Selective Hidden Random Fields: Exploiting Domain-Specific Saliency for Event Classification

Vidit Jain*
*University of Massachusetts Amherst
Amherst MA USA
vidit@cs.umass.edu

Amit Singhal†
†Eastman Kodak Company
Rochester NY USA

Jiebo Luo†

Abstract

Classifying an event captured in an image is useful for understanding the contents of the image. The captured event provides context to refine models for the presence and appearance of various entities, such as people and objects, in the captured scene. Such contextual processing facilitates the generation of better abstractions and annotations for the image. Consider a typical set of consumer images with sports-related content. These images are taken mostly by amateur photographers, and often at a distance. In the absence of manual annotation or other sources of information such as time and location, typical recognition tasks are formidable on these images. Identifying the sporting event in these images provides a context for further recognition and annotation tasks. We propose to use the domain-specific saliency of the appearances of the playing surfaces, and ignore the noninformative parts of the image such as crowd regions, to discriminate among different sports. To this end, we present a variation of the hidden-state conditional random field that selects a subset of the observed features suitable for classification. The inferred hidden variables in this model represent a selection criteria desirable for the problem domain. For sports-related images, this selection criteria corresponds to the segmentation of the playing surface in the image. We demonstrate the utility of this model on consumer images collected from the Internet.

1. Introduction

The ease and convenience of digital photography relative to film photography have resulted in an explosion in the number of personal photographs. The size of these collections presents difficulty in browsing and searching for specific images in these collections, obviating the need of a system that could select a subset of such collections based upon rich queries like "Alice playing tennis" or "Bob at a soccer game". In the absence of effective content-based image retrieval systems, retrieving images relevant to such queries is dependent on the meta-information associated with the images such as tag words. Manually annotating individual images with key words, however, is a tedious task, which motivates the need for the automatic generation of intelligent annotation for images.

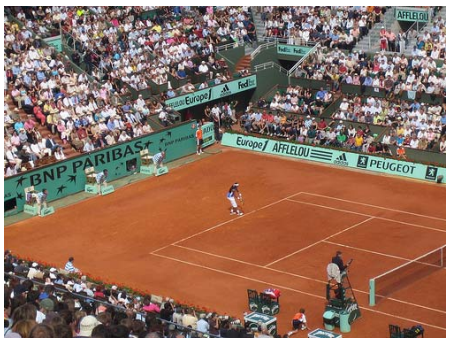

Figure 1. Most people will identify the sporting event in this image as "tennis", and associate tags like "French Open" and "clay court" with this image. A careful observer may also add "Rafael Nadal" and "serving" to the annotation. The scope of this work is to recognize the sporting event, which can be seen as providing context for subsequent tasks such as identifying people, venue, and activity.

Consider the image shown in Figure 1, which is a typical sample from a personal photo collection. Identifying the sporting event, i.e., tennis, in this image would provide context for recognizing the venue and the player. The caption of Figure 1 shows examples of the desired annotations, some of which are difficult to generate using the appearance of the corresponding image regions alone. In particular, at the current resolution, even a very sophisticated face recognizer [9] will find it difficult to identify the player in this image. Using the context that this image captures a game of tennis played on clay court in the year 2007, however, makes the identity of the player as Rafael Nadal more likely than Bjorn Borg or Tiger Woods.

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

Figure 2. A typical image in a personal collection is usually captured while sitting in the spectator area, and often has a wide-angle view of the sporting event with crowd covering most of the image region. While the resolution of such images is often too low to recognize sport accessories such as a ball or a racquet, the playing surface stands out as a reliable and robust source of information to identify the sporting event.

1.1. Playing Surface

Sports related images in a personal collection are usually captured while sitting in the stands, and are at a distance from the playing surface (see Figure 2 for an example). Spectators occupy significant areas in these images, yet provide little information about the sporting event. These image regions can be potentially distractive to the algorithms that analyze the general scene statistics to classify the sporting event shown in an image (basketball, in Figure 2). Ignoring the crowd regions in such images of most popular sports, it is reasonable to assume: (a) the playing surface for a single sport is consistent across different venues and over time; and (b) the markings on the ground are different for different sports, e.g., a lot of parallel lines on a (American) football field for yard markings as opposed to a diamond on a baseball field. These observations motivate the utility of identifying and characterizing the playing surface for recognizing the sport shown in an image.

To characterize it in terms of its distinctive markings, the playing surface needs to be segmented in a given image, which is a nontrivial task. We present a novel variation of the hidden-state conditional random fields, *selective hidden random fields*, which jointly segments a region of interest and selects the features computed on it to classify the event. This model can be applied to domains that require a selective processing of data. For example, an autonomous vehicle can use it to determine a parking lot or a crossing using the markings on the surface while ignoring the image regions that are unlikely to be labeled as road or horizontal.

1.2. Related Work

Image understanding is a well-studied field in computer vision. While a general-purpose system for interpreting the scene captured in an image is unlikely to materialize in the near future, some systems have achieved appreciable results in domain-specific settings. Recent advances were made by Hoiem et al. [7], however, the related work dates back decades to VISIONS [5]. The scope of our work is limited to the classification of popular sporting events that are played on a horizontal surface such as tennis and soccer; sports involving uneven terrain like golf and skiing are beyond the scope of this work.

Most of the research in event classification is focused on video data, where the trajectories of objects or people [8] are used to recognize different events. Specific to sports videos, several groups like Messer et al. [17] and Kittler et al. [12] developed systems for semantic annotation and summarization. Despite the abundance of sports-related images (as opposed to videos) on the Internet and in personal collections, little research has been done towards classifying sporting events or identifying players in still images. One of the most relevant works is Li et al.'s [14], where the scene structure and general image statistics are used to classify the images of eight sporting events like bocce, sailing, and snowboarding. Their work goes beyond event classification and makes annotation for image regions as well. Their work, however, relies on highly discriminative general scene statistics of different sporting events, which is not effective for discriminating among sports with very similar global appearances such as soccer and American football. For these highly similar event classes, some additional information about the scene context is required.

Berg et al. [2] and Jain et al. [10] used the context from the captions associated with images to cluster face images according to the identity of different people, but they ignored building a visual context. Because their data set was collected from news sources, the associated captions were very reliable and informative of the image content. Consumer images, on the other hand, do not have such information available with them, and thus require the context to be built from the visual cues only, which is the primary focus of this work.

To extract the distinct domain-specific features, the foreground object (playing surface, for sports images) needs to be segmented in an image. The training images are annotated with the class label but not the segmentation, which must be inferred. Hidden-state conditional random fields (HRF) [20] include latent variables to represent similar behavior. Several of its variants were found to be useful in different applications such as gesture recognition [21] and learning discriminative parts (LHRF) [11]. Both HRF and LHRF require that the segmentation of the foreground object be observed for the training images. The unavailability of these labels makes these models unsuitable for our data set. We present another model from this family for the selection of features that are part of the foreground and are

useful for the classification task. To represent this joint criterion, we model the segmentation label for the foreground as a hidden variable, and the event class as the unobserved variable. Note that for a general set of features, the hidden variables need not correspond to the selection criteria, but for our choice of features, the learned hidden variables have the desired semantics.

2. Methodology

Instead of using a decision theoretic framework for processing the data sequentially, we obtain several higher level abstractions such as lines and horizontal planes for an image and collectively use them to perform the required classification task. The abstractions used in our system are discussed in this section, and the joint probabilistic framework that uses these components is presented in Section 3.

2.1. Building Line Hypotheses

The markings on the playing surface are not very easy to recognize due to the perspective view, image resolution, occlusions, and other imaging artifacts. To approximate these markings, we use collections of straight lines and the pairwise interactions (intersections and parallelism) among them. Kosecka et al. [13] used straight lines to estimate the vanishing points in an image, which helps in building a hypothesis for the orientation (horizontal or vertical) of different image regions. We use their approach to determine long line segments in an image, and compute a histogram of orientations weighted by the length of the line segments. This histogram is rotated to center around the most frequent orientation to compensate for some changes in the viewing angle. The average orientation histograms for the lines detected on the entire image and the playing surface are shown in Figure 3.

2.2. Super-pixel Representation

Pixels are fundamental entities in an image, but a raw pixel-based representation is very high-dimensional. Pixels in an image with very similar appearances in terms of color or texture can often be collected into regions or segments, which reduces the dimensions of the image representation with a slight information loss. Selecting the permissible variations within a segment depends on the goal of the application. One such criterion is to group the pixels into a super-pixel representation based on the similarity in appearance. The resulting segments can be succinctly represented by computing statistical quantities like moments over them.

We are interested in extracting the image regions corresponding to the horizontal planes, crowd, players, and related semantic groups. Instead of designing specific analytical criteria for merging the pixels, we adopt a general-purpose technique to obtain super-pixels based on the color/appearance without using any related context. In particular, we use the mean-shift approach suggested by Comaniciu and Meer [4]. For each super-pixel, we compute several features to describe its location, shape, appearance, and geometry (see Table 1).

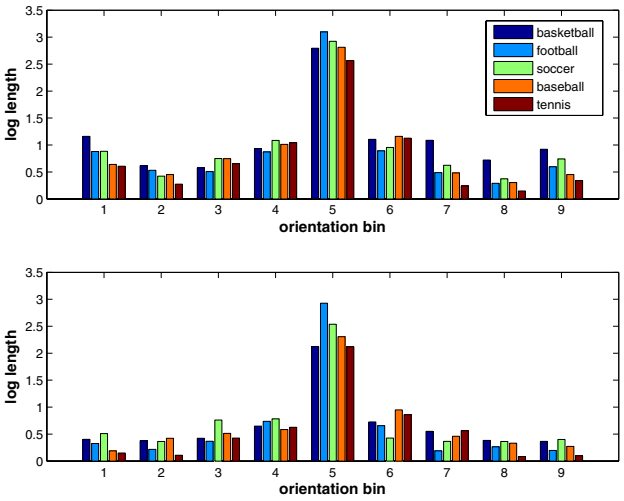

Figure 3. Average distribution of the cumulative length of lines detected in the entire image (top) and on the playing surface (bottom). The lines are clustered by their orientation, and the resulting histogram is rotated to center around the bin with maximum cumulative length. The distribution at the bottom is more useful for discrimination among the classes than is the distribution on the top.

Type	Features
Location and shape	
Position	normalized mean x,y
Shape	area, second moment
Appearance	
Color	mean(RGB), mean(HSV)
Texture	mean(8 DOOG filter responses)
Geometry	
Single line	histogram of lines in 9 orientations
Pair of lines	no. of intersections
	length of parallel lines

Table 1. Features computed for every super-pixel. DOOG filters refer to the difference of oriented Gaussian filters.

2.3. Inference of Surface Orientation

The super-pixel representation is limited by the spatial neighborhood of pixels in the image, so the super-pixels that do not share a boundary are not merged irrespective of the similarity in their appearances. A naïve algorithm can compare all pairs of super-pixels for a possibility of merging, but it is nontrivial to determine whether a super-pixel is part of the playing surface. Hoiem et al. [6] developed

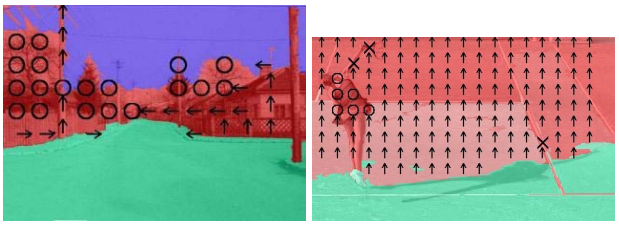

Figure 4. Surface orientation annotation obtained by Hoiem et al. [6]. Green color represents horizontal and red represents vertical. For this work, we ignore the subdivisions of the vertical surfaces: planar orientations (arrows), nonplanar solid ('x') and porous ('o'). The left image shows the results on an outdoor scene with buildings, and the right image shows the results on a sports-related image. While their results are very impressive on street scenes, we did not find the learned statistics to be useful in modeling the surface orientations in sports images.

a system that classifies the orientation of different surfaces in an image as horizontal and vertical. They obtained very impressive results on outdoor scenes containing streets and buildings. The features and statistics learned for their domain are not useful in analyzing the sports-related images, as the edges in the three principle coordinate axes are not uniformly present in all the sports images. Figure 4 shows examples of annotations obtained using their approach.

For every super-pixel, we compute the features shown in Table 1. We labeled all the super-pixels of a few images as *horizontal* or *non-horizontal* and used them to train a support vector machine (SVM) classifier. An SVM classifier is trained to minimize the cumulative error for all the training samples. Since the non-horizontal super-pixels occur about forty times more frequently than the horizontal super-pixels, the learned classifier generates a large number of false negatives for the minority class (horizontal, in this case). Since we want to minimize the average class error, we use the synthetic minority over-sampling technique (SMOTE) [3] to rectify the imbalance in the class frequencies. Using the output of this classifier, we define the following probability measure to represent the confidence in the predicted orientation:

$$p(\text{horizontal}|d) = \frac{\exp(d)}{1+\exp(d)}, \tag{1}$$

where d is the distance from the separating hyper-plane in the projected space obtained by the SVM classifier.

2.4. Interest Points and Regions

We avoid considering the complete set of image patches over all sizes, shapes, and resolution by sub-sampling the image using an interest point detector. In a detailed comparison of scale and affine invariant interest point detectors [18], Mikolajczyk and Schmid found the extrema of difference of Gaussian (DoG) operator [15], and maximally stable extrema regions (MSER) [16] to be useful for a variety of visual scene categories. Both of these detectors, however, generate only a few samples on relatively uniform regions like the horizontal playing surface in a sports-related image. The affine invariant Harris corner detector, on the other hand, responds to relatively local yet salient interest points on these image regions.

We use these detectors: DoG, MSER, and affine-Harris, to sub-sample an image, and use the scale invariant feature transform (commonly known as SIFT) [15] to represent the detected interest points. A SIFT descriptor requires the scale at which the interest point/region is detected. The MSER detector does not work in the scale space of an image, so we normalized the image region by fitting an ellipse, using the method of moments, to compute the appropriate SIFT descriptor. A similar normalization was used for affine-invariant Harris corners.

2.5. Bag of Visual Words

Representing an image as an unordered set of image patches or "bag of visual words" has been found to be useful in many computer vision tasks. The usual approach to obtain this representation is to cluster the interest point/region representation (discussed in previous section) into bins called "visual words"; a collection of these bins is called a "visual vocabulary". In this work, we cluster the SIFT descriptors into visual words to represent an image in terms of their occurrence frequencies in the image. We use the *k-means* algorithm with cosine distance measure for clustering these descriptors.

While this representation throws away the spatial information for these patches, the performance of systems using this type of representation on classification and recognition tasks [14] are impressive. For every image, we compute the visual words to add the global image statistics as context.

3. Selective Hidden Random Field

A selective hidden random field (SHRF) is a variation of the hidden-state random field that has binary hidden variables to represent a selection criterion for the features computed on the observed data. In the context of sports classification, these binary hidden variables are used to infer whether a given image region is a part of the horizontal playing surface. Figure 5 shows the factor graph representation of this model. x_i represents the information extracted from the i^{th} super-pixel of an image using different experts such as edge detectors, appearance features, and visual vocabulary. h_i is a latent binary variable representing the surface orientation of the i^{th} super-pixel, and s represents the sports label for the image. The boxes represent the factors computed for the connected variable nodes; factors of the same color are of the same form and share parameters. Given a set of observations $\mathbf{x}$ and the parameters θ, the

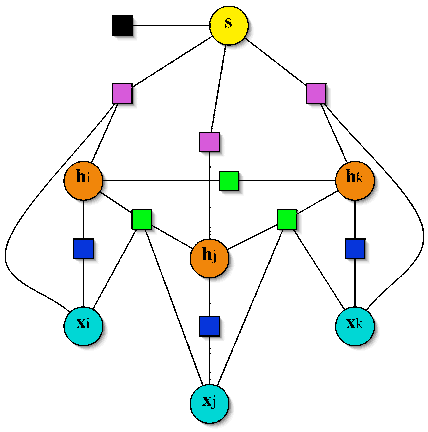

Figure 5. *Factor graph representation of Selective Hidden Random Field*. In this graphical model, the node s represents the sport label, and h_i and x_i represent the surface annotation and observed features for the i^{th} super-pixel, respectively. A blue box represents the local evidence for the surface annotation of a super-pixel, a green box denotes compatibility between the annotations for connected super-pixels, a purple box represents the contribution of a super-pixel towards the sport label for the image, and the black box represents the prior probabilities for different sporting events. Note that some of the edges (e.g., the edge connecting the green box between h_i and h_k with x_i) are omitted for clarity. *Best seen in color*.

conditional probability of a class (sports) label s is modeled as:

$$p(s|\mathbf{x},\theta) = \sum_{\mathbf{h}} p(s,\mathbf{h}|\mathbf{x},\theta) = \frac{\sum_{\mathbf{h}} \exp(\psi(s,\mathbf{h},\mathbf{x};\theta))}{\sum_s \sum_{\mathbf{h}} \exp(\psi(s,\mathbf{h},\mathbf{x};\theta))}, \tag{2}$$

where

$$\psi(s,\mathbf{h},\mathbf{x};\theta) = \sum_i \phi_i^b(x_i,h_i) + \sum_{i,j\in E} \phi_{ij}^g(h_i,h_j,x_i,x_j) + \sum_i \phi_i^p(s,h_i,x_i) + \phi^k(s). \tag{3}$$

In Equation 3, $(i,j) \in E$ implies that the i^{th} and j^{th} super-pixels share a boundary, and are similar in appearance. The factors ϕ have the following semantics:

- Horizontal plane hypothesis (**blue box**): For every super-pixel, we compute a value $f_1(x_i)$ reflective of its likelihood of having a horizontal orientation. In particular, we use the probability measure obtained from the surface classifier (Equation 1).

$$\phi^b(h_i,x_i) = \theta_b^T[\delta_{h_i=1} f_1(x_i) \;\; \delta_{h_i=0}(1-f_1(x_i))]^T. \tag{4}$$

- Neighborhood compatibility (**green box**): For every pair of connected[1] super-pixels, we compute a value $f_2(x_i,x_j)$ representing the similarity in appearance between the two. This helps us make consistent annotation for neighboring super-pixels. We use the cosine similarity between the feature vectors for the two super-pixels.

$$\phi^g(h_i,h_j,x_i,x_j) = \theta_g^T[\delta_{h_i=h_j} f_2(x_i,x_j) \;\; \delta_{h_i\neq h_j}(1-f_2(x_i,x_j))]^T \tag{5}$$

- Selection of super-pixels for features (**purple box**): The purpose of this potential is to select the super-pixels that are labeled as part of the playing surface, and use the features $f_4(x_i)$ computed on them to determine the sporting event.

$$\phi^p(\mathbf{s},h_i,x_i) = \mathbf{s}^T\theta_p^T[f_4(x_i)\delta_{h_i=1}]^T \tag{6}$$

- Prior information about the occurrence of sport (**black box**): This potential represents a prior information about the frequencies of different sporting events in a data set. For example, if the owner of the collection is passionate about soccer and tennis, but would rarely go to a baseball game, the corresponding prior would have more probability mass for soccer and tennis than baseball. For our data set, we are assuming a uniform prior.

$$\phi^k(s) = 1 \tag{7}$$

Given the parameters $\theta^T = [\theta_b^T \;\; \theta_g^T \;\; vec(\theta_p)^T]$ and observed image $\mathbf{x}$, the label is given by $argmax_s p(s|\mathbf{x},\theta)$ (specified in Equation 2). The parameters θ are estimated by maximizing

$$\mathcal{L}(\theta) = \sum_d \log p(s_d|\mathbf{x}_d,\theta) - \frac{1}{2\sigma^2}\|\theta\|^2, \tag{8}$$

where the first term is the conditional log-likelihood of the training data and the second term represents a Gaussian prior on the parameter values. We use a conjugate-gradient method for this optimization. Computing the gradient of $\mathcal{L}(\theta)$ and $p(s|\mathbf{x},\theta)$ involves the evaluation of quantities like the partition function $Z(s|\mathbf{x},\theta) = \sum_{\mathbf{h}} \exp(\psi(s,\mathbf{h},\mathbf{x};\theta))$. The presence of cycles in the connectivity graph prevents the use of exact methods for inference of these quantities. Thus, we resort to loopy belief propagation [19] for doing approximate inference in this graph.

[1] We apply different heuristics, such as sharing of boundary and threshold on similarity in appearance, to reduce the connectivity in graph, as opposed to a fully connected graph. This is done to ensure that the approximate inference algorithm converges.

Training set	Average class accuracy
original data	55.87 ± 3.70
data balanced by SMOTE [3]	89.58 ± 2.33

Table 2. In a given sports image, the regions that are part of the playing field are more in number than the regions that are not. Such imbalance in class frequencies often affect the performance of a classifier on the minority class if it is trained by minimizing the cumulative error. This table shows a big boost in average class accuracy for an SVM classifier for the playing surface when the class frequencies in the training data are balanced.

4. Data Set

We collected sports images from Flickr [1], an online photo management and sharing application that provides an API that supports multiple word queries for searching, listing, and downloading images. We used some sports team names and venues as queries to construct a data set of images of five popular sports: baseball, basketball, football, soccer, and tennis. We discarded the images without a significant view of the playing field, but did not restrict the images to include the entire view of the field. Some of the images include players, balls, or other objects, occluding the distinctive markings on the ground.

The data set contains 2449 images with roughly the same number of images for each sport. We split the data set into three parts: 50% for training, 25% for validation, and 25% for testing. The training and validation sets are used for tuning the parameters, and the test set is used for the final evaluation.

5. Experiments

We labeled all of the super-pixels of a few images (100 images each for training and testing) as horizontal or non-horizontal, to evaluate the SVM classifier for the playing surface (Section 2.3). Table 2 shows a significant improvement in the average class accuracy by removing the class imbalance in the training data.

To evaluate the proposed model, we consider the following approaches:

- **SVM**: We train a linear SVM for each of the three representations discussed in Section 2: visual vocabularies, line features, and super-pixel features, to obtain baseline accuracy values for comparison. Finally, we concatenate these representations together and train another SVM for it. For every choice of features, we compare the computation of features on (a) the entire image, and (b) the predicted horizontal surface. The obtained results verify our hypothesis about the utility of the playing surface.

- **CRF+SVM**: We build a CRF similar to our proposed model by modeling hidden h_i variables as target and removing the sports node s, to obtain the segmentation of the playing surface. We trained this model on the same examples used for training the SVM classifier for playing surface (Section 2.3). We observe an improvement of about four percent in the labeling accuracy by exploiting the consistency of labeling in the spatial neighborhood. Some example annotations are shown in the fifth row of Figure 6. This improvement in the segmentation performance motivates the probabilistic modeling of the consistencies across the neighbors.The super-pixels that are labeled as horizontal are used to compute the features that are fed into an SVM to classify the sporting event.

- **HRF**: We implemented a hidden random field similar to Quattoni et al.'s model [20] for gesture recognition with appropriate choice of potential functions. This method automatically learns the discriminative parts for different objects. This framework is not apt for our problem as the object of interest, i.e., the playing surface, is very similar in appearance across different sports, and the field markers – and not the image patches – are the discriminative components. The classification results obtained with this model are not competitive with other approaches discussed here, and the learned hidden variables do not correspond to the orientation of the image regions. Note that this implementation does not use the orientation labels obtained for a few images.

- **SHRF**: Our proposed model, selective hidden random field (see Section 3).

Rows 2-3 of Figure 6 show some of the features computed on the example images (shown in row 1) of different sports. The last four rows show a qualitative comparison of the predicted surface annotations obtained using the above-mentioned approaches. Table 3 compares the average class accuracies for 5-fold cross validation. The proposed model consistently outperforms the other approaches in average class accuracies and provides more accurate segmentation of the playing surface. Another key observation in the comparison between SHRF and CRF+SVM is that a joint segmentation-classification approach improves upon a sequential segmentation-classification approach.

6. Discussion

This paper makes three contributions: (a) It presents selective hidden random fields that simultaneously does the segmentation of the object of interest and uses it for classification. (b) It demonstrates the utility of exploiting domain-specific saliency for event classification, i.e., identifying

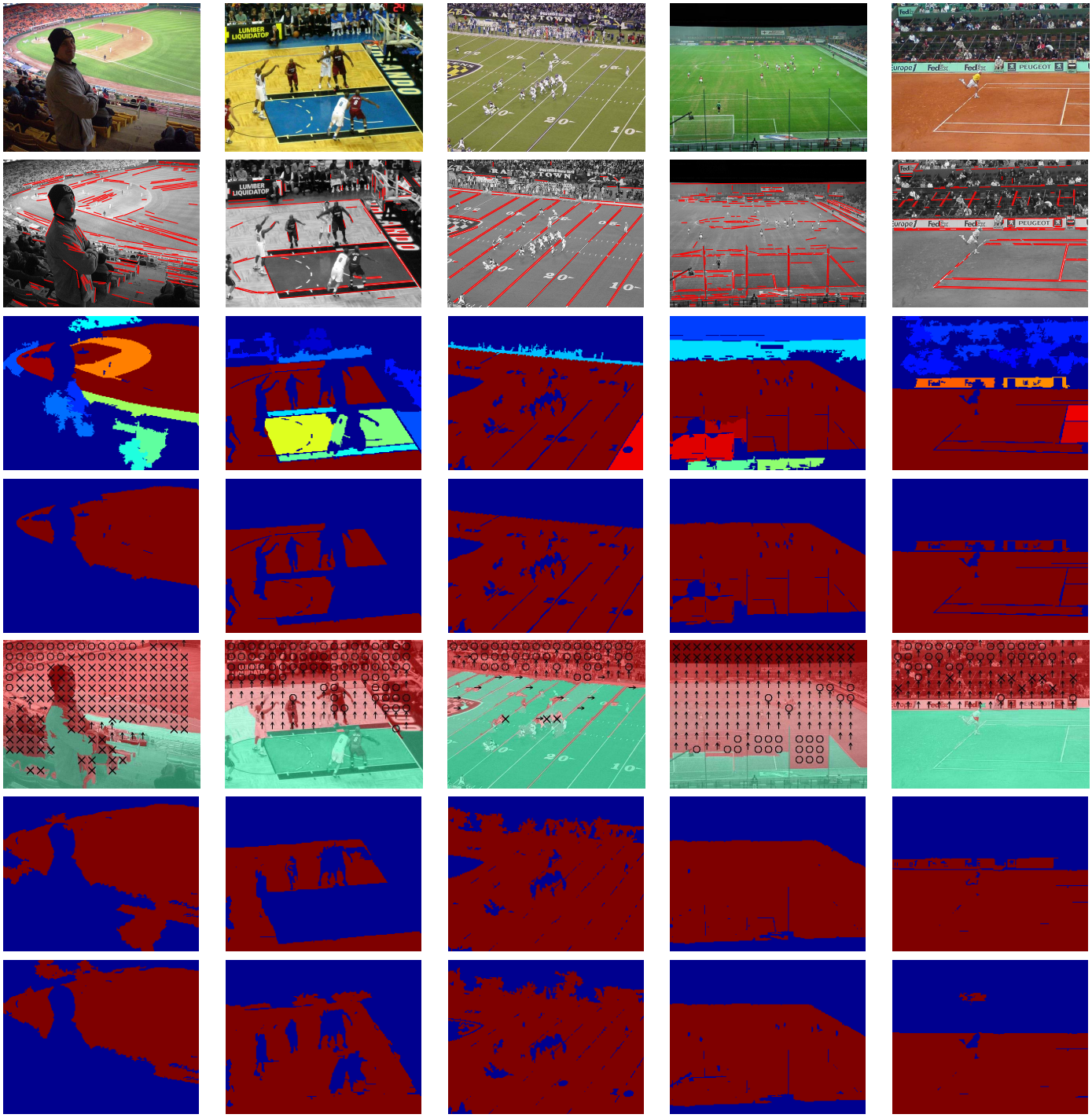

Figure 6. *Features and predicted surface orientation*: The top row shows example images of five different sports. The second row shows the line hypotheses generated for the example images, and the third row shows the probability of the orientation of different image regions to be horizontal; red represents horizontal and blue represents non-horizontal. The next four rows show the segmentation of the playing surface obtained using different approaches. The fifth row demonstrates that the features and statistics learned by Hoiem et al. [7] for street scenes, containing prominent horizontal and vertical surfaces in the same image, do not generalize to the sports images. Segmentations of the playing field obtained by an independent super-pixel model (SVM, row 4) are improved by including the dependencies on the neighbors (CRF, row 6) as the *holes* are filled and most of the field markers are correctly labeled. The segmentations obtained by our model (SHRF, row 7) are similar to those of CRF, and further improves the labeling for surfaces with very dissimilar appearance for neighboring super-pixels (see basketball image, column 2). The details of these methods are given in Section 5. *Best seen in color*.

Approach	Avg. class accuracy	
SVM		
	entire image	horizontal surface
Visual Vocabulary	41.56 ± 1.79	44.00 ± 7.02
Super-pixel features	52.07 ± 0.81	53.24 ± 1.01
Line features	50.83 ± 4.22	55.34 ± 3.16
All features	56.76 ± 3.40	59.92 ± 2.92
CRF + SVM	61.38 ± 2.01	
HRF	31.94 ± 4.19	
SHRF	$\mathbf{65.28 \pm 3.85}$	

Table 3. Average class accuracy for sports classification. The error terms correspond to 5-fold cross-validation experiments. This shows an improvement in performance by selecting the features on the horizontal surface for all the types of features. The difference is least significant in the case of visual vocabularies as the sampling rate is very low on the horizontal surfaces and they tend to represent the global image statistics. Among the type of features, line features gave the best performance. The results for HRF are not competitive with the other approaches.

and characterizing the playing surface to classify the sporting event. (c) It attempts to solve the recognition problem for popular sports that are more structurally challenging than the reported effort of Li et al. [14] (facilitated by the first two contributions).

Unlike many published data sets used for event classification, the events included in our data set have huge intra-class variations and inter-class similarities in the general scene statistics. In light of the difficulty of the data set, the improvement in performance – both accuracy and segmentation results – is remarkable, which justifies the use of the proposed model. Furthermore, we believe that the efficacy of the proposed model is not limited to this problem domain only. This model could be used for exploiting domain-specific contextual cues for other settings such as autonomous vehicle navigation, where an appropriate formulation of the selection criterion would replace the existing criterion for segmenting the playing surface.

References

[1] Flickr website: http://www.flickr.com.

[2] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. Learned-Miller, and D. A. Forsyth. Names and faces in the news. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 848–854, 2004.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002.

[4] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[5] A. Hanson and E. Riseman. VISIONS: A computer system for interpreting scenes. In *Computer Vision Systems*. Academic Press, 1978.

[6] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. In *IEEE International Conference on Computer Vision*, 2005.

[7] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[8] S. Hongeng, R. Nevatia, and F. Bremond. Video-based event recognition: activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding*, 96(2):129–162, November 2004.

[9] V. Jain, A. Ferencz, and E. Learned-Miller. Discriminative training of hyper-feature models for object identification. In *British Machine Vision Conference*, pages 357–366, 2006.

[10] V. Jain, E. Learned-Miller, and A. McCallum. People-LDA: Anchoring topics to people using face recognition. In *IEEE International Conference on Computer Vision*, 2007.

[11] A. Kapoor and J. Winn. Located hidden random fields: Learning discriminative parts for object detection. In *European Conference on Computer Vision*, 2006.

[12] J. Kittler, K. Messer, W. Christmas, B. Levienaise-Obadia, and D. Koubaroulis. Generation of semantic cues for sports video annotation. In *IEEE International Conference on Image Processing*, 2001.

[13] J. Kosecka and W. Zhang. Video compass. In *European Conference on Computer Vision*, 2002.

[14] L. J. Li and L. Fei-Fei. What, where and who? classifying event by scene and object recognition. In *IEEE International Conference on Computer Vision*, 2007.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[16] J. Matas, O. Chum, U. Martin, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *British Machine Vision Conference*, volume 1, pages 384–393, London, 2002.

[17] K. Messer, W. J. Christmas, and J. Kittler. Automatic sports classification. In *IEEE International Conference on Pattern Recognition*, 2002.

[18] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[19] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Conference on Uncertainty in Artificial Intelligence*, 1999.

[20] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *Advances in Neural Information Processing Systems*. MIT Press, 2005.

[21] S. B. Wang, A. Quattoni, L.-P. Morency, and D. Demirdjian. Hidden conditional random fields for gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.