

8. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying Aesthetics in Photographic Images Using a Computational Approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 288–301. Springer, Heidelberg (2006)
9. Obrador, P., Saad, M.A., Suryanarayan, P., Oliver, N.: Towards Category-Based Aesthetic Models of Photographs. In: Schoeffmann, K., Merialdo, B., Hauptmann, A.G., Ngo, C.-W., Andreopoulos, Y., Breiteneder, C. (eds.) MMM 2012. LNCS, vol. 7131, pp. 63–76. Springer, Heidelberg (2012)
10. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: Proceedings of the International Conference on Multimedia, pp. 83–92. ACM (2010)
11. Rigau, J., Feixas, M., Sbert, M.: Conceptualizing birkhoff's aesthetic measure using shannon entropy and kolmogorov complexity. In: Computational Aesthetics in Graphics, Visualization, and Imaging (2007)
12. Redi, M., Merialdo, B.: Saliency moments for image categorization. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR 2011 (2011)
13. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
14. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3485–3492. IEEE (2010)
15. Wong, L., Low, K.: Saliency-enhanced image aesthetics class prediction. In: 2009 16th IEEE International Conference on Image Processing (ICIP). IEEE (2009)
16. Wang, W., Yu, Y.: Image emotional semantic query based on color semantic description. In: Proceedings of 2005 International Conference on Machine Learning and Cybernetics, vol. 7, pp. 4571–4576. IEEE (2005)
17. Li, C., Chen, T.: Aesthetic visual quality assessment of paintings. IEEE Journal of Selected Topics in Signal Processing 3, 236–252 (2009)
18. Leslie, L., Chua, T., Ramesh, J.: Annotation of paintings with high-level semantic concepts using transductive inference and ontology-based concept disambiguation. In: Proceedings of the 15th International Conference on Multimedia. ACM (2007)
19. Haralick, R.M., Shapiro, L.G.: Computer and Robot Vision, 1st edn. Addison-Wesley Longman Publishing Co., Inc., Boston (1992)
20. Desnoyer, M., Wettergreen, D.: Aesthetic image classification for autonomous agents. In: Proc. ICPR. Citeseer (2010)
21. Michelson, A.: Studies in optics. Dover Pubns. (1995)
22. Birkhoff, G.: Aesthetic measure (1933)
23. Won, C., Park, D., Park, S.: Efficient use of mpeg-7 edge histogram descriptor. Etri Journal 24, 23–30 (2002)
24. Ruderman, D.: The statistics of natural images. Network: Computation in Neural Systems 5, 517–548 (1994)
25. Delezoide, B., Precioso, F., Redi, M., Merialdo, B., Granjon, L., Pellerin, D., Rombaut, M., Jégou, H., Vieux, R., Mansencal, B., et al.: Irim at trecvid 2011: Semantic indexing and instance search. TREC Online Proceedings (2011)
26. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007. IEEE (2007)
27. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2. IEEE (2003)
28. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 2, pp. 524–531. IEEE (2005)