This CVPR2013 Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Automatic Recognition of Offensive Team Formation in American Football Plays

Indriyati Atmosukarto
ADSC, Singapore
indria@adsc.com.sg

Bernard Ghanem
KAUST, Saudi Arabia
bernard.ghanem@kaust.edu.sa

Shaunak Ahuja
ADSC, Singapore
shaunak.ahuja@adsc.com.sg

Karthik Muthuswamy
NTU, Singapore
KART0028@e.ntu.edu.sg

Narendra Ahuja
UIUC, USA
n-ahuja@illinois.edu

Abstract

Compared to security surveillance and military applications, where automated action analysis is prevalent, the sports domain is extremely under-served. Most existing software packages for sports video analysis require manual annotation of important events in the video. American football is the most popular sport in the United States, however most game analysis is still done manually. Line of scrimmage and offensive team formation recognition are two statistics that must be tagged by American Football coaches when watching and evaluating past play video clips, a process which takes many man hours per week. These two statistics are also the building blocks for more high-level analysis such as play strategy inference and automatic statistic generation. In this paper, we propose a novel framework where given an American football play clip, we automatically identify the video frame in which the offensive team lines in formation (formation frame), the line of scrimmage for that play, and the type of player formation the offensive team takes on. The proposed framework achieves 95% accuracy in detecting the formation frame, 98% accuracy in detecting the line of scrimmage, and up to 67% accuracy in classifying the offensive team's formation. To validate our framework, we compiled a large dataset comprising more than 800 play-clips of standard and high definition resolution from real-world football games. This dataset will be made publicly available for future comparison.

1. Introduction

Sports coaches and analysts often analyze large collections of video to extract patterns and develop strategies in their respective sport. Before any extensive video analysis can be conducted, these sports experts must annotate the video data to provide context to the video clips. It is quite natural for humans to commit errors during annotation due to the high visual similarity, repetitive nature, and sheer volume of video data. If they have the budget, which few do, American Football (hereafter referred to as 'football') teams employ highly experienced staff (video coaches) specifically for video annotation. Even dedicated video coaches are prone to error when manually annotating football clips and extracting important details from them.

In football, there are two general football play types: offense/defense (o/d) and special teams. In this paper, we concentrate on two important details which must be recorded in every o/d play, namely the line of scrimmage and offensive formation type. The line of scrimmage is an imaginary line along the width of the field, upon which the ball is placed before the beginning of each play. A formation is defined as the spatial configuration of a team's players before a play starts. Before each play starts, the players go to a pre-determined place on the field on their side of the line of scrimmage. This is when a team is said to be in *'formation'*. The play begins when a player lifts the ball off the ground ('snaps') and ends when the player with the ball either scores or is brought to the ground ('tackled'). Automatically detecting the line of scrimmage in a play clip is a challenging computer vision problem primarily due to high player occlusion and appearance similarity. Automatically detecting the offense formation in a play clip is difficult primarily due to significant occlusion, intra-class variation, inter-class similarity, and player appearance similarity.

Detecting the line of scrimmage and the offensive team formation are two fundamental building blocks without which any future action recognition and play understanding work is incomplete. The results of this solution can be used for more extensive analysis of football games such as offensive and defensive personnel detection, play detection, and ultimately playbook inference.

In this paper, we propose a novel framework to automatically identify the line of scrimmage and offensive team formation in football (refer to Figure 1). The proposed framework uses the gradient information of the video frames projected onto the field using the method proposed in [7]. Pro-

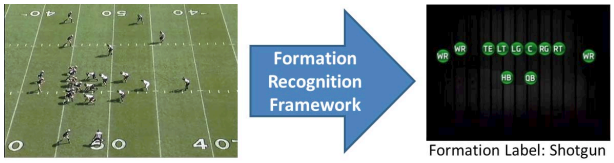

Figure 1. The goal of our work is to automatically recognize and label offensive team formation that occur at the beginning of an American Football play.

jecting the frames not only removes camera motion but also allows player locations to be normalized on to the field of play irrespective of the view recorded by the camera. This solution is robust enough to identify features from any video as it is agnostic to colour information.

Our proposed framework has four stages described as follows: (1) we automatically identify the frame in which the formation occurs, (2) we use the identified formation frame to automatically estimate the line of scrimmage, (3) we use the estimated line of scrimmage to differentiate players from the two teams and automatically identify which team is on offense, and (4) we use features extracted from the offensive side of the line of scrimmage to classify the offensive team's formation in the play clip.

The rest of the paper is organized as follows. We briefly discuss some existing work in football play analysis and formation detection. In Section 3, we illustrate our proposed framework and describe each of the framework modules in detail. In Section 4, we validate our framework by extensively testing our methods on different American football play datasets.

2. Related Work

Most of the work on American football video analysis has been on recognizing and classifying the different types of plays[12, 13, 11]. Little work has been done in automated offensive team formation recognition. Given a formation image, Hess et al. [10, 9] used a mixture-of-parts pictorial structure model (MoPPS) to localize and identify football players in order to recognize the formation in the image. Their work was tested on a limited dataset, the learning method for MoPPS model was computationally expensive and parameters of the model were hand-coded. In addition, their work assumes that the formation image is given, however in most cases, video coaches work with play clips, and would have to go through the whole play clip to find the formation image. Our work automatically finds the formation frame in the play clip, identifies the line of scrimmage, and labels the formation in the detected formation frame. Related work on soccer team formations [2, 1, 14] are not directly applicable to the American football domain as soccer tends to be more fluid and dynamic, changing over the course of the game, while American football games tend to be more structured and inherently repetitive.

3. Proposed Framework

Our proposed solution aims to identify the formation of the team playing offense in a game of football. Figure 2 shows the block diagram of our overall proposed framework. The input to our framework is a single football video play clip. The proposed approach consists of two major modules: *pre-processing* of the play video clip and *formation recognition* of the offensive team in the play clip. This paper focuses on the formation recognition module.

Given a registered play clip, the framework will first identify the frame in which the two teams are lined up in formation. Once this frame is detected, the field line that separates the two teams at formation time (otherwise known as the line of scrimmage) is determined in the formation frame. We utilize the spatial distribution of the two teams to identify the offensive team. Finally, we extract features from the offense region to train a linear SVM to classify the different offensive formations.

3.1. Pre-processing

Pre-processing of a given play clip includes registering the video to the reference field image and foreground/background identification. We use the robust registration method in [7] to map each frame of each play clip to the field image (refer to Figure 2). Since camera motion (zoom, pan and tilt) and viewpoint vary between plays and within the same play, registering frames unto the same coordinate system (the overhead field image) is crucial for formation recognition. The registration method matches entire images or image patches. The matching process computes an optimal homography between every pair of frames in the play clip by assuming that outlier pixels are sufficiently sparse in each frame. For a given play clip, an overhead view of the playing field is not necessarily given, which precludes matching any frame of the clip to the reference. In this case, user interaction is required to register a single frame to the reference, whereby the user selects at least four corresponding pairs of points between the two images. Coupling this homography with the homographies computed between consecutive frames of the clip, we can register the entire play clip unto the reference field image. Note that user interaction is only required once for each different field.

After registering the clip to the field, we proceed to discriminate and extract the foreground (players) from the background (e.g. field patterns such as markers, lines, and logos). One way this can be done is through player detection and classification (as in [3]), where a classifier's confidence at each region of each frame designates the likelihood that a player exists at that region. However, this process requires training samples for players from different teams and from different camera viewpoints. Also, since the video frames are captured from far-field, the player resolution is

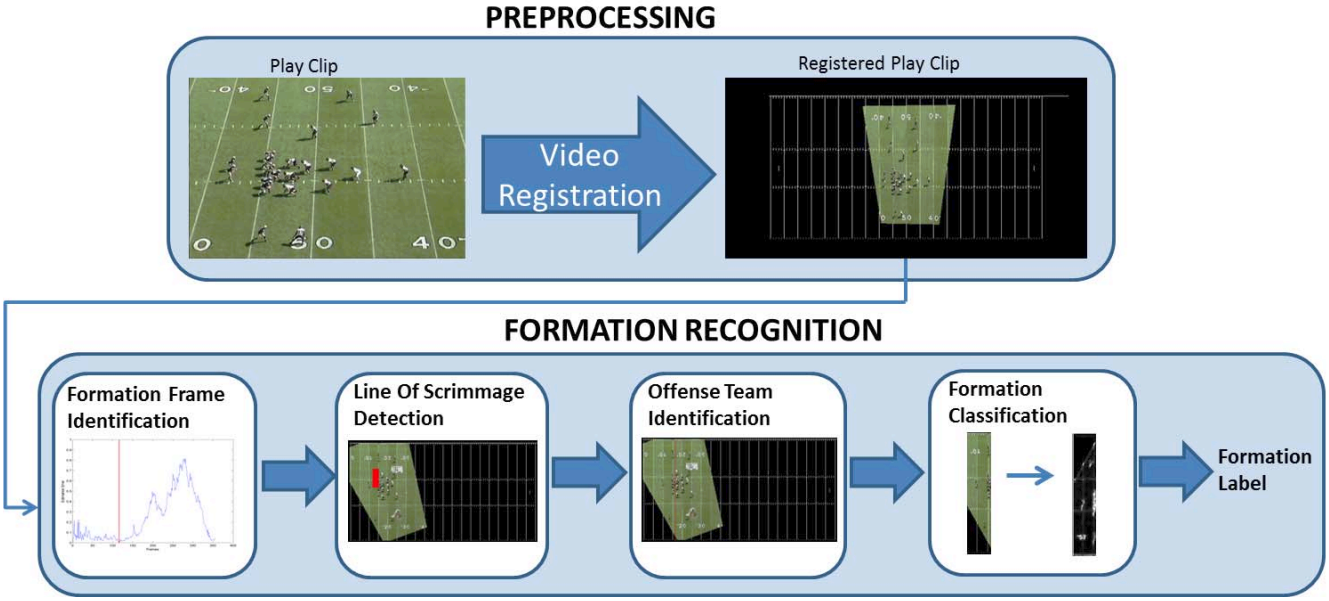

Figure 2. Overall framework for automatic offensive team formation recognition in American Football play.

usually low, so learning appearance-based classifiers leads to substantial false positives, especially at markers and mid-field logos. So, to avoid view dependent classification from far-field, we resort to background subtraction after all the frames of a play clip are registered to the same field image. This process can also be viewed as video stitching, where the panoramic generated by the stitch is the background with the rest constituting the foreground. Although many background subtraction methods exist in the literature (e.g. [6, 5]), the majority assumes that the camera to be static and thus do not address the problem of camera motion. In our case, since frames are registered in the same coordinate system, conventional background subtraction methods can be applied; however, special care has to be taken because not *all* pixels are necessarily visible in all frames. Instead, we proceed to extend the robust registration work in [7] to allow for robust video stitching in the presence of incomplete data and sparse error. This sparse error corresponds to pixels that do not satisfy the homography mapping, specifically pixels belonging to players on the field. Sparsity here is a valid assumption, since the frames are imaged in the far-field and players generally constitute a small fraction of the pixels in a frame.

Robust Video Stitching To mathematically formalize the stitching problem, we are given a set of F frames of a play clip registered to the field image. The projection of the k^{th} frame into the reference frame is denoted as $\vec{\mathbf{i}}_k \in \mathbb{R}^M$. As such, each of the mapped frames into the reference cover a subset of the whole panoramic view of the field. This panoramic image contains M pixels. We denote this panoramic image (sometimes known as the intrinsic image) in vector form as $\vec{\mathbf{v}}$. The goal of our analysis is to reconstruct the panoramic $\vec{\mathbf{v}}$ in the presence of sparse error and frame-to-frame global illumination change, modeled by $\vec{\mathbf{u}} \in \mathbb{R}_+^F$. To solve for $\vec{\mathbf{v}}$, we formulate the video stitching problem as a rank-1 matrix completion problem with sparse error. This problem is stated in Eq. (1), where $\mathbf{I} \in \mathbb{R}^{F\times M}$ is the concatenation of all the mapped frames (refer to Figure 3 for an illustration), $\mathbf{W} \in \mathbb{R}^{F\times M}$ is the concatenation of pixel weights (weighted observable supports) for all the mapped images in the reference image, and $\mathbf{E} \in \mathbb{R}^{F\times M}$ is the sparse error matrix. Here, we denote $\mathbf{W} \circ \mathbf{I}$ as the Hadamard product between the two sparse matrices $\mathbf{W}$ and $\mathbf{I}$. In the simplest case $\mathbf{W}$ is a sampling matrix, where each element is a binary value indicating whether or not the pixel is visible in that frame. More generally, $\mathbf{W}$ can be considered a weighting matrix for each frame or even for each pixel of each frame in the clip.

$$\min_{\vec{\mathbf{u}},\vec{\mathbf{v}},\mathbf{E}} \quad \|\mathbf{E}\|_{1,1} \tag{1}$$
$$\text{subject to:} \quad \mathbf{W} \circ \mathbf{I} = \mathbf{W} \circ \left(\vec{\mathbf{u}}\vec{\mathbf{v}}^T + \mathbf{E}\right)$$

In fact, Eq. (1) can be viewed as the rank-1 version of robust PCA (RPCA [4]). Apart from being a valid assumption for video, where frame-to-frame lighting variations are not significant, the rank-1 model for $\mathbf{W} \circ \mathbf{I}$ precludes the computational infeasibility of RPCA on such large scale data, especially since RPCA involves expensive SVD operations. To put this in context, F is usually on the order of hundreds of frames, while M is in the order of 10^5-10^6 pixels. Eq. (1) is non-convex due to the non-convex equality constraint. In this way, the formulated optimization problem is similar to GRASTA [8], except there is no tracking of an orthonormal subspace basis. Ideally, without illumination change

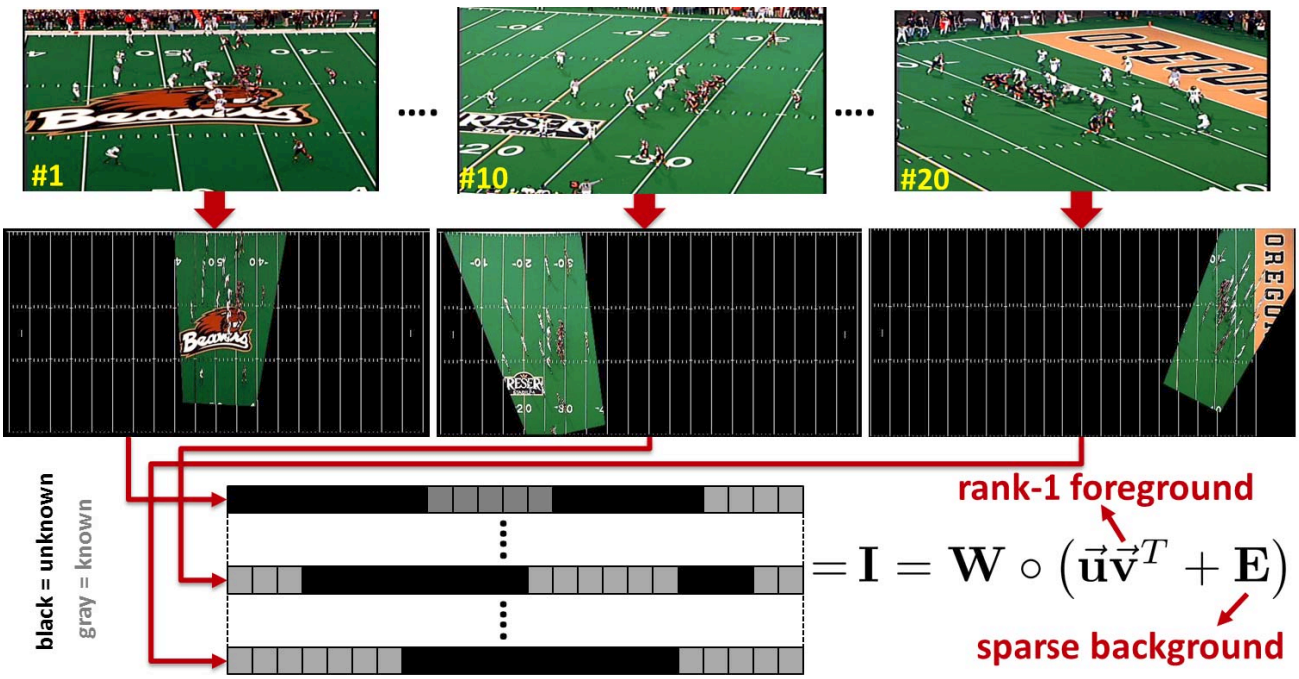

Figure 3. Robust video stitching for foreground/background estimation. Each frame in the play clip is projected into the overhead field reference using the registration method in [7]. The projection of frame j (j^{th} row of $\mathbf{I}$) constitutes a small part of the panoramic image $\vec{\mathbf{v}}$ and contributes sparse error (j^{th} row of $\mathbf{E}$). The visible pixels of the panoramic in each frame of the clip are defined in a binary matrix $\mathbf{W}$. The global illumination variation in the clip is addressed by the positive coefficients $\vec{\mathbf{u}}$.

due to camera settings, we have $\vec{\mathbf{u}} = \vec{\mathbf{1}}$. Therefore, to solve Eq. (1), we resort to an alternating descent approach, which alternates between fixing $\vec{\mathbf{u}}$ and updating $\vec{\mathbf{E}}$ and $\vec{\mathbf{v}}$ via the inexact augmented Lagrangian method (IALM) and fixing $\vec{\mathbf{v}}$ and $\vec{\mathbf{E}}$ and updating $\vec{\mathbf{u}}$ using a similar method. Although in general, this strategy does not guarantee the global solution (and a local solution at best); however, for most cases when the image sequences are captured by the same camera over time, the local solution that is obtained by initializing $\vec{\mathbf{u}} = \vec{\mathbf{1}}$ leads to a reasonable solution. For the case of spatially varying illumination change from image to image, there is a need to produce a higher rank approximation to the error-less $\mathbf{I}$, which can be approximated greedily by recursively solving Eq. (1) with $\mathbf{I}$ replaced by $\mathbf{I} - \vec{\mathbf{u}}\vec{\mathbf{v}}^T$. We do this to keep the computational complexity of the stitching process at a minimum. It is easly shown that each IALM iteration only involves simple and highly parallelizable matrix operations (e.g. matrix subtraction and addition) and *no* SVD operations are needed. We provide all the optimization details of this method in the **supplementary material**. As compared to RPCA, one interesting property of this solution is that it can be executed both in batch (all frames together) and incremental (e.g. one frame at a time) modes.

In Figure 4, we show an example of applying robust video stitching to a football play clip, where the different frames are stitched together to generate the panorama and their respective sparse errors. Clearly, the error values are high at player locations and low in the background. As a result, the video stitching method is able to weigh each pixel in a frame as either background or foreground, thus essentially identifying only the moving objects or players in the play clip. This is especially useful for removing large midfield logos and field markers, which tend to be misclassified as moving players. It is worthwhile to note that the background is separated from the foreground even though the foreground is static for many frames, especially at the beginning of the play. As such, the output of the registratio/stitching module is subsequently used in the formation recognition module described next.

3.2. Formation Frame Detection

The first task in the formation recognition module is to identify the frame in which the teams position themselves in a predefined formation before the start of play. Intuitively, the formation frame is the frame with the least player motion. At this frame, the frame-to-frame pixel displacement in the formation frame has the least motion magnitude among all other frames in the play. Given that the video frames are registered to a reference field image, the frame with the least motion magnitude is easily identified by calculating the frame-to-frame mean-squared error (MSE) for all pixels across the entire video sequence followed by a 5 tap median filter. We identify the formation frame as the last frame after which the MSE values are substantial and

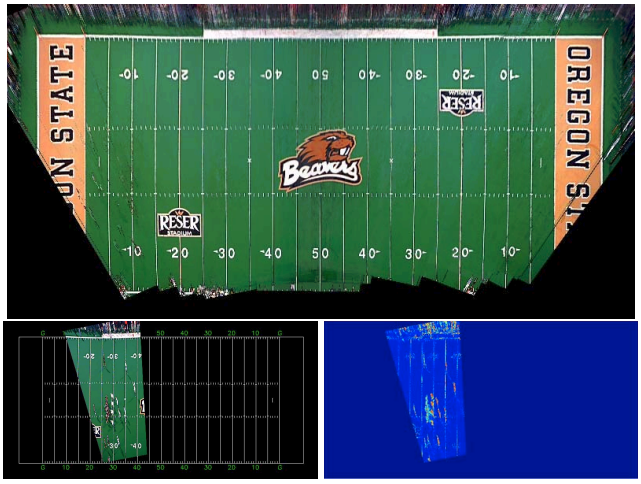

Figure 4. Robust video stitching for foreground/background estimation. The top row shows the stitched panoramic $\vec{\mathbf{v}}$ (or background) computed by solving Eq. (1). A frame in the play clip and its corresponding error (or foreground) are shown in the bottom row. Colder colors designate smaller error values.

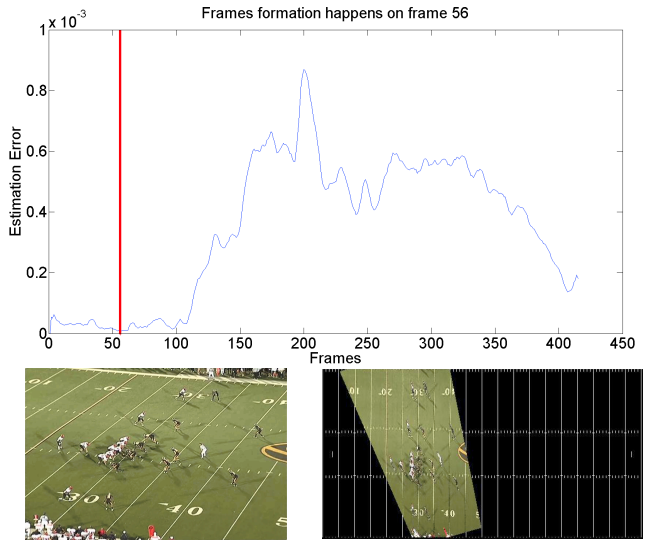

Figure 5. Formation frame detection. MSE values of all frames in a sample play clip are plotted in the top row. The index of the formation frame is designated in *red*. The detected formation frame and its projection are shown in the bottom row.

monotonically increasing. This is done to ensure that the motion of players in the frames is due to all players moving and *not* due to man-in-motion events that may occur before the actual start of the play. To formalize, we build an SVM classifier on the MSE differences between frames of a play and learn its parameters using a small set of formation frames labelled by a sports expert. In Figure 5(a), we plot MSE values of all the frames in a given play clip. The index of the formation frame is marked by the red line. Figure 5(b) & (c) show the detected formation frame in the clip and its projection on the field respectively. In what follows, the detected formation frame is used to detect the line of scrimmage and recognize the play formation of the offensive team.

3.3. Line of scrimmage detection

By reducing the effect of background pixels including the logos on the field, player (foreground) density is highest at the line of scrimmage relative to any other area in the projected formation frame. We exploit this information to identify the line of scrimmage. Colour can be seen as a feature for identifying players on offense, since the two teams wear different colored jerseys. A possible way to identify the team on offense would be to require the user to label at least one offensive player per play clip. However, most standard techniques for identifying players using colour models such as hue and saturation would fail to identify the players alone, owing to the fact that hue and saturation values also exist on the field leading to a number of false positives. Over and above that, opposing teams wear complementary colours such as a black jersey with white shorts and white jersey with black shorts making even semi-supervised methods ineffective in identifying players on offense. In this paper, we take a color agnostic approach and exploit the fact that the line of scrimmage is the region of the field in the formation frame that has the highest player density. In our case, this means that the line of scrimmage is the region where the image gradient density of the foreground is highest. Figure 6 shows the gradient intensity of the projected formation frame, weighted by the foreground (error) values obtained in the pre-processing step. Note that we only use the gradient information in the y-(vertical) direction as this would avoid including the gradient information from vertical lines such as that of yard lines. A sliding window approach is used to find the region with the highest density in the gradient image.

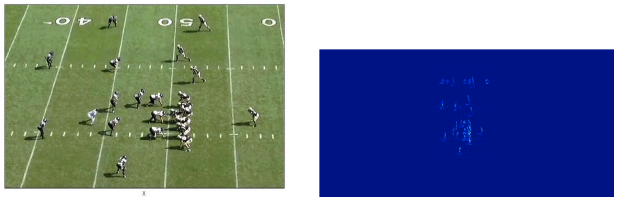

Figure 6. Example of a formation frame (left) and its corresponding weighted gradient information (right).

After each play, the ball is placed somewhere in the middle of the width of the field closest to where the previous play ended. As such, to account for the variation in the offensive team position at the beginning of a play, we use a sliding window. The window's size was set to cover the offensive team and changes according to the size of the field image. For our experiments on the standard definition dataset, we set the window size to be 20×160 pixels

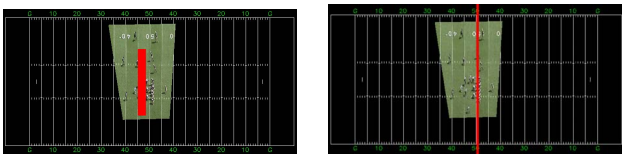

Figure 7. A sliding window (shown in red on the left) is used to calculate the density of gradient intensity. The corresponding detected line of scrimmage is marked in red on the right.

respectively for field size image of 800 $\times$ 400. The window slides across the entire length of the projected formation frame region, not across the whole field. We calculate the sum of gradient intensities in each window in Eq (2) and determine the line of scrimmage in Eq (3). Figure 7(a) shows an example of the sliding window, while Figure 7(b) shows the detected line of scrimmage marked in red.

$$Dens(y) = \sum_{i=1}^{M_w} \sum_{j=1}^{N_w} \left| \frac{\partial I_w}{\partial x} \right| \tag{2}$$

$$\text{line-of-scrimmage} = \arg\max_y Dens(y) \tag{3}$$

3.4. Offensive Team Identification

Football presents many technical hurdles for analysis. This is especially due to frequent and substantial player occlusion. However, football is highly repetitive and structured in nature. We exploit this repetition to distinguish offense from defense. The offensive team tends to be compactly positioned at the line of the scrimmage when in formation. There are at least five offensive players near the ball ('offensive line') in close proximity in front of the quarterback. The defense on the other hand, is usually more spread out, especially along the line of scrimmage. Therefore, the team on offense is determined as the side of the line of scrimmage (left or right) where the spatial distribution of players (foreground) on that side has minimum variance. The foreground distribution on either side of the line of scrimmage is estimated by a spatial pmf as shown in Eq (4). Here, we model the probability that a player exists at pixel (x, y) as a function of the gradient intensity and foreground (error) value at (x, y).

$$p(x, y|d) = \frac{\partial I(x,y)/\partial x}{\sum_{(x,y)\in d} \partial I(x,y)/\partial x}; \;\; [d \in \{\text{left}, \text{right}\}] \tag{4}$$

The side d with the minimum variance is determined to be offense.

$$o = \underset{d\in\{\text{left,right}\}}{\arg\min} \; \sigma\left(p(x, y|d)\right) \tag{5}$$

3.5. Formation Classification

At this stage, the formation frame, line of scrimmage, and the offensive team of each play clip have been extracted. We proceed to learn a discriminative model to classify a play clip as one of five major formation types for offense. These formation types (classes) and the ground truth labeling of a play clip dataset are obtained from a football expert. We distinguish between the five formation classes by learning a multi-class linear SVM classifier on a training subset of the play clip dataset. The feature used to learn the SVM is the spatial distribution of gradient intensity for the offensive side of the line of scrimmage. Since the location of an offensive team can vary on the field, we make the feature translation invariant by centering the spatial distribution at the field pixel (on the offensive side) with maximum gradient intensity. Note that we reflect the formation frame about the line of scrimmage to ensure that the offensive team is always on the left side of the line of scrimmage. The SVM classifier parameters are learned using a popular and publicly available SVM solver. Results of this classification are presented in the next section.

4. Experimental Results

In this section, we present experimental results that validate the effectiveness of our formation recognition framework. We tested our approach on three separate datasets.

4.1. Datasets

We evaluate the performance of our detection and classification framework by applying it to real-world football videos. There is no publicly available dataset that accurately represents video captured and used by American Football teams. Typically, a football coach would segment the video of a whole game into single play clips, identify their respective formation frames, and then classify the formation type. So to evaluate our framework, we test it on three separate datasets, two of which we compiled ourselves and plan to make publicly accessible. To the best of our knowledge, our dataset is the largest dataset for American football plays.

The two datasets we compiled are denoted SD dataset and HD dataset. They comprise 274 and 541 distinct offensive plays respectively. In both cases, the football game videos were collected from several colleges and show the sideline view of the game. The videos were manually segmented such that each clip shows a single play. In dataset SD (standard definition), play clips are standard resolution 640 $\times$ 480. In dataset HD, they are high-definition 1440 $\times$ 1080 resolution. All play clips were registered to the football field reference image using the video registration and stitching method discussed in Section 3.1. The groundtruth data for the formation frame, line of scrimmage, and formation label were annotated by a sports domain expert with more than 6 years of experience in studying and analyzing football games. The third dataset, denoted OSU dataset, was obtained from the Oregon State University Digital Scout Project [10]. This dataset consists of 51 formation frame images and their corresponding homography matrices for projection to the reference field image.

There exist more than a hundred known formation types

(classes) in the sport of football. However, due to the limited amount of data available, we categorize the plays in each dataset into two different levels of categorization. Table 1 shows the hierarchy of the formation classes. At the coarse level of classification, the plays were classified into three different categories: *Ace*, *IForm*, or *ShotGun*. These top labels in the taxonomy were determined by the location of the quarterback (QB) and running backs (RB). Ace formation means there is one running back directly behind the quarterback, who is very close to the ball. IForm is similar to Ace, except there are two running backs directly behind the quarterback instead of one. In Shotgun formation, the quarterback is several yards away from the ball at the start of the play. At the second level, the classes were determined by the number of running backs (RB) and wide receivers (WR). The *Ace* formation consists of one subclass: Ace.3WR. The *IForm* class is subdivided into IForm.2WR and IForm.3WR, while the *ShotGun* formation consists of five subclasses: Shotgun.2WR, Shotgun.3WR.1RB, Shotgun.3WR.2RB, Shotgun.4WR.1RB, and Shotgun.5WR. The small inter-class variability in the datasets makes formation classification a challenging problem. Figure 8 shows examples of the different formations.

4.2. Quantitative Results

We tested and evaluated each of the modules in our framework separately.

Formation Frame Detection: To measure the performance of our formation frame detection method in Section 3.2, a formation frame is accurately labelled if the detected formation frame is within 1 second (30 frames) of the groundtruth formation frame index. This evaluation criterion was deemed acceptable by the sports expert. In this setup, the formation frame was accurately detected 94.53% of the time on the SD dataset with 274 play clips. We could not test this module on the OSU dataset as the dataset consists only of frame images and not play video clips.

Line of Scrimmage Detection: After identifying the formation frame, we use the foreground (error) values for each pixel in the projected formation frame (as described in Section 3.1) as well as the gradient intensity to automatically detect the line of scrimmage, as described in Section 3.3. In fact, weighting the pixels of the formation frame with the foreground values significantly reduces false positive detection arising from regions of high gradient density such as field logos or markers.

To test our automatic line of scrimmage detection method, we use the groundtruth formation frame in the datasets. For the SD dataset, we achieve 97.5% accuracy in detecting the line of scrimmage within 10 pixels of its actual location on the field. This translates to within a yard accuracy. The detection results improve significantly on the HD dataset, where we achieve a 97.9% detection accuracy within 3 pixels on an HD reference field. This translates to a 6 inch accuracy on the physical field. As mentioned earlier, the OSU dataset only contains formation images and no play clips. Hence, we could not directly apply the robust registration/stitching method in Section 3.1 to obtain the per-pixel foreground values in the formation frame. Instead, we apply the video stitching method on the given formation frames in the dataset to generate an overhead panoramic stitching of the field. The resulting panoramic image of the OSU field is shown in Figure 4(top). This image is subtracted from each of the projected formation frames to produce the foreground error values (i.e. background subtraction). By learning a multi-class linear SVM classifier, we achieve an accuracy of 90.2% on the OSU dataset.

Table 1. Hierarchy of the formation classes in the dataset.

Top level label	Second level label	# instances
Ace	Ace.3WR	30
IForm	IForm.2WR	40
	IForm.3WR	19
Shotgun	Shotgun.2WR.1RB	21
	Shotgun.3WR.1RB	110
	Shotgun.3WR.2RB	85
	Shotgun.4WR.1RB	106
	Shotgun.5WR	26

Formation Classification: To validate our formation classification method (described in Sections 3.4&3.5), we test on the union of the three datasets. Special care has to be taken because of the imbalance in the combined dataset. We adopt the oversampling paradigm to alleviate this imbalance. The hierarchy of the classes and the number of instances in each class are reported in Table 1.

All classification model training and testing are performed using 5-fold cross validation. We perform two levels of classification. For top level classification, we train a classifier to classify formations into one of three super classes: Ace, IForm, and ShotGun. The classification rate for the top level classification is 67.1%. At the next level of classification, we train two classifiers, one each for labeling the IForm and ShotGun formations into their subclass labels respectively. The classification rate for labeling IForm formations into two subclasses, IForm.2WR and IForm.3WR, is 65.01%, while the classification rate for classifying ShotGun formations into the 5 different subclasses is 36.43%. Figure 9 shows the confusion matrix for classification of the ShotGun formations into the subclasses. The classification rate is lower compared to the classification of the IForm formation because the differences between the subclasses are very subtle with the distinction being just the placement of one player in the formation. We also compared the performance of the hierarchical classifier to a simple multi-class classifier. When using a flat multi-class classifier, we obtained an overall classification accuracy of 31.1%. The con-

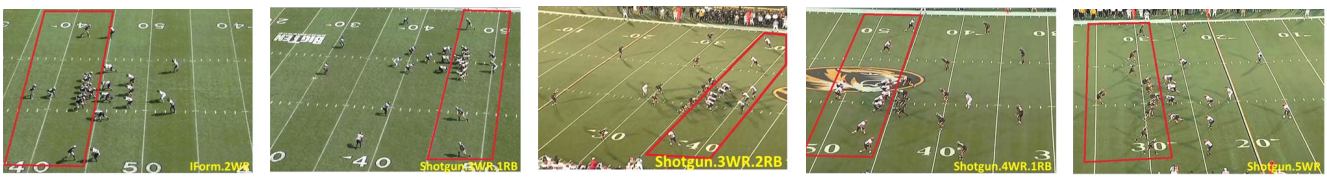

Figure 8. Examples of 5 offensive team formation classes (red bounding box). It can be seen that the inter-class variation between the different formations is very small, making the classification problem a challenging task.

fusion matrix is shown in Figure 10. Again the confusion mainly occurs within the ShotGun formation as the difference between the classes is very subtle.

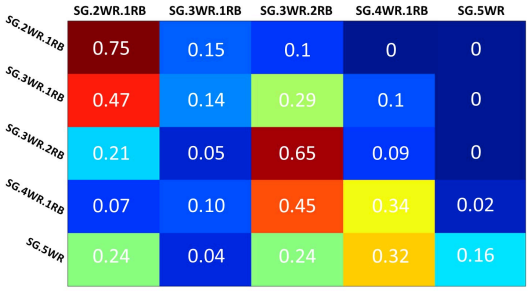

Figure 9. Confusion matrix for classification of ShotGun formations into five subclasses.

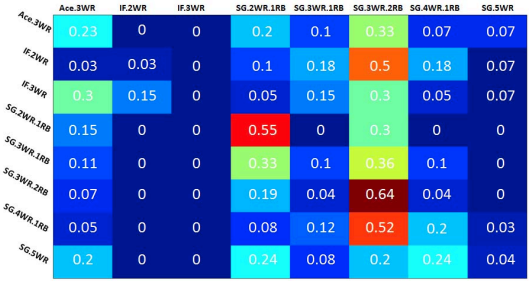

Figure 10. Confusion matrix for simple multi-class classification using all 8 formation classes.

5. Conclusion

In this paper, we propose a framework for automatic recognition of offensive team formation in American football plays. We show promising results that our method is able to automatically label formations on three different datasets. The framework that we have developed serves as a building block that is part of our tool prototype for automatic analysis of American football games. Future work utilizing this formation classification framework include automatic personnel identification or identifying all players that are on the field, automatic play classification, and strategical playbook inference of a team.

6. Acknowledgement

This study is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapores Agency for Science, Technology and Research (A*STAR).

References

[1] Ayanegui-Santiago and Huberto. Recognizing team formations in multiagent systems: Applications in robotic soccer. In *In Lecture Notes in Computer Science*, volume 5796, pages 163–173, 2009.

[2] M. Beetz, N. von Hoyningen-Huene, B. Kirchlechner, S. Gedikli, F. Siles, M. Durus, and M. Lames. Aspogamo: Automated sports game analysis models. *International Journal of Computer Science in Sport*, 8(1), 2009.

[3] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *PAMI*, 33(9):1806–1819, 2011.

[4] E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3), 2011.

[5] F. De la Torre and M. Black. A framework for robust subspace learning. *IJCV*, 54(1-3):117–142, 2003.

[6] X. Ding, L. He, and L. Carin. Bayesian robust principal component analysis. *Image Processing, IEEE Transactions on*, 20(12):3419–3430, 2011.

[7] B. Ghanem, T. Zhang, and N. Ahuja. Robust video registration applied to field-sports video analysis. In *ICASSP*, 2012.

[8] J. He, L. Balzano, and A. Szlam. Incremental gradient on the grassmannian for online foreground and background separation in subsampled video. *CVPR*, 2012.

[9] R. Hess and A. Fern. Toward learning mixture-of-parts pictorial structures. In *ICML Workshop on Constrained Optimization and Structured Output Spaces*, 2007.

[10] R. Hess, A. Fern, and E. Mortensen. Mixture-of-parts pictorial structures for objects with variable part sets. In *ICCV*, 2007.

[11] R. Li and R. Chellapa. Recognizing offensive strategies from football videos. In *ICIP*, 2010.

[12] B. Siddiquie, Y. Yacoob, and L. S. Davis. Recognizing plays in american football videos. Technical report, University of Maryland, 2009.

[13] E. Swears and A. Hoogs. Learning and recognizing complex multi-agent activities with applications to american football plays. In *WACV*, 2012.

[14] U. Visser, C. Drcker, S. Hbner, E. Schmidt, and H.-G. Weland. Recognizing formations in opponent teams. In *In Lecture Notes in Computer Science*. Springer-Verlag, 2001.