What, Where and Who? Telling the Story of an Image by Activity Classification, Scene Recognition and Object Categorization

Li Fei-Fei and Li-Jia Li

Abstract We live in a richly visual world. More than one third of the entire human brain is involved in visual processing and understanding. Psychologists have shown that the human visual system is particularly efficient and effective in perceiving high-level meanings in cluttered real-world scenes, such as objects, scene classes, activities and the stories in the images. In this chapter, we discuss a generative model approach for classifying complex human activities (such as croquet game, snowboarding, etc.) given a single static image. We observe that object recognition in the scene as well as scene environment classification of the image facilitate each other in the overall activity recognition task. We formulate this observation in a graphical model representation where activity classification is achieved by combining information from both the object recognition and the scene classification pathways. For evaluating the robustness of our algorithm, we have assembled a challenging dataset consisting real-world images of eight different sport events, most of them collected from the Internet. Experimental results show that our hierarchical model performs better than existing methods.

1 Introduction and Motivation

One of the most remarkable feats of the human visual system is how rapidly, accurately and comprehensively it can recognize and understand the complex visual world. The various types of tasks related to understanding what we see in the visual world is called "visual recognition". When presented with a real-world image, such as the top image of Fig.1, what do you see? It is a colorful image. On the top of the

Li Fei-Fei
Dept. of Computer Science, Stanford University, USA, e-mail: feifeili@cs.stanford.edu

Li-Jia Li
Dept. of Computer Science, Stanford University, USA e-mail: lijiali@cs.stanford.edu

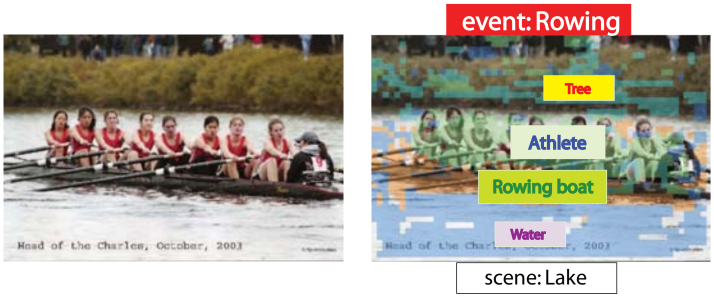

event: Rowing
Tree
Athlete
Rowing boat
Water
scene: Lake
Head of the Charles, October, 2003
Head of the Charles, October, 2003

Fig. 1 Telling the *what, where and who* story. Given an *event* (rowing) image such as the one on the left, our system can automatically interpret what is the event, where does this happen and who (or what kind of objects) are in the image. The result is represented in the figure on the right. A red name tag over the image represents the event category. The scene category label is given in the white tag below the image. A set of name tags are attached to the estimated centers of the objects to indicate their categorical labels. As an example, from the bottom image, we can tell from the name tags that this is a rowing sport event held on a lake (scene). In this event, there are rowing boat, athletes, water and trees (objects).

picture is mostly green color while the lower half is dominated by light blue, red and darker colors. There are salient edges in the foreground of the pictures, rows of round shapes paint a vivid visual picture in our mind. The different attributes of the images we describe, such as colors, edges, shapes, and textures, have been important research topics in the computer vision field. Recognizing these components of an image provide very important and useful information in a large number of practical applications. But this is not the level we communicate on and remember the visual world. It is also not the kind of description we would provide to a blind person. For most of us, this picture can be interpreted as a rich amount of semantically meaningful information. Now imagine the same scene, but this time I will describe it as a rowing event taking place on a lake. The water is clean and blue. Lush green trees stand along the shore of the lake in the background. A team of women athletes in red vests is training on a rowboat, accelerating to the right. I hope this time your mental imagery is much more vivid and meaningful than the first time. This is also the most natural way for most of us to interpret and describe our visual world. This kind of semantic interpretation of the visual world is called high-level visual recognition, part of the larger field known as vision. Vision is one of the most fundamental and important functionalities of an intelligence system. Humans rely on vision to survive, socialize and perform most of their daily tasks.

Recently, a psychophysics study has shown that in a single glance of an image, humans can not only recognize or categorize many of the individual objects in the scene, tell apart the different environments of the scene, but also perceive complex activities and social interactions [1]. In computer vision, a lot of progress has been made in object recognition and classification in recent years (see [2] for a review). A number of algorithms have also provided effective models for scene environment categorization [3, 4, 5, 6]. But little has been done in event recognition in static images. In this work, we define an *event* to be a semantically meaningful human activity, taking place within a selected environment and containing a number of necessary objects. We present a first attempt to mimic the human ability of recognizing an event and its encompassing objects and scenes. Fig.1 best illustrates the goal of this work. We would like to achieve event categorization by as much semantic level image interpretation as possible. This is somewhat like what a school child does when learning to write a descriptive sentence of the event. It is taught that one should pay attention to the 5 W's: who, where, what, when and how. In our system, we try to answer 3 of the 5 W's: *what* (the **event** label), *where* (the **scene** environment label) and *who* (a list of the **object categories**).

Similar to object and scene recognition, event classification is both an intriguing scientific question as well as a highly useful engineering application. From the scientific point of view, much needs to be done to understand how such complex and high level visual information can be represented in efficient yet accurate way. In this work, we propose to decompose an event into its scene environment and the objects within the scene. We assume that the scene and the objects are independent of each other given an event. But both of their presences influence the probability of recognizing the event. We made a further simplification for classifying the objects in an event. Our algorithm ignores the positional and interactive relationships among the objects in an image. In other words, when athletes and mountains are observed, the event of rock climbing is inferred, in spite of whether the athlete is actually on the rock performing the climbing. Much needs to be done in both human visual experiments as well as computational models to verify the validity and effectiveness of such assumptions. From an engineering point of view, event classification is a useful task for a number of applications. It is part of the ongoing effort of providing effective tools to retrieve and search semantically meaningful visual data. Such algorithms are at the core of the large scale search engines and digital library organizational tools. Event classification is also particularly useful for automatic annotation of images, as well as descriptive interpretation of the visual world for visually-impaired patients.

2 Overall Approach

Our model integrates scene and object level image interpretation in order to achieve the final event classification. Let's use the sport game polo as an example. In the foreground, a picture of the polo game usually consists of distinctive objects such as horses and players (in polo uniforms). The setting of the polo field is normally a grassland. Following this intuition, we model an event as a combination of scene and a group of representative objects. The goal of our approach is not only to classify the images into different event categories, but also to give meaningful, semantic labels to the scene and object components of the images.

3 Literature Review

While our approach is an integrative one, our algorithm is built upon several established ideas in scene and object recognition. To the first order of approximation, an event category can be viewed as a scene category. Intuitively, a snowy mountain slope can predict well an event of skiing or snow-boarding. A number of previous works have offered ways of recognizing scene categories [4, 5, 6]. Most of these algorithms learn global statistics of the scene categories through either frequency distributions or local patch distributions. In the scene part of our model, we adopt a similar algorithm as Fei-Fei et al. [6]. In addition to the scene environment, event recognition relies heavily on foreground objects such as players and ball for a soccer game. Object categorization is one of the most widely researched areas recently. One could grossly divide the literature into those that use generative models (e.g. [7, 8, 9]) and those that use discriminative models or methods (e.g. [10, 11]). Given our goal is to perform event categorization by integrating scene and object recognition components, it is natural for us to use a generative approach. Our object model is adapted from the bag of words models that have recently shown much robustness in object categorization [12, 13, 14]. As [15] points out, other than scene and object level information, general layout of the image also contributes to our complex yet robust perception of a real-world image. Much can be included here for general layout information, from a rough sketch of the different regions of the image to a detailed 3D location and shape of each pixels of the image. We choose to demonstrate the usefulness of the layout/geometry information by using a simple estimation of 3 geometry cues: sky at infinity distance, vertical structure of the scene, and ground plane of the scene [16]. It is important to point out here that while each of these three different types of information is highly useful for event recognition (scene level, object level, layout level), our experiments show that we only achieve the most satisfying results by integrating all of them (Sec.7).

Several previous works have taken on a more holistic approach in scene interpretation [17, 18, 19, 20]. In all these works, global scene level information is incorporated in the model for improving better object recognition or detection. Mathematically, our approach is closest in spirit with Sudderth et al [19]. We both learn a generative model to label the images. And at the object level, both of our models are based on the bag of words approach. Our model, however, differs fundamentally from the previous works by providing a set of integrative and hierarchical labels of an image, performing the *what*(event), *where*(scene) and *who*(object) recognition of an entire scene.

4 The Integrative Model

Given an image of an event, our algorithm aims to not only classify the type of event, but also to provide meaningful, semantic labels to the scene and object components of the images.

To incorporate all these different levels of information, we choose a generative model to represent our image. Fig.2 illustrates the graphical model representation. We first define the variables of the model, and then show how an image of a particular event category can be generated based on this model. For each image of an event, our fundamental building blocks are densely sampled local image patches (sampling grid size is 10×10). In recent years, interest point detectors have demonstrated much success in object level recognition (e.g. [21, 22, 23]). But for a holistic scene interpretation task, we would like to assign semantic level labels to as many pixels as possible on the image. It has been observed that tasks such as scene classification benefit more from a dense uniform sampling of the image than using interest point detectors [5, 6]. Each of these local image patches then goes on to serve both the scene recognition part of the model, as well as the object recognition part. For scene recognition, we denote each patch by X in Fig.2. X only encodes here appearance based information of the patch (e.g. a SIFT descriptor [21]). For the object recognition part, two types of information are obtained for each patch. We denote the appearance information by A, and the layout/geometry related information by G. A is similar to X in expression. G in theory, however, could be a very rich set of descriptions of the geometric or layout properties of the patch, such as 3D location in space, shape, and so on. For scenes subtending a reasonably large space (such as these event scenes), such geometric constraint should help recognition. In Sec.6, we discuss the usage of three simple geometry/layout cues: verticalness, sky at infinity and the ground-plane.[1]

We now go over the graphical model (Fig.2) and show how we generate an event picture. Note that each node in Fig.2 represents a random variable of the graphical model. An open node is a latent (or unobserved) variable whereas a darkened node is observed during training. The lighter gray nodes (event, scene and object labels) are only observed during training whereas the darker gray nodes (image patches) are observed in both training and testing.

1. An **event** category is represented by the discrete random variable E. We assume a fixed uniform prior distribution of E, hence omitting showing the prior distribution in Fig.2. We select $E \sim p(E)$. The images are indexed from 1 to I and one E is generated for each of them.
2. Given the event class, we generate the **scene** image of this event. There are in theory S classes of scenes for the whole event dataset. For each event image, we assume only one scene class can be drawn.

    - A scene category is first chosen according to $S \sim p(S|E, \psi)$. S is a discrete variable denoting the class label of the scene. ψ is the multinomial parameter

[1] The theoretically minded machine learning readers might notice that the observed variables X, A and G occupy the same physical space on the image. This might cause the problem of “double counting”. We recognize this potential confound. But in practice, since our estimations are all taken placed on the same “double counted” space in both learning and testing, we do not observe a problem. One could also argue that even though these features occupy the same physical locations, they come from different “image feature space”. Therefore this problem does not apply. It is, however, a curious theoretical point to explore further.

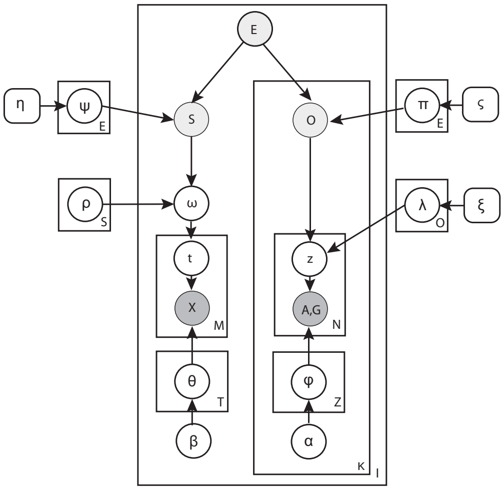


Fig. 2 Graphical model of our approach. E, S, and O represent the event, scene and object labels respectively. X is the observed appearance patch for scene. A and G are the observed appearance and geometry/layout properties for the object patch. The rest of the nodes are parameters of the model. For details, please refer to Sec.4

that governs the distribution of S given E. ψ is a matrix of size $E \times S$, whereas η is an S dimensional vector acting as a Dirichlet prior for ψ.

- Given S, we generate the mixing parameters ω that governs the distribution of scene patch topics $\omega \sim p(\omega|S,\rho)$. Elements of ω sum to 1 as it is the multinomial parameter of the latent topics t. ρ is the Dirichlet prior of ω, a matrix of size $S \times T$, where T is the total number of the latent topics.
- A patch in the scene image is denoted by X. To generate each of the M patches
  - Choose the latent topic $t \sim \text{Mult}(\omega)$. t is a discrete variable indicating which latent topic this patch will come from.
  - Choose patch $X \sim p(X|t,\theta)$, where θ is a matrix of size $T \times V_S$. V_S is the total number of vocabularies in the scene codebook for X. θ is the multinomial parameter for discrete variable X, whereas β is the Dirichlet prior for θ.

3. Similar to the scene image, we also generate an **object** image. Unlike the scene, there could be more than one objects in an image. We use K to denote the number of objects in a given image. There is a total of O classes of objects for the whole dataset. The following generative process is repeated for each of the K objects in an image.

- An object category is first chosen according to $O \sim p(O|E,\pi)$. O is a discrete variable denoting the class label of the object. A multinomial parameter π governs the distribution of O given E. π is a matrix of size $E \times O$, whereas ς is a O dimensional vector acting as a Dirichlet prior for π.
- Given O, we are ready to generate each of the N patches A,G in the k^{th} object of the object image
  - Choose the latent topic $z \sim \text{Mult}(\lambda|O)$. z is a discrete variable indicating which latent topic this patch will come from, whereas λ is the multinomial

parameter for z, a matrix of size $O \times Z$. K is the total number of objects appear in one image, and Z is the total number of latent topics. ξ is the Dirichlet prior for λ.

- Choose patch $A, G \sim p(A, G|t, \varphi)$, where φ is a matrix of size $Z \times V_O$. V_O is the total number of vocabularies in the codebook for A, G. φ is the multinomial parameter for discrete variable A, G, whereas α is the Dirichelet prior for φ. Note that we explicitly denote the patch variable as A, G to emphasize on the fact it includes both appearance and geometry/layout property information.

Putting everything together in the graphical model, we arrive at the following joint distribution for the image patches, the event, scene, object labels and the latent topics associated with these labels.

$$\begin{aligned} & p(E, S, \mathbf{O}, \mathbf{X}, \mathbf{A}, \mathbf{G}, \mathbf{t}, \mathbf{z}, \omega|\rho, \varphi, \lambda, \psi, \pi, \theta) = \\ & p(E) \cdot p(S|E, \psi) p(\omega|S, \rho) \prod_{m=1}^{M} p(X_m|t_m, \theta) p(t_m|w) \\ & \cdot \prod_{k=1}^{K} p(O_k|E, \pi) \prod_{n=1}^{N} p(A_n, G_n|z_n, \varphi) p(z_n|\lambda, O_k) \end{aligned} \tag{1}$$

where $\mathbf{O}, \mathbf{X}, \mathbf{A}, \mathbf{G}, \mathbf{t}, \mathbf{z}$ represent the generated objects, appearance representation of patches in the scene part, appearance and geometry properties of patches in the object part, topics in the scene part, and topics in the object part respectively. Each component of Eq.1 can be broken into

$$p(S|E, \psi) = \mathrm{Mult}(S|E, \psi) \tag{2}$$

$$p(\omega|S, \rho) = \mathrm{Dir}(\omega|\rho_{j\cdot}), S = j \tag{3}$$

$$p(t_m|\omega) = \mathrm{Mult}(t_m|\omega) \tag{4}$$

$$p(X_m|t, \theta) = p(X_m|\theta_{j\cdot}), t_m = j \tag{5}$$

$$p(O|E, \pi) = \mathrm{Mult}(O|E, \pi) \tag{6}$$

$$p(z_n|\lambda, O) = \mathrm{Mult}(z_n|\lambda, O) \tag{7}$$

$$p(A_n, G_n|z, \varphi) = p(A_n, G_n|\varphi_{j\cdot}), z_n = j \tag{8}$$

where "·" in the equations represents components in the row of the corresponding matrix.

4.1 Labeling an Unknown Image

Given an unknown event image with unknown scene and object labels, our goal is: 1) to classify it as one of the event classes (*what*); 2) to recognize the scene environment class (*where*); and 3) to recognize the object classes in the image (*who*).

We realize this by calculating the maximum likelihood at the event level, the scene level and the object level of the graphical model (Fig.2).

At the object level, the likelihood of the image given the object class is

$$p(I|O) = \prod_{n=1}^{N} \sum_{j} P(A_n, G_n|z_j, O) P(z_j|O) \tag{9}$$

The most possible objects appear in the image are based on the maximum likelihood of the image given the object classes, which is $O = \mathrm{argmax}_O p(I|O)$. Each object is labeled by showing the most possible patches given the object, represented as $O = \mathrm{argmax}_O p(A, G|O)$.

At the scene level, the likelihood of the image given the scene class is:

$$p(I|S, \rho, \theta) = \int p(\omega|\rho, S) (\prod_{m=1}^{M} \sum_{t_m} p(t_m|\omega) \cdot p(X_m|t_m, \theta)) d\omega \tag{10}$$

Similarly, the decision of the scene class label can be made based on the maximum likelihood estimation of the image given the scene classes, which is $S = \mathrm{argmax}_S p(I|S, \rho, \theta)$. However, due to the coupling of θ and ω, the maximum likelihood estimation is not tractable computationally [24]. Here, we use the variational method based on Variational Message Passing [25] provided in [6] for an approximation.

Finally, the image likelihood for a given event class is estimated based on the object and scene level likelihoods:

$$p(I|E) \propto \sum_{j} P(I|O_j) P(O_j|E) P(I|S) P(S|E) \tag{11}$$

The most likely event label is then given according to $E = \mathrm{argmax}_E p(I|E)$.

5 Learning the Model

The goal of learning is to update the parameters $\{\psi, \rho, \pi, \lambda, \theta, \beta\}$ in the hierarchical model (Fig.2). Given the event E, the scene and object images are assumed independent of each other. We can therefore learn the scene-related and object-related parameters separately.

We use Variational Message Passing method to update parameters $\{\psi, \rho, \theta\}$. Detailed explanation and update equations can be found in [6]. For the object branch of the model, we learn the parameters $\{\pi, \lambda, \beta\}$ via Gibbs sampling [26] of the latent topics. In such a way, the topic sampling and model learning are conducted iteratively. In each round of the Gibbs sampling procedure, the object topic will be sampled based on $p(z_i|\mathbf{z}_{\setminus i}, A, G, O)$, where $\mathbf{z}_{\setminus i}$ denotes all topic assignment except the current one. Given the Dirichlet hyperparameters ξ and α, the distribution of topic given object $p(z|O)$ and the distribution of appearance and geometry words given topic $p(A,G|z)$ can be derived by using the standard Dirichlet integral formulas:

$$p(z=i|\mathbf{z}_{\setminus i}, O=j) = \frac{c_{ij}+\xi}{\Sigma_i c_{ij}+\xi \times H} \tag{12}$$

$$p((A,G)=k|\mathbf{z}_{\setminus i}, z=i) = \frac{n_{ki}+\varphi}{\Sigma_k n_{ki}+\varphi \times V_O} \tag{13}$$

where c_{ij} is the total number of patches assigned to object j and object topic i, while n_{ki} is the number of patch k assigned to object topic i. H is the number of object topics, which is set to some known, constant value. V_O is the object codebook size. And a patch is a combination of appearance (A) and geometry (G) features. By combining Eq.12 and 13, we can derive the posterior of topic assignment as

$$\begin{aligned} p(z_i|\mathbf{z}_{\setminus i}, A, G, O) = p(z=i|\mathbf{z}_{\setminus i}, O) \times \\ p((A,G)=k|\mathbf{z}_{\setminus i}, z=i) \end{aligned} \tag{14}$$

Current topic will be sampled from this distribution.

6 System Implementation

Our goal is to extract as much information as possible out of the event images, most of which are cluttered, filled with objects of variable sizes and multiple categories. At the feature level, we use a grid sampling technique similar to [6]. In our experiments, the grid size is 10×10. A patch of size 12×12 is extracted from each of the grid centers. A 128-dim SIFT vector is used to represent each patch [21]. The poses of the objects from the same object class change significantly in these events. Thus, we use rotation invariant SIFT vector to better capture the visual similarity within each object class. A codebook is necessary in order to represent an image as a sequence of appearance words. We build a codebook of 300 visual words by applying K-means for the 200000 SIFT vectors extracted from 30 randomly chosen training images per event class. To represent the geometry/layout information, each pixel in an image is given a geometry label using the codes provided by [18]. In this approach, only three simple geometry/layout properties are used. They are: ground plane, vertical structure and sky at infinity. Each patch is assign a geometry membership by the major vote of the pixels within.

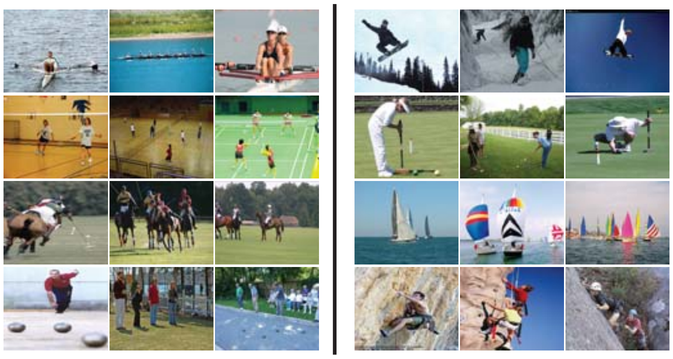


Fig. 3 Our dataset contains 8 sports event classes: rowing (250 images), badminton (200 images), polo (182 images), bocce (137 images), snowboarding (190 images), croquet (236 images), sailing (190 images), and rock climbing (194 images). In this figure, each triplet is randomly selected from our dataset. Our examples here demonstrate the complexity and diversity of this highly challenging dataset.

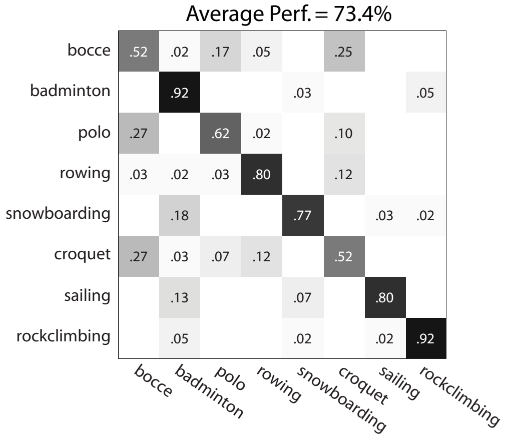
7 Experiments and Results

7.1 Dataset

As the first attempt to tackle the problem of static event recognition, we have no existing dataset to use and compare with. Instead we have compiled a new dataset containing 8 sports event categories collected from the Internet: bocce, croquet, polo, rowing, snowboarding, badminton, sailing, and rock climbing. The number of images in each category varies from 137 (bocce) to 250 (rowing). As shown in Fig. 3, this event dataset is a very challenging one. Here we highlight some of the difficulties.

- The background of each image is highly cluttered and diverse;
- Object classes are diverse;
- Within the same category, sizes of instances from the same object are very different;
- The pose of the objects can be very different in each image;
- Number of instances of the same object category change diversely even within the same event category;
- Some of the foreground objects are too small to be detected.

We have also obtained a thorough groundtruth annotation for every image in the dataset (in collaboration with Lotus Hill Research Institute [27]). This annotation provides information for: event class, background scene class(es), most discernable object classes, and detailed segmentation of each objects.



	bocce	badminton	polo	rowing	snowboarding	croquet	sailing	rockclimbing
bocce	.52	.02	.17	.05		.25		
badminton		.92			.03			.05
polo	.27		.62	.02		.10		
rowing	.03	.02	.03	.80		.12		
snowboarding		.18			.77		.03	.02
croquet	.27	.03	.07	.12		.52		
sailing		.13			.07		.80	
rockclimbing		.05			.02		.02	.92

Fig. 4 Confusion table for the 8-class event recognition experiment. The average performance is 73.4%. Random chance would be 12.5%.

7.2 Experimental Setup

We set out to learn to classify these 8 events as well as labeling the semantic contents (scene and objects) of these images. For each event class, 70 randomly selected images are used for training and 60 are used for testing. We do not have any previous work to compare to. But we test our algorithm and the effectiveness of each components of the model. Specifically, we compare the performance of our full integrative model with the following baselines.

- A *scene only* model. We use the LDA model of [6] to do event classification based on scene categorization only. We “turn off” the influence of the object part by setting the likelihood of O in Eq.11 to a uniform distribution. This is effectively a standard “bag of words” model for event classification.
- An *object only* model. In this model we learn and recognize an event class based on the distribution of foreground objects estimated in Eq.9. No geometry/layout information is included. We “turn off” the influence of the scene part by setting the likelihood of S in Eq.11 to a uniform distribution.
- A *object* + *geometry* model. Similar to the object-only model, here we include the feature representations of both appearance (A) and geometry/layout (G).

Except for the LDA model, training is supervised by having the object identities labeled. We use exactly the same training and testing images in all of these different model conditions.

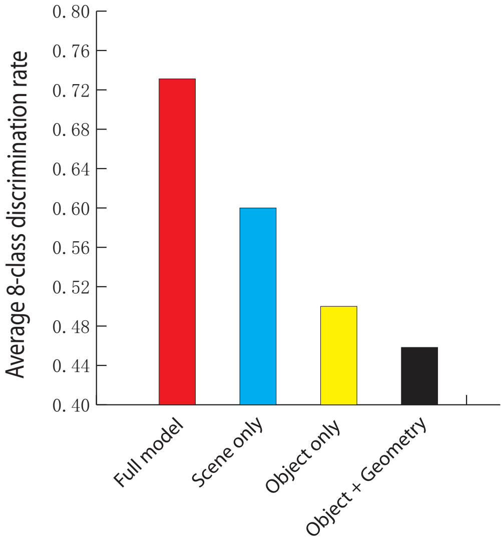


Fig. 5 Performance comparison between the full model and the three control models (defined in Sec.7.2). The x-axis denotes the name of the model used in each experiment. The ‘full model’ is our proposed integrative model (see Fig.2). The y-axis represents the average 8-class discrimination rate, which is the average score of the diagonal entries of the confusion table of each model.

7.3 Results

We report an overall 8-class event discrimination of 73.4% by using the full integrative model. Fig.4 shows the confusion table results of this experiment. In the confusion table, the rows represent the models for each event category while the columns represent the ground truth categories of events. It is interesting to observe that the system tends to confuse bocce and croquet, where the images tend to share similar foreground objects. On the other hand, polo is also more easily confused with bocce and croquet because all of these events often take places in grassland type of environments. These two facts agree with our intuition that an event image could be represented as a combination of the foreground objects and the scene environment.

In the control experiment with different model conditions, our integrative model consistently outperforms the other three models (see Fig.5). A curious observation is that the *object* + *geometry* model performs worse than the *object only* model. We believe that this is largely due to the simplicity of the geometry/layout properties. While these properties help to differentiate sky, ground from vertical structures, they also introduce noise. As an example, water and snow are always incorrectly classified as sky or ground by the geometry labeling process, which deteriorates the result of object classification. However, the scene recognition alleviates the confusion among water, snow, sky and ground by encoding explicitly their different appearance properties. Thus, when the scene pathway is added to the integrated model, the overall results become much better.

Finally, we present more details of our image interpretation results in Fig.6. At the beginning of this chapter, we set out to build an algorithm that can tell a *what*, *where* and *who* story of the sport event pictures. We show here how each of these W's is answered by our algorithm. Note all the labels provided in this figure are automatically generated by the algorithm, no human annotations are involved.

8 Conclusion

Semantic interpretation of the visual world is an indispensable functionality of the future generations of artificial intelligence system. This project aims to contribute to both the scientific questions of image modeling and the technological advancement of visual intelligence. One of the most important applications is personal assistance to visually-impaired or blind people. Currently, other than specific domain applications such as texts and faces, little technology is available to assist them to interpret and analyze the visual environment in a comprehensive and meaningful way. Semantic understanding of images could serve to advance the state of the art assistance in this domain. It will also improve real-word applications that require advanced visual recognition tools. One example is the increasing need for sophisticated and meaningful sorting and searching tools for large image datasets, such as personal photo collections and images on the internet. Our model is, of course, just the first attempt for such an ambitious goal.

References

1. L. Fei-Fei, A. Iyer, C. Koch, and P. Perona. What do we see in a glance of a scene? *Journal of Vision*, 7(1):10, 1–29, 2007. http://journalofvision.org/7/1/10/, doi:10.1167/7.1.10.
2. L. Fei-Fei, R. Fergus, and A. Torralba. Recognizing and learning object categories. Short Course CVPR: http://people.csail.mit.edu/torralba/shortCourseRLOC/index.html, 2007.
3. M. Szummer and R. Picard. Indoor-outdoor image classification. In *Int. Workshop on Content-based Access of Image and Vedeo Databases*, Bombay, India, 1998.
4. A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. Journal of Computer Vision.*, 42, 2001.
5. J. Vogel and B. Schiele. A semantic typicality measure for natural scene categorization. In *DAGM'04 Annual Pattern Recognition Symposium*, Tuebingen, Germany, 2004.
6. L. Fei-Fei and P. Perona. A Bayesian hierarchy model for learning natural scene categories. *CVPR*, 2005.
7. M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. European Conference on Computer Vision*, volume 2, pages 101–108, 2000.
8. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. Computer Vision and Pattern Recognition*, pages 264–271, 2003.
9. M. P. Kumar, P. H. S. Torr, and A. Zisserman. Obj cut. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1*, pages 18–25, Washington, DC, USA, 2005. IEEE Computer Society.
10. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. Computer Vision and Pattern Recognition*, volume 1, pages 511–518, 2001.
11. H. Zhang, A. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. *Proc. CVPR*, 2006.
12. G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

13. J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. In *Proc. International Conference on Computer Vision*, 2005.
14. L.-J. Li, G. Wang, and L. Fei-Fei. Optimol: automatic online picture collection via incremental model learning. In *Proc. Computer Vision and Pattern Recognition*, 2007.
15. J. Wolfe. Visual memory: what do you know about what you saw? *Curr. Bio.*, 8:R303–R304, 1998.
16. D. Hoiem, A. Efros, and M. Hebert. Automatic photo pop-up. *Proceedings of ACM SIGGRAPH 2005*, 24(3):577–584, 2005.
17. K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the trees:a graphical model relating features, objects and scenes. In *NIPS (Neural Info. Processing Systems)*, 2004.
18. D. Hoiem, A. Efros, and M. Hebert. Putting Objects in Perspective. *Proc. IEEE Computer Vision and Pattern Recognition*, 2006.
19. E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Learning hierarchical models of scenes, objects, and parts. In *Proc. International Conference on Computer Vision*, 2005.
20. Z. Tu, X. Chen, A. Yuille, and S. Zhu. Image Parsing: Unifying Segmentation, Detection, and Recognition. *International Journal of Computer Vision*, 63(2):113–140, 2005.
21. D. Lowe. Object recognition from local scale-invariant features. In *Proc. International Conference on Computer Vision*, 1999.
22. G. Dorko and C. Schmid. Object class recognition using discriminative local features. *IEEE PAMI*, submitted.
23. S. Obdrzalek and J. Matas. Object recognition using local affine frames on distinguished regions. *Proc. British Machine Vision Conference*, pages 113–122, 2002.
24. D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
25. J. Winn and C. M. Bishop. Variational message passing. *J. Mach. Learn. Res.*, 6:661–694, 2004.
26. S. Krempp, D. Geman, and Y. Amit. Sequential learning with reusable parts for object detection. Technical report, Johns Hopkins University, 2002.
27. Z.-Y. Yao, X. Yang, and S.-C. Zhu. Introduction to a large scale general purpose groundtruth dataset: methodology, annotation tool, and benchmarks. In *6th Int'l Conf on EMMCVPR*, 2007.

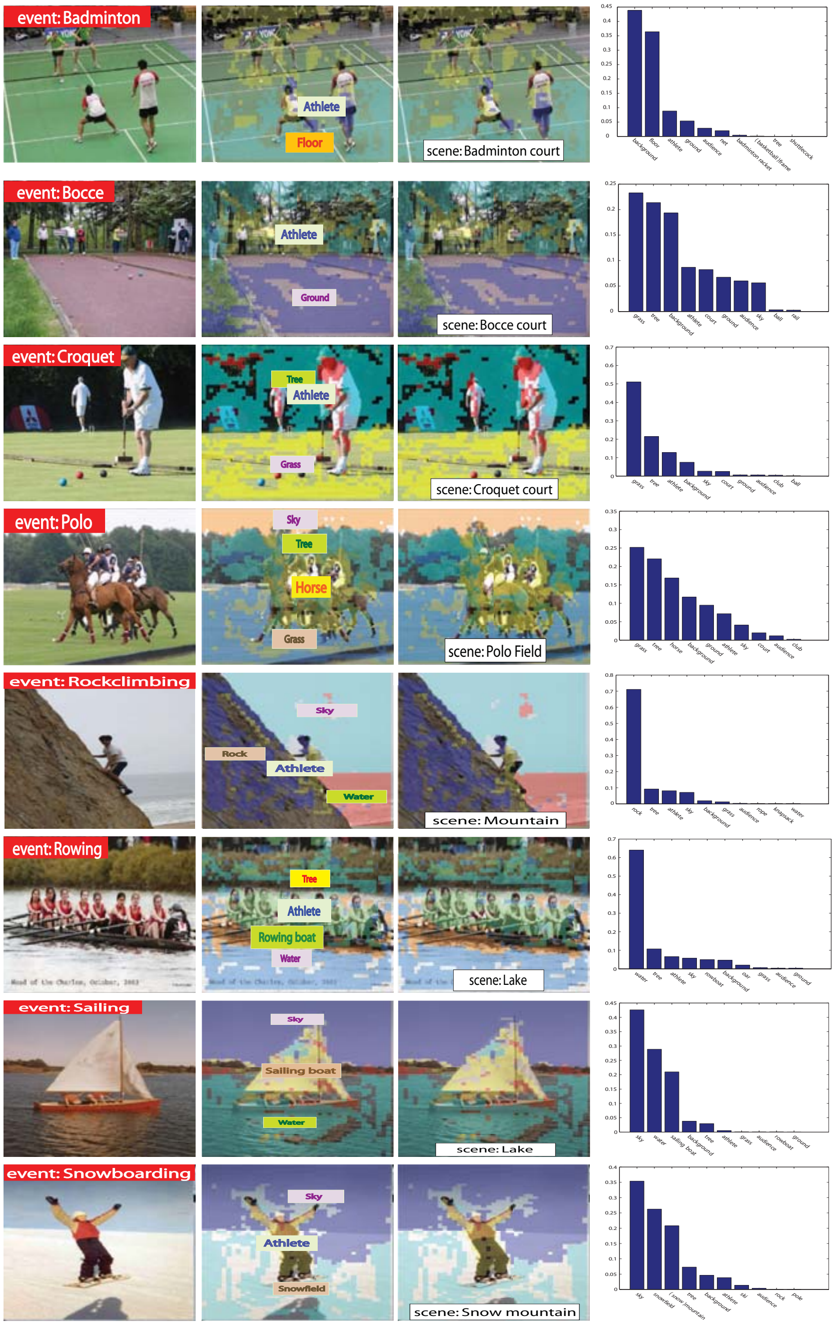


Fig. 6 (This figure is best viewed in color and with PDF magnification.) Image interpretation via event, scene, and object recognition. Each row shows results of an event class. **Column 1** shows the event class label. **Column 2** shows the object classes recognized by the system. Masks with different colors indicate different object classes. The name of each object class appears at the estimated centroid of the object. **Column 3** is the scene class label assigned to this image by our system. Finally **Column 4** shows the sorted object distribution given the event. Names on the x-axis represents the object class, the order of which varies across the categories. y-axis represents the distribution.