Recognizing Plays in American Football Videos

Behjat Siddiquie
University of Maryland
College Park, MD 20742
behjat@cs.umd.edu

Yaser Yacoob
University of Maryland
College Park, MD 20742
yaser@umiacs.umd.edu

Larry S. Davis
University of Maryland
College Park, MD 20742
lsd@cs.umd.edu

Abstract

We address the problem of recognizing American Football plays in video. In contrast to recent work on activity recognition this is a much more challenging problem as it involves the actions of multiple players. We propose a method which builds on recent advances in activity recognition, such as using shape and motion based spatio-temporal features and building a space-time representation of the video. Furthermore, we use Multiple Kernel Learning to effectively combine different features. We also propose an extension to the Multiple Kernel Learning method, which optimizes the number of kernels selected, thereby improving efficiency. We demonstrate our approach on a challenging dataset consisting of a variety of football plays and obtain promising results, in the process we also discover some interesting aspects of different types of football plays.

1. Introduction

Recognizing human actions in videos is an important research area in the field of computer vision. Activity recognition has a wide range of applications such as video indexing and retrieval, video surveillance and sports video analysis. As a result it has received considerable interest over the last few years. A common approach for action recognition involves extracting local features from videos, representing the video in terms of these local features and, finally classification.

Much of the work on action recognition has focussed on two datasets, the KTH and Weizmann datasets [23, 9, 24, 21]. These datasets contain scenarios where actions are performed by a single person against a homogeneous background, viewed with a static camera. Though there has also been some work on recognizing human actions in more realistic settings such as movies or videos from YouTube [13, 14, 18], these approaches are still limited to recognizing activities where at most one or two people are involved. In this paper we investigate the problem of recognizing activities in which multiple people are involved. Specifically, we focus on identifying the type of play in American Football videos. Currently, football coaches spend significant time studying football videos to gauge the strengths and weakness of their own team as well as future opponents. Our research has potential applications in supporting query and retrieval on football videos which would considerably reduce the amount of manual work currently involved in the analysis of these videos.

We classify a football play into one of seven play types. Compared to single person action recognition problems, this is a much more challenging problem as the type of a play is influenced by the actions of multiple people and the movement of the ball. Unlike previous approaches dealing with multi-agent activities [8, 17], we do not track the motion of individual players nor do we employ any high level reasoning. Instead, we use only low-level features and rely on a machine learning based feature selection method to identify the discriminative features for classification of the football plays.

We utilize spatio-temporal features to represent both shape and motion, employing spatio-temporal pyramids to build a space-time representation and use Multiple Kernel Learning [12] to effectively combine different features. We also propose an extension to an existing Multiple Kernel Learning method [22], where we reformulate the problem to sparsify the weights assigned to the kernels, thereby improving its efficiency. Additionally, we also propose recognizing the play type, as early in the execution of the play as possible.

The next section describes related work. Section 3 describes our method. Section 4 describes the experiments and results, which is followed by the conclusion.

2. Related Work

There has been a large body of work on action recognition. Though most of it has focussed on recognizing actions of a single human in relatively controlled settings [4, 23, 9, 5, 24, 21], there has also been work on recognizing human activities in realistic situations [13, 14, 18]. Two of the key insights from these efforts have been to use multiple heterogeneous features and to combine them using an effective classification model to integrate their complementary information.

A variety of local feature descriptors have been developed for describing actions in videos. Spatio-temporal features, such as histograms of space time gradients, have been

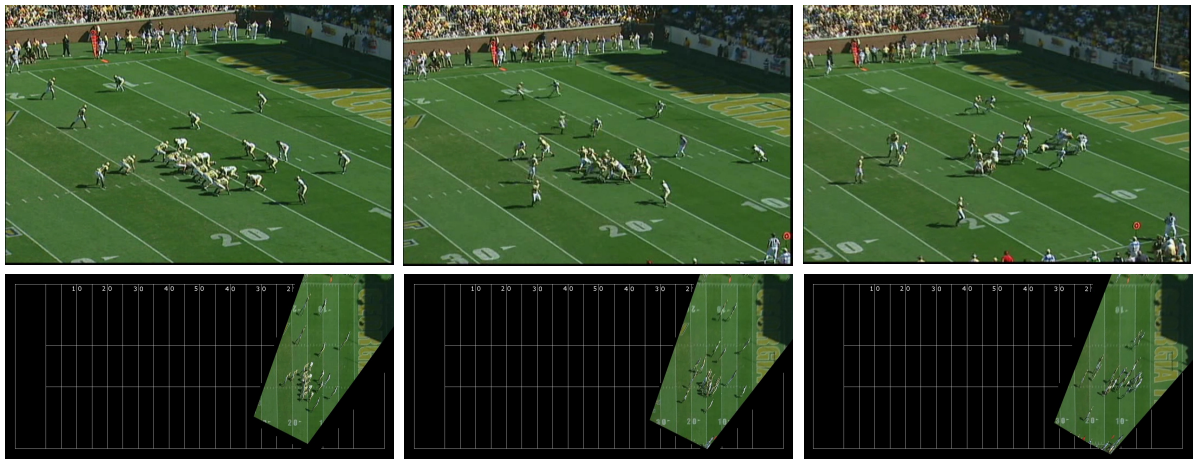


Figure 1. **Preprocessing:** frames from video(top row) are stabilized and then warped to a virtual football field(bottom row). The feature extraction is done on the warped video.

shown to be effective for action recognition [13, 14, 4, 24]. Optical flow based features are used to represent the local motion patterns in video sequences. In [13, 14], histograms of optical flow have been used, while [23, 9] employed a biologically inspired model to encode dense flow information. Static features, that capture the human pose at each time instant have also been used [18]. Typically, the Bag-of-Features model is used to represent the occurrence of these local features, by constructing a vocabulary of visual words for quantizing them. This Bag-of-Features model can be further augmented by encoding the spatio-temporal relationships between the local features. Methods such as spatio-temporal pyramid matching [13], which provide a coarse description of the spatio-temporal layout of the video, have been shown to improve recognition.

Combing multiple features has been shown to be effective for improving action recognition performance. In general, appearance and motion based features contain complementary information and several techniques have been proposed for combining them. Fanti et al. [5], proposed a mixture of both static and dynamic features for action recognition. Schindler et al. [23] and Jhuang et al. [9], also combine shape and motion features to improve recognition accuracy. In [14, 18], AdaBoost is used to combine multiple features for recognizing actions in realistic settings. Laptev et al. [13], combine kernels computed from HOG and HOF features over different spatio-temporal grids.

A number of techniques have been proposed to learn the optimal combination of a set of kernels, computed from multiple features, for SVM-based classification. Lanckriet et al. [12], introduced the Multiple Kernel Learning(MKL) method to learn a set of linear combination weights for combining multiple kernels and the SVM parameters for the resulting kernel simultaneously in a semidefinite programming framework. Rakotomamonjy et al. [22], substantially increased the efficiency of MKL by reformulating it using a 2-norm regularization as a convex optimization problem. Additionally, their formulation utilized an l_1-norm constraint for favoring sparse kernel combinations. Varma and Ray [26], combined multiple features using MKL and showed a considerable increase in the classification accuracy on several object recognition datasets. In [6], a fast multiple learning method was proposed for learning kernel weights over the codebook of visual words.

Activity recognition for multiple people has focussed on analysis of sports video, [7, 15, 19]. Other work has dealt with recognizing abnormal activities in video [28, 10, 1]. Classification of American Football plays has been proposed in [25, 8, 17]. In [25], a non-stationary kernel hidden markov model is used for recognizing football plays. Intille and Bobick [8], recognized football plays by using Bayesian networks for modeling the interactions between the players. In [17], multiple person activity is modeled as a four-dimensional object-time interaction tensor that is reduced to a discriminative temporal interaction matrix, which is then classified using a probabilistic framework. However, both [8, 17], have been demonstrated on data with human annotated player trajectories and player roles, for which the amount of manual work required is significant. It is unclear if they will maintain their performance when these tasks are automated, as tracking and identity maintenance would be extremely hard in football games due to occlusions and players wearing similar clothing. In contrast, in our method the only annotation required is the class labels for the training set.

3. Classification for Play Recognition

We evaluate our method on a dataset consisting of 78 videos of play instances collected from NCAA football games. Previous work on this dataset has been presented in [25, 17]. The videos are similar to typically broadcast sports videos. They taken from a camera located in the stadium, which pans to keep the players within the field of view and also frequently zooms in and out, during the course of the play. These videos cannot be directly processed as the camera motion severely degrades the quality

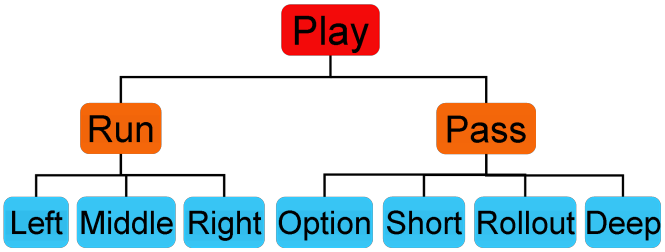


Figure 2. The class hierarchy of the football plays.

of the spatio-temporal features. Therefore, we first stabilize each video to negate the effect of the camera motion and then warp the videos to a virtual football field to ensure that all the videos have a consistent spatial representation. Figure 1 illustrates this process for frames from one of the videos of our dataset. The plays consist of seven different classes - *left-run, middle-run, right-run, short-pass, option-pass, rollout-pass, deep-pass*. The hierarchy of the play classes is shown in Figure 2.

3.1. Feature Extraction

Existing works on action recognition [4, 13, 24], use local spatio-temporal features extracted from interest points. Local spatio-temporal features have been shown to perform well for action recognition and they are known to be robust to background clutter and scale and illumination changes. Typically, a set of spatio-temporal interest points are first identified using interest point detectors, and the video is then represented by spatio-temporal features extracted from 3D space-time blocks centered at the interest points.

We adopt a similar approach for feature extraction. However, instead of sampling features from sparse interest points, we extract them from a dense 3D grid. Due to small errors in the stabilization, there is some jitter introduced into the videos and hence dense features prove to be more robust. The size of each 3D space-time block is (V_x, V_y, V_t) and they are sampled on a dense 3D grid $(V_x/2, V_y/2, V_t/2)$ apart, providing a 50% overlap between neighboring blocks. As all the videos are warped to the same virtual field, they have a consistent scale and hence we use a fixed scale for the 3D blocks. To characterize local appearance and shape, we use histograms of oriented gradients(HOG). Each 3D block is partitioned into a grid with $n_x \times n_y \times n_t$ spatio-temporal sub-blocks, and HOG features are computed for all the sub-blocks and are normalized and concatenated. To represent local motion patterns, Histograms of Optical Flow(HOF) features are extracted from each 3D spatio-temporal block in a similar manner. We set $V_x, V_y = 32$ pixels, chosen so as to roughly correspond to the size of a human in the videos and $V_t = 8$ frames.

3.2. Bag-of-Features

We combine feature vectors from all the videos and cluster them using *k-means* to create a Bag-of-Features vocabulary, which consists of the cluster centers. Separate vocabularies, consisting of 150 visual words each, are created for the HOG and HOF features. This quantizes the feature space and enables representation of each video in terms of the vocabulary. Given a test video, the local features extracted from the space-time blocks are assigned to the closest visual words by Euclidean matching. The occurrences of each of the visual words can then be used to represent the video.

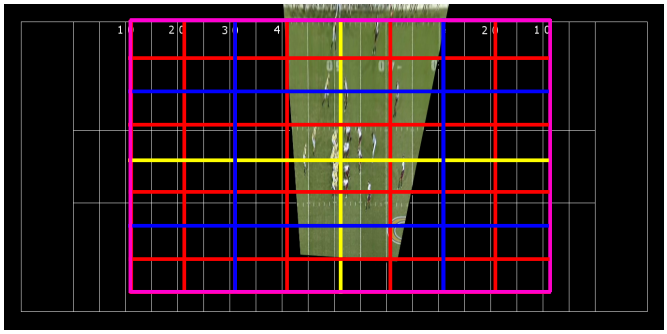


Figure 3. The spatial grid used for spatio-temporal pyramid match. The different colors denote the pyramidal blocks at different levels. Note that there are also 3 levels in the temporal dimension(not shown).

3.3. Spatio-Temporal Pyramid Match

The Bag-of-Features representation does not capture any information regarding the spatial and temporal distribution of the visual words. To overcome this problem Lazebnik et al. [16] proposed the spatial pyramid match method, which encodes the spatial geometry of the scene at a coarse level and has proven to be very effective for object and scene recognition. Given the spatial distributions of a visual word in two images, it computes a measure of the spatial correlation between the two distributions by repeatedly subdividing the image space into smaller sub-images and computing the correspondence between the distributions in all those sub-images. This method can also be applied for action recognition in video by extending it to the temporal dimension [13]. Using the Spatio-Temporal Pyramid Match a pair of videos will be highly similar, if they have similar space-time distributions of visual words. For a pair of videos x_i and x_j and a given visual word w_l, the Spatio-temporal pyramid match computes a similarity kernel $\Phi_l(x_i, x_j)$, between them. Taking into account all visual words, this similarity measure can be written as:

$$\Phi(x_i, x_j) = \sum_k \Phi_l(x_i, x_j) \tag{1}$$

Here all the visual words are given equal weights. It has been shown that the resultant Φ forms a mercer kernel and hence can be used for classification using kernel based methods, such as SVMs. To ensure that Spatio-Temporal Pyramid match captures semantically relevant information, instead of pyramidally subdividing the video equally in all

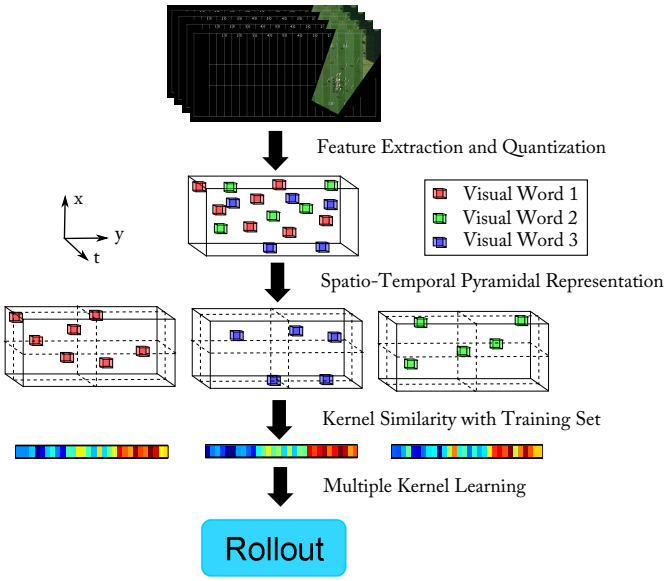


Figure 4. **Classification Procedure:** In the first stage the features are extracted and quantized in terms of the visual vocabulary. Next the Spatio-Temporal Pyramid Match is used to obtain the kernel similarities to the training set with respect to all the visual words. Finally Multiple Kernel Learning is used to combine the kernel distances and obtain the class label.

dimensions, we position the pyramidal grid such that it is spatially centered at the point in the field where the snap takes place(Fig. 3). This ensures that the regions behind and in front of the scrimmage line[1] and the areas to the left and the right of the point of snap[1] are well demarcated. We use a pyramidal grid consisting of 4 levels in the spatial dimension(Fig. 3) and 3 levels in the temporal dimension, resulting in a $8 \times 8 \times 4$ spatio-temporal grid at the finest level. These parameters have been empirically determined.

3.4. Multiple Kernel Learning

Spatio-Temporal Pyramid match combines information from multiple features channels and also from the different visual words of each feature channel. Instead of weighting all the kernels uniformly, as in Equn. 1, one would like to find a set of optimum weights d_l that maximize the discriminative power of the similarity metric Φ with respect to the classification problem. The optimal kernel can be computed as a convex combination of the basis kernels:

$$\Phi(x_i, x_j) = \sum_{k=1}^{K} d_k \Phi_k(x_i, x_j), \quad \sum_{k=1}^{K} d_k = 1 \quad , d_k \geq 0 \qquad (2)$$

where x_i are the data samples(videos), $\phi_k(x_i, x_j)$ is the kth kernel and d_k are the weights given to each visual word (kernel). Learning the classifier model parameters and the kernel combination weights in a single optimization problem is known as the Multiple Kernel Learning(MKL) problem [12]. There have been a number of formulations for the MKL problem, as noted in Section 2. The MKL optimization equation is given by:

$$\begin{aligned} \min \quad & \sum_k \frac{1}{d_k} w_k w_k^T + C \sum_i \xi_i \\ \text{such that} \quad & y_i \sum_k \phi_k(x_i) + y_i b \geq 1 - \xi_i \ \ \forall i \\ & \xi_i \geq 0 \ \ \forall i, \ \ d_k \geq 0 \ \forall k, \ \ \sum_k d_k = 1 \end{aligned} \qquad (3)$$

where b is the bias, ξ_i is the slack afforded to each data point and C is the regularization parameter. The solution to the above MKL formulation is based on an iterative two stage method [22]. In the first stage, the kernel weights d_k are fixed and the above equation reduces to the standard SVM optimization problem, which can be solved with any SVM solver. In the second stage, the SVM parameters are fixed and a projected gradient descent is performed to minimize the objective function with respect to the kernel weights. The two stages are repeated until convergence. MKL results in learning higher weights for the more discriminative kernels and assigns low weights to the redundant ones, leading to a significant improvement in classification performance. Unlike [26], where Multiple Kernel Learning has been used for combining kernels computed over multiple features and spatial pyramidal levels, we learn a linear combination of kernels computed over the visual words from different feature channels. Figure 4 illustrates our entire recognition framework.

3.5. Sparse Multiple Kernel Learning

The Multiple Kernel Learning formulation in equation 3, imposes an l_1 norm constraint on the kernel weights d_k. This constraint, apart from ensuring that $\sum d_k = 1$, also restricts the search space and results in an efficient solution. In the MKL formulation of [11], an l_2 norm has been used for constraining the kernel weights. The comparison of the effect of l_1-vs-l_2 norm constraints on MKL has been previously studied [27]. Typically, the l_1 norm constraint ensures a sparser solution and is more effective in the presence of noisy kernels. While on the other hand, l_2 norm based approaches, though less sparse and susceptible to noise, are known to more effectively combine kernels containing orthogonal information. In our case, kernels are computed from each visual word using spatio-temporal pyramid match. The quantization of the feature descriptor in terms of visual words usually leads to lots of redundant visual words and a very small number of discriminative visual words [2]. Hence, we argue that, a sparse solution to the kernel weights is better suited to our approach. We propose a new formulation of MKL with an approximate l_0 norm constraint, which assigns non-zero weights to only the

[1]The scrimmage line refers to the imaginary line parallel to the yard lines where the play starts. The snap refers to the act of starting the play and the point of snap refers to the position of the ball at the time of the start of play.

most discriminative visual words(kernels) leading to a very sparse and hence more computationally efficient solution.

The l_0 norm constraint on the kernel weights can be written as: $\|d_k\|_0 \leq r$. This corresponds to imposing a limit r, on the total number of kernels with non-zero weights, and hence can be used for restricting the number of kernels selected by MKL. However, imposing the l_0 norm constraint is in general NP-hard and this is typically overcome by an approximation to the l_0 norm. In [3], a method has been proposed for approximating the l_0 condition by imposing a reweighted l_1 norm penalty and has been applied for signal reconstruction. We adapt this method for the purpose of sparsifying the MKL solution. The MKL objective function(Eq. 3) is modified by adding an extra regularization term to get:

$$\begin{aligned} \min \quad & \sum_k \frac{1}{d_k} w_k w_k^T + C \sum_i \xi_i + \beta \sum_k t_k d_k \\ \text{such that} \quad & y_i \sum_k \phi_k(x_i) + y_i b \geq 1 - \xi_i \ \ \forall i \\ & \xi_i \geq 0 \ \ \forall i, \ \ d_k \geq 0 \, \forall k, \ \ \sum_k d_k = 1 \end{aligned} \tag{4}$$

Where β is the regularization parameter and the t_ks are "adaptive weights" assigned to the kernel weights. This formulation can be solved by modifying the MKL solution, described earlier. The regularization term is an adaptively weighted l_1 norm penalty on the kernel weights. At the end of each iteration of the MKL optimization algorithm, the weights t_k are reset as $t_k^{n+1} = \frac{1}{d_k^n + \epsilon}$, to adaptively approximate the l_0 norm condition. This ensures that, t_k is low when the corresponding kernel weight is high and vice versa, which penalizes the kernels with low weights and drives them towards zero, thereby ensuring a sparse solution. The parameter ϵ can be modified to control the sparsity. This sparse MKL method can be used for selecting a very small subset of discriminative visual words, while excluding the large majority of redundant visual words present in the vocabulary. This significantly reduces the number of kernel computations required by Spatio-Temporal Pyramid Match, leading to improved efficiency. In subsection 4.5, we demonstrate the effectiveness of our sparse MKL method.

4. Experiments and Results

4.1. Classification Results

Our dataset consists of seven classes, which have a well defined semantic hierarchy(Fig. 2). It has been previously shown that hierarchical classification in a semantically relevant manner can increase performance [20]. Hence we adopt a simple hierarchical classification scheme, consisting of two levels and three base classifiers. In the first level, we train a classifier to classify a play as a run or pass. The second level consists of two base classifiers, one each for classifying the run and pass plays into their respective subclasses. We also compare the performance of this hierarchical classifier to a simple multi-class classifier. We now describe the performance of base classifiers and the combined classifier. All the experiments consisted of 5-fold cross validation, repeated 50 times with different randomly selected training and test videos, and the average per-class recognition rate was recorded.

	Pass	Run
Pass	**87.1**	12.9
Run	8.3	**91.7**

Table 1. Confusion matrix for the run-vs-pass classification.

	Left-Run	Middle-Run	Right-Run
Left-Run	**89.9**	10.0	0.1
Middle-Run	23.9	**68.1**	8.0
Right-Run	0.0	14.4	**85.6**

Table 2. Confusion matrix for classification of the run plays.

	Short	Option	Rollout	Deep
Short-Pass	**81.4**	10.0	8.3	0.3
Option-Pass	46.0	**42.8**	10.8	0.4
Rollout-Pass	6.8	12.8	**69.2**	11.2
Deep-Pass	1.8	9.0	10.7	**78.5**

Table 3. Confusion matrix for classification of the pass plays.

Run-vs-Pass: For Run-vs-Pass classification at the top level using the base classifier, the recognition rate is 89.4%. The confusion matrix is shown in Table 1. In the next two subsections we analyze the affect of different features on the performance of this classifier.

Run Plays: At the next level, for the classification of run plays into *left-run*, *middle-run* and *right-run*, we obtain a classification accuracy of 81.2%. From the confusion matrix(Table 2), it is clear that most of the error occurs between left-run/middle-run and middle-run/right-run. This is because of the lack of a hard separating boundary between those two pairs of classes.

Pass Plays: For the classification of the pass plays into *short-pass*, *option-pass*, *rollout-pass* and *deep-pass* plays, the recognition rate is 70.3%. The confusion matrix is shown in Table 3. The recognition rate is lower compared to the classification of the run plays because the distinctions between these classes are subtle and often result from the actions of a particular player at a given instant of time. The deep-pass class has a relatively high recognition as it has a larger spatial extent on account of the long passes and hence is easier to distinguish.

Overall Results: When using the hierarchical classifier, we obtain an average classification accuracy of 71.9% for all the seven classes. The confusion matrix is shown in Figure 5. On performing a flat classification, the recognition

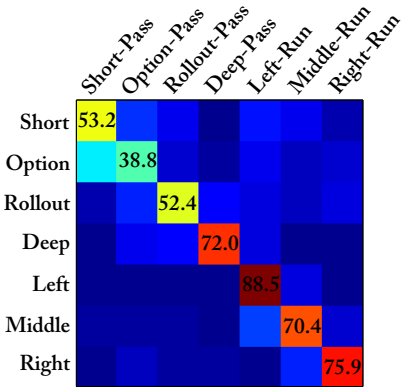


Figure 5. Confusion matrix for the overall classification.

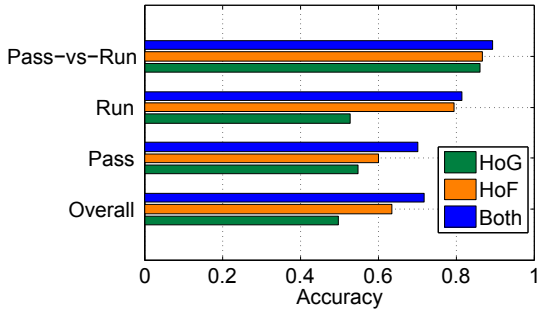


Figure 6. Performance of each of the base classifiers and the overall classifiers using HoG and HoF features separately and when combined together.

rate is 67.1%. It is clear that the hierarchical classification results in an improvement in the results. These results are substantially better than the 58% recognition rate obtained by [25] on the same dataset. Our results also compare favorably to Li et al. [17] who obtain a recognition rate of 87.9% on an easier subset of this data consisting of just 5 classes, moreover they have also used human annotations for the player trajectories and the player roles.

4.2. Features

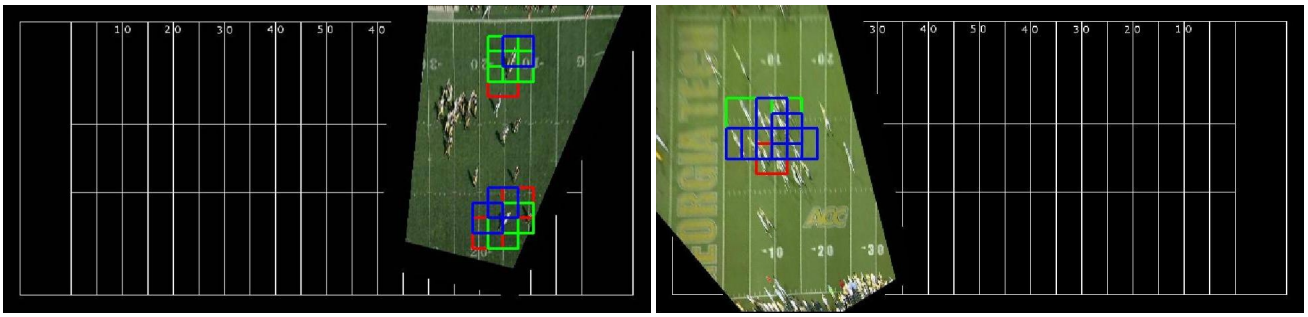
The Histograms of Oriented Gradients(HOG) based visual words can encode local actions such as standing, walking and running and can also differentiate between the actions of an individual player and the actions of a group of players together. These characteristics make HOG features useful for distinguishing between pass and run plays. On the other hand, the Histogram of Optical Flow(HOF) based features capture direction of local motion and are useful for discriminating between the left-middle-right run plays. To evaluate the performance of the HOG and HOF features, individually as well as when combined, we use MKL for learning a combination of kernels computed from the HOG visual words, the HOF visual words and all the visual words together. Figure 6 shows the performance of each of the base classifiers and the hierarchical classifier using HOG, HOF and both of them combined. It is clear that combining the these two features improves results. This is in agreement with [13, 18, 23], who suggested that shape and motion based features contain complementary information and combining them improves the action recognition accuracy.

4.3. Discriminative Visual Words

To gain further insight into the classification, we analyzed the weights assigned to different kernels by MKL. Discriminating kernels are assigned high weights by MKL and hence they represent the set of distinguishing features(visual words). For the run-vs-pass classification, the discriminative features corresponded to one or two players running along the field(Fig. 7(a)). In pass plays, players from the attacking team frequently run along the field to receive a pass and players from the defending team run along with them intending to prevent them from receiving the pass. This kind of activity rarely happens in run plays and is therefore an important distinguishing characteristic between run and pass plays. Other discriminative features corresponded to groups of players running together in the same direction(Fig. 7(b)), which occurs in run plays but is uncommon in pass plays. The spatial density of the occurrence of the top five discriminative visual words in the pass and run plays is shown in Fig. 8. In case of the pass plays it is interesting to see that the density of the distribution correlates very well with the motions of the receivers and it is clear that the spatial distribution of the visual words is also an important discriminating factor between the pass and run plays, hence validating our incorporation of the Spatio-Temporal pyramidal framework. These findings show that our recognition method, without any kind of domain knowledge, is able to identify such high level characteristic features of the data and incorporates them into the learning of the classifier.

4.4. Predicting the Play in Advance

Schindler et al. [23], have shown that very few frames are sufficient for accurately recognizing basic human actions. In our case, since there are complex interactions between multiple players involved, it might take even a human observer a few seconds to discern the type of play. Nevertheless, we evaluate our approach for recognizing a play, given only its first few frames, thereby predicting the outcome of the play in advance. Our experimental approach is as follows. We use only the features extracted from the first f frames of the training videos to build a classifier C_f, which is evaluated by testing it on the first f frames of the test videos. The results, shown in Figure 9, plot the overall accuracy as well as accuracy of the individual classes, of C_f versus the number of frames observed, f. In all the cases, the initial results are close to random, but as time progresses and more information becomes available, the recognition accuracy rapidly increases. For the run-vs-pass classification(Fig. 9(a)), we achieve a recognition rate of 80% by the 50th frame, even though the average length of a video is about 111 frames. Similarly for classification of the run plays we have an accuracy of about 70% by the 70th



(a) Pass Play (b) Run Play

Figure 7. The locations of the discriminative visual words. The squares denote the spatial locations of the 3D patches from which the discriminative visual words were extracted at a particular time instant. Each color denotes a different visual word. In a Pass Play(a), the discriminative visual words correspond to receivers running to receive a pass, while in a Run Play(b), they represent a group of players running together.

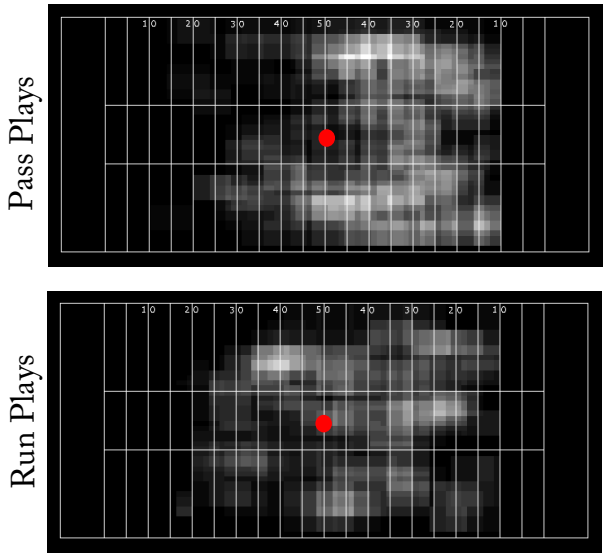
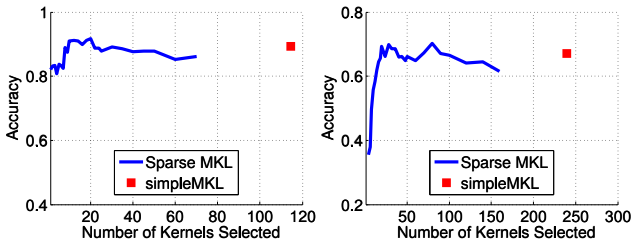


Figure 8. The spatial distribution of the top five discriminative visual words in pass plays(top) and run plays(bottom). All the plays have been centered so that their scrimmage line is aligned with the 50 yard line and that the snap takes place on the red dot in the center. In case of the pass plays, the distribution has a well defined spatial pattern.

frame(Fig. 9(b)). In the pass play classification(Fig. 9(c)) as well as the overall classification(Fig. 9(d)), the recognition rate reaches the peak level within a very few frames. These results show that our approach can predict the play type well in advance of the end of the play.

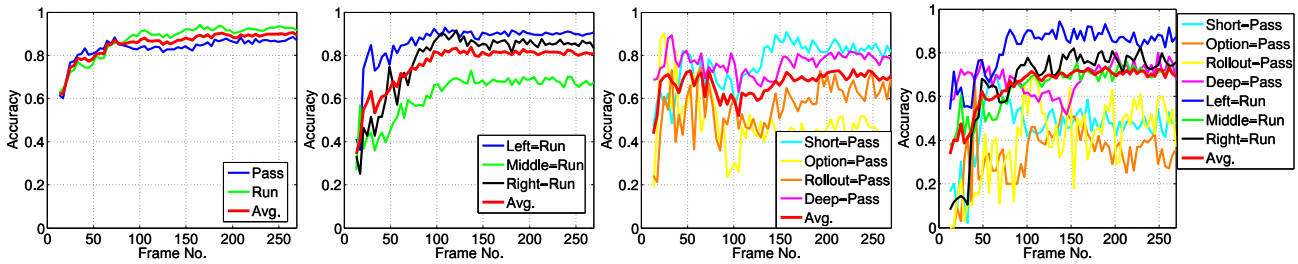
4.5. Sparse Kernel Learning

We study the affect of sparsity in kernel selection on the classification performance, using the sparse Multiple Kernel Learning technique described in 3.5. We evaluate our method on the Run-vs-Pass classification which is a binary classification problem as well as on the overall classification which is a multi-class problem. The sparsity can be controlled by varying the parameter ϵ, decreasing ϵ increases sparsity and vice-versa. In case of the Run-vs-Pass classification, we vary ϵ, keeping the other parameters fixed and plot(Figure 10(a)) the recognition accuracy versus the number of visual words(kernels) with non-zero weights. We compare the performance with simpleMKL [22] which uses the l_1 norm constraint. It is clear that even with a very small number of visual words it is possible to achieve high classification accuracy. In case of the multi-class problem, the individual one-vs-all classifiers independently select different sets of kernels and hence all the selected kernels are not utilized by all the individual classifiers. To overcome this problem, we use a two stage approach, the sparse Multiple Kernel Learning method is used only for selecting the discriminative kernels and then simpleMKL is used to perform Multiple Kernel Learning on only the selected kernels. Figure 10(b) plots the accuracy vs the number of kernels selected, and here again very few kernels are sufficient to achieve a high recognition rate. The results support our hypothesis that a very few carefully selected visual words are sufficient for good discrimination and this can be exploited by sparse approaches for improving efficiency during the classification phase by reducing the number of kernel computations required.



(a) Run-vs-Pass (b) Overall

Figure 10. The performance of the sparse Multiple Kernel Learning method as a function of the number of kernels selected. The number of kernels(visual words) required for effective classification can be reduced by a factor of about 10 compared to simpleMKL [22]. In case of the overall classification(b), a flat one-vs-all classifier is used.



(a) Run-vs-Pass (b) Run Plays (c) Pass Plays (d) All Plays

Figure 9. The recognition results of the base classifiers and the overall classifier as a function of the number of frames.

5. Summary

We have proposed a learning based approach for recognizing plays in American football games. Our method, based on discriminative feature selection framework, is able to identify high level properties of the data and utilize them for learning the classifier. We have demonstrated the effectiveness of our method on a challenging dataset of football videos. We have also proposed a sparse Multiple Kernel Learning method and shown that one can achieve high classification accuracy with a small number of suitably chosen visual words. We are currently looking into the possibility of generalizing our method for other sports and common activities.

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on PAMI*, 2008. 2

[2] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. *CVPR*, 2008. 4

[3] E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted l_1 minimization. *J. Fourier Anal. Appl.*, 2007. 5

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. *ICCV VS-PETS*, 2005. 1, 2, 3

[5] C. Fanti, L. Zelnik-Manor, and P. Perona. Hybrid models for human motion recognition. *CVPR*, 2005. 1, 2

[6] J. He, S.-F. Chang, and L. Xie. Fast kernel learning for spatial pyramid matching. *CVPR*, 2008. 2

[7] C. Huang, H. Shih, and C. Chao. Semantic analysis of soccer video using dynamic bayesian network. *IEEE Transactions on Multimedia*, 2006. 2

[8] S. Intille and A. Bobick. Recognizing planned, multiperson action. *CVIU*, 2001. 1, 2

[9] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. *ICCV*, 2007. 1, 2

[10] J. Kim and K. Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. *CVPR*, 2009. 2

[11] M. Kloft, U. Brefeld, P. Laskov, , and S. Sonnenburg. Non-sparse multiple kernel learning. *NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, 2008. 4

[12] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 2004. 1, 2, 4

[13] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *CVPR*, 2008. 1, 2, 3, 6

[14] I. Laptev and P. Perez. Retrieving actions in movies. *ICCV*, 2007. 1, 2

[15] M. Lazarescu and S. Venkatesh. Using camera motion to identify different types of american football plays. *ICME*, 2003. 2

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *CVPR*, 2006. 3

[17] R. Li, R. Chellappa, and S. K. Zhou. Learning multi-modal densities on discriminative temporal interaction manifold for group activity recognition. *CVPR*, 2009. 1, 2, 6

[18] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. *CVPR*, 2009. 1, 2, 6

[19] T. Liu, W. Ma, and H. Zhang. Effective feature extraction for play detection in american football video. *MMM*, 2005. 2

[20] M. Marszałek and C. Schmid. Semantic hierarchies for visual object recognition. *CVPR*, 2007. 5

[21] J. Niebles and L. Fei-Fei. A heirarchical model of shape and appearance for human action classification. *CVPR*, 2007. 1

[22] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. More efficiency in multiple kernel learning. *ICML*, 2007. 1, 2, 4, 7

[23] K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? *CVPR*, 2008. 1, 2, 6

[24] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. *ACM-Multimedia*, 2007. 1, 2, 3

[25] E. Swears and A. Hoogs. Learning and recognizing american football plays. *Snowbird Learning Workshop*, 2009. 2, 6

[26] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. *ICCV*, 2007. 2, 4

[27] F. Yan, K. Mikolajczyk, J. Kittler, and M. Tahir. A comparison of l_1 norm and l_2 norm multiple kernel svms in image and video classification. *CBMI*, 2009. 4

[28] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. *CVPR*, 2004. 2