Image understanding - a brief review of scene classification and recognition

Vineeta Singh, Deeptha Girish, and Anca Ralescu
EECS Department, ML 0030
University of Cincinnati
Cincinnati, OH 45221, USA
singhvi@mail.uc.edu, girishde@mail.uc.edu, Anca.Ralescu@uc.edu

Abstract

With over 40 years of history, image understanding, in particular, scene classification and recognition remains central to machine vision. With an abundance of image and video databases, it is necessary to be able to sort and retrieve the images and videos in a way that is both efficient and effective. This is possible only if the categories of images and/or their context are known to a user. Hence, the ability to classify and recognize scenes accurately is of utmost importance. This paper presents a brief survey of the advances in scene recognition and classification algorithms.

Depending on its goal, image understanding(IU) can be defined in many different ways. However, in general, IU means describing the image content, the objects in it, location and relations between objects, and most recently, describing the events in an image. In (Ralescu 1995) IU is equated with producing a verbal description of the image content. Scene analysis (as part of IU) and categorization is a highly useful ability of humans, who are able to categorize complex natural scenes containing animals or vehicles very quickly (Thorpe, Fize, and Marlot 1996), with little or no attention (Li et al. 2003). When a scene is presented to humans, they are able to quickly identify the scene, i.e., within a short period of exposure ($<$ 100 ms). How do humans perform all of these tasks the way they do, is yet to be fully understood. To date, the classic text by Marr (Marr 1982) remains one of the sources of understanding the human vision systems.

Many researchers have tried to imbibe this incredible capability of the human vision system into their algorithms for image processing, scene understanding and recognition. In the presence of a wealth of literature on this and related subjects, surveys of the field, even a limited one, as the present one necessarily is (due to space constraints) are bound to be very useful, by reviewing the methods for scene recognition and classification.

Perhaps, the first issue to consider is the concept of *scene* as a *technical* concept to capture the *natural* concept. According to Xiao et al. (Xiao et al. 2010) a *scene* is a place in which a human can act within, or a place to which a human being could navigate. Therefore, *scene recognition* and *scene classification* algorithms must delve into understanding the *semantic context* of the scene. According to how a scene is recognized in an image, scene recognition algorithms can be broadly divided into two categories.

- Scene recognition based on object detection.
- Scene recognition using low-level image features

Scene recognition using object recognition (SR-OR)

Using object recognition for scene classification is a straight-forward and intuitive approach to scene classification and it can assist in distinguishing very complex scenes which might otherwise prove difficult to do using standard low level features.

In the paper by Li-Jia Li et al. (Li et al. 2010) the authors argue that although "robust low-level image features have been proven to be effective representations for scene classification; but pixels, or even local image patches, carry little semantic meanings. For high level visual tasks, such low-level image representations are potentially not enough. " To combat this drawback of local features, they propose a high-level image representation, called the *Object Bank*(OB), where an image is represented by integrating the response of the image to various object detectors. These object detectors or filters are blind to the testing dataset or visual task. Using OB representation, superior performances on high level visual recognition tasks can be achieved with simple regularized logistic regression. Their algorithm uses the current state-of-the-art object detectors of Felzenszwalb et al. (Felzenszwalb et al. 2010), as well as the geometric context classifiers (stuff detectors) of Hoeim et al. (Hoiem, Efros, and Hebert 2005) for pre-training the object detectors.

OB offers a rich set of object features, while presenting a challenge – curse of dimensionality due to the presence of multiple class of objects within a single image, which then yields feature vectors of very high dimension. The performance of the system plateaus at a point when the number of object detection filters is too high. According to the authors, the system performance is best, when the number of object filters is moderate.

Scene recognition using low-level image features (SR-LLF)

Many of the papers in scene recognition are built around the question, *'Can we recognize the context of a scene without having first recognized the objects that are present?* There are a lot of reasons for avoiding object recognition for the purpose of scene recognition. While there are many robust OR algorithms, using SR-OR can be problematic because the OR portion of the algorithm is treated as a black box, and therefore, the OR errors propagate to the SR segment. OR also faces problems due to lighting conditions and occlusion. To avoid this many studies tend to use low level feature for scene understanding.

The challenge in SR-LLF is to find low-level features in the image that can be used successfully to infer its semantic context. Among the many features can be extracted from the image for the purpose of scene recognition, texture, orientation, and color have been used extensively in literature, implemented with different data sets and with different classifiers.

In (Renninger and Malik 2004) an algorithm which mimics the humans' ability to identify scenes with limited exposure is presented. The algorithm is based on a simple texture analysis of the image which can provide a useful cue to rapid scene identification. The relevant features within a texture are the first order statistics of *textons* which determine strength of texture discrimination. This idea is derived from Juselz's work (Julesz 1981) (Julesz 1986) (For a discussion on Juselz's work see (Marr 1977)). According to Juselz, textons are the elements in the image that govern our perception of texture. They are calculated by convolving the image with certain filter banks. The textons based model learns the local texture features which correspond to various scene categories, which is done by filtering a set of 250 training images and then learning the prototypical distributions. The number of occurrences of each feature within a particular image is stored as an histogram, creating a holistic texture descriptor for the image. To learn the most prevalent features, they use k-means clustering to find the 100 prototypical responses. When identifying a new test image, its histogram is matched against stored examples. It is concluded that early scene identification can be explained with a simple texture recognition model. This model leads to similar identifications and confusions as a human subject.

The same objective, i.e., understanding human perception of scenes is pursued in (Gorkani and Picard 1994). The paper investigates the measure of *dominant perceived orientation* developed to match the output of a human study involving 40 subjects. These global multi-scale orientation features were used to detect vacation photos belonging to "city/suburb". The authors state that orientation is an important feature for texture recognition and discrimination. The algorithm finds the local orientation and its strength at each pixel of the image. The implementation extracts orientation information over multiple scales using a steerable pyramid (Rock 1990) and then combines the orientations from these different scales and decides which one is dominant perceptually. The reported results show that the orientation features achieves agreement with the human classification in 91 of the 98 scenes (i.e., approximately 92.93%)

The paper by Guérin-Dugué et al. (Guérin-Dugué and Oliva 2000) uses an approach similar to Gorkani and Picard (Gorkani and Picard 1994) but extend this approach with more categories and introduce the selection of the optimal scale for this categorization task. They use local dominant orientation (LDO) features for classifying real-world scenes into four categories (outdoor urban scenes, indoor, closed landscapes and open landscapes). Instead of using the LDO features directly, they propose compact coding in a few features by Fourier series decomposition, and introduce the spatial scale parameter to optimize the categorization. For each scale and spatial location the dominant orientation and it's strength are estimated. The best discrimination ratios were obtained with a representation at a median spatial scale or when combining two different scales.

The paper (Csurka et al. 2004) presents a bag of keypoints approach to visual categorization. The procedure starts by detection and description of image patches. Then a vocabulary of image descriptors is created after applying the vector quantification algorithm. SIFT decriptors (Lowe 1999) are used as features for this algorithm. This is followed by constructing a bag of key-points which counts the number of patches assigned to each cluster. Finally, a multiclass classifier (SVM) is implemented, treating the bag of points as the feature vector and thus, determining which category the image belongs to.

It is clear that counts, or histograms, suggest that scene recognition and analysis could benefit from probabilistic approaches. Indeed, some algorithms use probability models to describe the scene based on the extracted features.

The paper *A Bayesian Hierarchical Model for Learning Natural Scene Categories* (Fei-Fei and Perona 2005) uses low level texture features as image descriptors. Each patch of the input image is represented using a code word (similar to bag of keywords approach). The code word is taken from a *codebook* – a large vocabulary of code words – obtained from 650 training examples from 13 categories (with around 50 images for each category). In this framework, initially the local regions are clustered into different intermediate themes and then into categories. The learning algorithm for achieving the model that best represents the distribution of code words to represent scenes is a modified Latent Dirichlet model(Blei, Ng, and Jordan 2003). Unlike traditional scene models where there is a hard assignment of an image to one theme, the algorithm produces a collection of themes that could be associated with an image.

In (Singhal, Luo, and Zhu 2003) a probabilistic approach is used for content detection within the scene. The labels generated are very similar to scene labels. The authors present a holistic approach to determine the scene content, based on a set of individual *material detection* algorithms as well as probabilistic spatial context models. Material detection is the problem of identifying key semantic objects such as sky, grass, foliage, water, and snow in images. In order to detect materials the algorithm combines low-level features with unique region analysis and inputs this to a classifier to obtain individual material belief maps. To avoid mis-

classification of materials in images they devise a spatial context aware material detection system which constraints the beliefs to conform to the probabilistic spatial context models.

The bag of keypoints model (Sivic and Zisserman 2009) corresponds to a histogram of the number of occurrences of particular image patterns in a given image. Most papers mentioned above use this concept in some form. This is adapted from the bag of words model in natural language processing.

In (Lazebnik, Schmid, and Ponce 2006) the authors argue that in spite of impressive levels of performance, the bag of features model represents the image as an *orderless* collection of local features, thereby disregarding all information about the spatial layout of the features. To overcome this aspect, they devise a method for recognizing scene categories based on the approximate global geometry correspondence. They compute a *spatial pyramid* by partitioning the image into increasingly fine sub-regions and computing histograms of local features found in each sub-region. The *spatial pyramid* is an extension of the orderless bag of features model of image representation, which is improved upon by the introduction of a kernel based recognition method. This method works by computing a rough geometric correspondence on a global scale using an approximation technique adapted from the pyramid matching scheme of (Grauman and Darrell 2007). This method involves repeatedly subdividing the image and computing histograms of local features at increasingly fine resolutions. The spatial pyramid approach can be thought of as an alternative formulation of locally orderless image where a fixed hierarchy of rectangular windows is defined. The *spatial pyramid framework* is based on the idea that the best results will be achieved when multiple resolutions are combined in a principled way. The features calculated are subdivided into *weak features*, oriented edge points, and *strong features*, SIFT descriptors. K-means clustering is performed on a random subset of patches from the training set to form a visual vocabulary. Multiclass classification is done with the support vector machine (SVM) , trained using the one versus all rule.

Though fewer algorithms use color based features, in certain cases this descriptor is very powerful in discriminating scenes.

Color descriptors can be used for scene and object recognition (Van De Sande, Gevers, and Snoek 2010) in order to increase illumination invariance and discriminative power. From theoretical and experimental results it is shown that, invariance to light intensity changes and light color changes affect category recognition. Various color descriptors were analyzed and evaluated. Color descriptors based on histograms, color moments moment invariants and color SIFT were used as descriptors, and it was concluded that SIFT based descriptors performed considerably better that histogram and moment based descriptors.

Indoor-Outdoor classification

In the paper by Szummer et al. (Szummer and Picard 1998) the authors show that high-level scene properties can be inferred from classification of low-level features specifically for indoor - outdoor scene retrieval problem. Their algorithm extracts three types of features: $1)$ histogram in the ohta color space $2)$ multi-resolution simultaneous autoregressive model parameters $3)$ coefficients of shift invariant DCT. They exhibit that performance is improved by computing features on sub-blocks, classifying these sub-blocks and then combining these results by stacking.

This paper (Serrano, Savakis, and Luo 2004) by Serrano et al. uses simplified low level features to predict the semantic category of scenes. This is integrated probabilistically using Bayesian network to give a final indoor/outdoor classification. Low-dimensional color and wavelet texture features are used to classify scenes using the support vector machine (SVM). These wavelet texture features are used here instead of the popular MSAR texture features to reduce the computational complexity.

Other approaches

Various other approaches exist in the literature of scene recognition, as reviewed below.

Semantic Typicality

The concepts of typicality and prototype have made a significant impact in cognitive science. See for example the work pioneered by Eleanor Rosch and her collaborators, (Rosch 1973), (Rosch and Mervis 1975), (Rosch et al. 1976). In computer vision, (Vogel and Schiele 2004) introduces an interesting concept of *semantic typicality* in categorizing of real world natural scenes. The proposed typicality measure is used to grade the similarity of an image with respect to a scene category. Typicality is defined as a measure for the uncertainty of annotation judgment. This is an important concept because many natural scenes are ambiguous and the categorization accuracy sometimes reflects the opinion of a particular person who performed the annotation. Therefore, the authors believe that attention should be directed to modeling the typicality of a particular scene after manual annotation. The semantic typicality measure is used to find the similarity of natural real-world scenes with respect to six scenes including coasts, rivers/lakes, forests, plains, mountains and sky/clouds.

The typicality based approach is evaluated on an image database of 700 natural scenes. The attribute score is a representation which is predictive of typicality. Typicality is a function of frequency of occurrence, that is, the items deemed most typical have attributes that are very common to the category. Local semantic concepts act as scene category attributes. They are calculated from the sub-regions which are represented by a combined 84-bin linear histogram in the HSI color space, and a 72-bin edge direction histogram. Classification is done by a k-nearest neighbor classifier. The categorization experiment was carried out using manually annotated images from the database. By analyzing the semantic similarities and dissimilarities of the aforementioned categories a set of nine local semantic concepts emerged as being most discriminant: sky, water, grass, trunks, foliage, fields, rocks, flowers, and sand. The local semantic concepts were extracted on a 10×10 grid of image sub-regions and

the frequency of occurrence in a particular image was represented by concept occurrence vector. For each category, a *category prototype* is defined as the most typical example for that category, which constructed as the means over the concept occurrence vectors of the category members. The image typicality was measured by computing the Mahalanobis distance between the images' concept occurrence vector and the prototypical representation in order to classify the image as a particular scene.

Configural Recognition

The goal of (Lipson, Grimson, and Sinha 1997) is to classify scenes based on their content. Most of the solutions that are available for scene recognition rely on color histograms and local texture statistics. The authors state that these features cannot capture a scenes' global configuration. To overcome this they present a novel approach, which they call *configural recognition* for encoding scene class structure in images. The configural recognition scheme encodes class models as a set of salient low resolution image regions and salient qualitative relations between the regions. An example of qualitative relationships are: 'given three regions, a blue region(A), a white region (B) and a gray region (C), Snow-capped mountains always have region A above region B which is above region C'.

The class models are described using seven types of relative relationships between image patches. Each of them has the following values: *Less than, greater than*, or *equal to*. The relationships encoded are relative color between image regions, the relative luminance between the patches, the spatial relationships (relative horizontal and vertical descriptions) and the relative size of the patch. Based on this, each region in the image is grouped into directional equivalence classes, such as *above* and *below*.

The generated model acts as deformable templates. When compared with the image, the model can be deformed by moving the patches around so that the model best matches the image in terms of relative luminance and photometric attributes. An improvement to this system can be made where instead of hand crafting the models,an automated process could take a set of example images and generate a set of templates which describe the relevant relationships between the pictures in the example set.

A fuzzy part based model was described in (Miyajima and Ralescu 1993) and fuzzy sets were also widely and effectively used for spatial descriptors in an image. A very powerful formal model, based on fuzzy sets, for the description of spatial relations in an image was introduced in (Miyajima and Ralescu 1994), (Miyajima and Ralescu 1994) and further extended by (Bloch 1999). A comparison of the fuzzy approaches for the description of directional relative position between object in an image can be found in (Bloch and Ralescu 2003), and a review of these approached can be found in (Bloch 2005). Furthermore, more recently, fuzzy spatial relations were integrated in deformable models and applied to MRI images (Colliot, Camara, and Bloch 2006).

In (Ralescu and Baldwin 1987) a new approach for concept learning from examples and counter-examples with applications to a vision learning system, later extended to a general concept learning problem (Ralescu and Baldwin 1989), was developed. It makes use of Conceptual Structures (Sowa 1983) for knowledge representation, and support logic programming (Baldwin 1986) for inference. Examples of a concept (e.g., a 'car') are used to construct a *memory aggregate*(MA), which rather than average all examples, keeps track of various probability distributions of the object features. Counter-examples, i.e., descriptions that are very similar to a concept, but fail to be instances of that concept, are used in a similar manner to construct a counter example memory aggregate (CMA). Matching between conceptual structures describing an object candidate and the MA and CMA produce supports for and against the recognition of a concept. The result is therefore, qualified by a *support pair*, whose values $(1, 1)$ mean complete recognition, $(0, 0)$ complete rejection, $(0, 1)$ total uncertainty.

Deformable part based models

In (Pandey and Lazebnik 2011), the author comments that weakly supervised discovery of common visual structure in highly variable and that cluttered images present a major problem in recognition. In order to address this problem, the authors propose using *deformable part-based models* (DPM) with latent SVM training. For scene recognition, deformable part-based models capture recurring visual elements and salient objects. The DPM represents an object by low-resolution root filters and a set of high resolution part filters in a flexible spatial configuration. The image is represented by a variation of histogram of oriented gradients (HOG) features which are used to classify scenes using linear SVM.

Covariance descriptor

The paper (Yang et al. 2016) proposes a supervised collaborative kernel coding method based on *covariance descriptor* ($covd$) for scene level geographic image classification. Covariance descriptor is a covariance matrix of different features such as color, spatial location, and gradient that is rotation and scale invariant but it lies in the Riemannian space (i.e., non- Euclidean space) and therefore, the traditional computational and mathematical models used in the Euclidean space cannot be used.

The major contribution of this paper is that they propose a supervised kernel coding model that transforms $covd$ into a discriminative feature representation and obtain a corresponding linear classifier. The method can be seen as a three step process. The first step is to extract the $covd$ features from the geographical scene image. In the second step supervised collaborative kernel coding involving dictionary coefficients in the coding representation phase and linear classification phase is performed. Lastly, in the classification stage, based on dictionary coefficients and learned linear classifier a label vector is derived. A novel objective function is proposed to combine the collaborative kernel coding phase and the classification phase. This method gives satisfying performance on high resolution aerial image dataset proving to be an efficient method for scene level geographic image classification.

Shape of the scene

The paper (Oliva and Torralba 2001) takes a very different approach to scene recognition: rather than looking at the scene as a configuration of objects the paper proposes to consider the scene as an individual object, with a unitary shape. A computational model to find the shape of the scene using a few perceptual dimensions specifically dedicated to describing spatial properties of the scene is proposed. It is shown that the holistic spatial scene properties, called *Spatial Envelope*(SE) properties may be reliably estimated using spectral and coarsely localized information.

Given and environment V, its spatial envelope $SE(V)$ is defined as a composite set of boundaries such as walls, section, elevation etc. that define the shape of the space. A group of 17 observers were asked to categorize 81 images into categories based on some global aspect. Based on the classification results, the criterion for classification of scenes was agreed to be based upon the *degree of naturalness*, *degree of openess*, *degree of roughness*, *degree of expansion* and *degree of ruggedness*. Therefore, the purpose of the spatial envelope model is to show that modeling these five spatial properties is adequate to understand the high-level description of the scene. Their algorithm learns the spectral signatures (the global energy spectrum and the spectogram) of basic scene categories from labeled training data. A learning algorithm (regression) is then used to find the relation between the global features and spectral features.

Beyond Scene recognition

Certain algorithms detect scenes and then use scene recognition as a prior in order to find more structure in the image, thus motivating further study in the field of scene analysis.

In *Using Forest to see trees* (Murphy et al. 2003) an intuitive approach to detect the presence of object based on the detected scenes is presented. The approach is suggested by psychological evidence that people perform rapid global scene analysis before and conducting more detailed local object analysis. Based on this the authors propose to use the whole image as a global feature in order to overcome ambiguities which might occur at the local level. They extend the notion of *gist* from (Oliva and Torralba 2006) by combining the prior suggested by the gist to the output of bottom-up local object detectors which are trained using boosting. They also use the same set of features for object detection in the image. The image is divided into patches at different scales (image pyramid) and each patch is convolved with 13 zero-mean filters which include oriented edges, a Laplacian filter, corner detectors and long edge detectors. This is represented by two statistics, variance, and kurtosis derived from the histogram of image patches at two scales and with 30 spatial masks. The kurtosis is omitted for scene recognition. The features are further reduced in dimensionality using PCA to give the PCA-*gist*. A one vs all binary classifier is trained for recognizing each type of scene using boosting applied to the gist. They further this by using *scene* as a latent common cause upon which the presence of the object is conditionally dependent on.

Understanding whole image, or *holistic* scene understanding is is described in (Yao, Fidler, and Urtasun 2012). The idea is to jointly evaluate and makes conclusions about location, regions, class and spatial information of objects, presence of a class in an image and also the scene type. The idea is to recover and connect the multiple different aspects of a scene. This problem is framed as a prediction problem in a graphical model defined over hierarchies of regions of different sizes, auxiliary variables encoding scene type, presence of a given class in a scene and correctness of bounding boxes obtained by the object detector. Class labels of image segments at two different levels of segmentation hierarchy, namely *segments* and large *super segments* is proposed. Binary variables indicate which classes are present in images and multi-labeled variable represents scene type. Segments and super segments are used to assign semantic class labels to each pixel in an image. Super segments are used to create long range dependencies and they also prove to be more efficient computationally. The *holistic loss function* is defined, which is a weighted sum of losses from each task. State of the art performance is achieved in MSRC-21 benchmark and the approach is much faster than the existing approaches.

Another very interesting approach is present in (Li and Fei-Fei 2007), which goes beyond scene recognition to event recognition. An *event in a static image is defined as a human activity taking place in the specific environment*. The objective is to recognize/classify the event in the image as well as provide a number of semantic labels to the object and scene environment within the image. It is assumed that conditioned on the event, scene and objects are independent of each other, but both their presence influences the probability of predicting the event. For scene recognition they adopt a model similar to the Bayesian model of (Fei-Fei and Perona 2005). Scene recognition heavily influences the event recognition, an in fact, as a first approximation event recognition is essentially scene recognition. The robust bag of words model is used in order to recognize objects. In addition to scene and object recognition they understand the importance of layout of the image in accurately identifying the event. They use some simple geometric cues to define the layout of the image and manage to provide integrative and hierarchical labels to an image by performing the what (event), where (scene) and who (object) recognition of the entire scene by using a generative model in order to represent the image.

An extensive Scene Understanding (SUN) database consisting of 899 categories and 130519 images is created in (Xiao et al. 2010). This work is motivated by the authors' belief that the existing data sets for scene classification fail to capture the richness and diversity of daily life environments. The authors claim to have built the most complete dataset with a number of different scene image categories with different functionalities that are important enough to have unique identities. They measure human performance on scene classification and compare it with the state of the art algorithms, using the SUN database. Both human and algorithm results had errors, with the humans erring between semantically similar categories, while algorithms erring between semantically unrelated scenes due to spurious visual matches. It was also recorded that the best features agree more with correct human classifications and make the same mistakes as humans do. The computational algorithms need

a much larger number of features to performs as well as humans. They also propose the notion of recognizing scene type within images rather than labeling an entire image with a scene because the real world often contains combination of scenes. This is a new interesting idea and could also be one of the directions in which the future scene recognition algorithms can progress.

Conclusion

It can be seen, even based on the limited number of papers reviewed here, that image understanding, scene recognition can be approached from various different directions. At a very high level, the approaches can be divided into two main categories - using low-level features, and using object recognition. However, many other techniques are integrated into each of these approaches, including probabilistic, and/or fuzzy techniques, in order to deal with the uncertainty which often attends the result of image understanding. When it come to evaluating low level feature approach and object recognition approach, the goal of the image understanding must be taken into account. Scene recognition performs better when low level features are used. Local features help override the effects of occluded objects, low lighting conditions. Most commonly used features for scene detection include texture, texture orientation and strength, 'Gist' of the image, SIFT descriptor, edge orientation, histograms in different color space (e.g., Ohta, HSI, RGB), histograms of angles between segmented regions, coefficients of shift-invariant DCT. These features can be successfully mapped into semantic image descriptors.

References

Baldwin, J. F. 1986. Support logic programming. In *Fuzzy sets theory and applications*. Springer. 133–170.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Bloch, I., and Ralescu, A. 2003. Directional relative position between objects in image processing: a comparison between fuzzy approaches. *pattern Recognition* 36(7):1563–1582.

Bloch, I. 1999. Fuzzy relative position between objects in image processing: a morphological approach. *IEEE transactions on pattern analysis and machine intelligence* 21(7):657–664.

Bloch, I. 2005. Fuzzy spatial relationships for image processing and interpretation: a review. *Image and Vision Computing* 23(2):89–110.

Colliot, O.; Camara, O.; and Bloch, I. 2006. Integration of fuzzy spatial relations in deformable modelsapplication to brain mri segmentation. *Pattern recognition* 39(8):1401–1414.

Csurka, G.; Dance, C.; Fan, L.; Willamowski, J.; and Bray, C. 2004. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, 1–2. Prague.

Fei-Fei, L., and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, 524–531. IEEE.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2010. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence* 32(9):1627–1645.

Gorkani, M. M., and Picard, R. W. 1994. Texture orientation for sorting photos" at a glance". In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, 459–464. IEEE.

Grauman, K., and Darrell, T. 2007. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research* 8(Apr):725–760.

Guérin-Dugué, A., and Oliva, A. 2000. Classification of scene photographs from local orientations features. *Pattern Recognition Letters* 21(13):1135–1140.

Hoiem, D.; Efros, A. A.; and Hebert, M. 2005. Automatic photo pop-up. *ACM transactions on graphics (TOG)* 24(3):577–584.

Julesz, B. 1981. Textons, the elements of texture perception, and their interactions. *Nature* 290(5802):91–97.

Julesz, B. 1986. Texton gradients: The texton theory revisited. *Biological cybernetics* 54(4):245–251.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. 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, 2169–2178. IEEE.

Li, L.-J., and Fei-Fei, L. 2007. What, where and who? classifying events by scene and object recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 1–8. IEEE.

Li, F. F.; VanRullen, R.; Koch, C.; and Perona, P. 2003. Natural scene categorization in the near absence of attention: further explorations. *Journal of Vision* 3(9):331–331.

Li, L.-J.; Su, H.; Fei-Fei, L.; and Xing, E. P. 2010. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, 1378–1386.

Lipson, P.; Grimson, E.; and Sinha, P. 1997. Configuration based scene classification and image indexing. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1007–1013. IEEE.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, 1150–1157. Ieee.

Marr, D. 1977. Artificial intelligence – a personal view. *Artificial Intelligence* 9(1):37–48.

Marr, D. 1982. Vision: A computational approach.

Miyajima, K., and Ralescu, A. 1993. Modeling of natural objects including fuzziness and application to image under-

standing. In *Fuzzy Systems, 1993., Second IEEE International Conference on*, 1049–1054. IEEE.

Miyajima, K., and Ralescu, A. 1994. Spatial organization in 2d segmented images: representation and recognition of primitive spatial relations. *Fuzzy Sets and Systems* 65(2-3):225–236.

Murphy, K.; Torralba, A.; Freeman, W.; et al. 2003. Using the forest to see the trees: a graphical model relating features, objects and scenes. *Advances in neural information processing systems* 16:1499–1506.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision* 42(3):145–175.

Oliva, A., and Torralba, A. 2006. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research* 155:23–36.

Pandey, M., and Lazebnik, S. 2011. Scene recognition and weakly supervised object localization with deformable part-based models. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 1307–1314. IEEE.

Ralescu, A. L., and Baldwin, J. F. 1987. Concept learning from examples with applications to a vision learning system. In *Alvey Vision Conference*, 1–8.

Ralescu, A. L., and Baldwin, J. F. 1989. Concept learning from examples and counter examples. *International Journal of Man-Machine Studies* 30(3):329–354.

Ralescu, A. L. 1995. Image understanding = verbal description of the image contents. *SOFT, Journal of the Japanese Society for Fuzzy Theory* 7(4):739–746.

Renninger, L. W., and Malik, J. 2004. When is scene identification just texture recognition? *Vision research* 44(19):2301–2311.

Rock, I. 1990. The perceptual world. *Scientific American* 127.

Rosch, E., and Mervis, C. B. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology* 7(4):573–605.

Rosch, E.; Mervis, C. B.; Gray, W. D.; Johnson, D. M.; and Boyes-Braem, P. 1976. Basic objects in natural categories. *Cognitive psychology* 8(3):382–439.

Rosch, E. H. 1973. Natural categories. *Cognitive psychology* 4(3):328–350.

Serrano, N.; Savakis, A. E.; and Luo, J. 2004. Improved scene classification using efficient low-level features and semantic cues. *Pattern Recognition* 37(9):1773–1784.

Singhal, A.; Luo, J.; and Zhu, W. 2003. Probabilistic spatial context models for scene content understanding. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, I–I. IEEE.

Sivic, J., and Zisserman, A. 2009. Efficient visual search of videos cast as text retrieval. *IEEE transactions on pattern analysis and machine intelligence* 31(4):591–606.

Sowa, J. F. 1983. Conceptual structures: information processing in mind and machine.

Szummer, M., and Picard, R. W. 1998. Indoor-outdoor image classification. In *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*, 42–51. IEEE.

Thorpe, S.; Fize, D.; and Marlot, C. 1996. Speed of processing in the human visual system. *nature* 381(6582):520.

Van De Sande, K.; Gevers, T.; and Snoek, C. 2010. Evaluating color descriptors for object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 32(9):1582–1596.

Vogel, J., and Schiele, B. 2004. A semantic typicality measure for natural scene categorization. In *Joint Pattern Recognition Symposium*, 195–203. Springer.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, 3485–3492. IEEE.

Yang, C.; Liu, H.; Wang, S.; and Liao, S. 2016. Scene-level geographic image classification based on a covariance descriptor using supervised collaborative kernel coding. *Sensors* 16(3):392.

Yao, J.; Fidler, S.; and Urtasun, R. 2012. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 702–709. IEEE.