This is the published version:


Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30044893

©1998 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Copyright: 1998, IEEE
Combining NL processing and video data to query American Football

Mihai Lazarescu, Svetla Venkatesh, Geoff West
Curtin University of Technology
GPO Box U1987, Perth 6001, W.A.
lazaresc@cs.curtin.edu.au

Terry Caelli
Ohio State University
1216 Kinnear Rd., Columbus, Ohio 43212
caelli@cfm.ohio-state.edu

Abstract

In this paper we explore the use of natural language understanding and image processing to index and query American Football tapes. We present a model for representing spatio-temporal characteristics of multiple objects in dynamic scenes in this domain, and a recognition system which uses the model to recognise American Football plays.

1. Introduction

Natural language understanding and image understanding are two of the main areas of research in artificial intelligence. The main benefit of using both image information and natural language is that it allows for information to be fused from both sources. Furthermore it allows semantic understanding of video at a higher level than would be possible by looking at video images alone.

VITRA (VIsual TRANslator) [8, 3] is one of the projects which tries to integrate computer vision and the generation of natural language expressions for the description of image sequences in traffic scenes (CITYTOUR system [14, 15, 9, 7]) and soccer sequences (SOCCER system [14, 15, 21, 20, 19]). Another project which combines natural language understanding with image processing is Informedia: News-on-Demand [17, 13, 1, 2, 6, 18]. This project creates a digital library of text, images, video and audio data available for full content search and retrieval. News-on-Demand [4, 12] is a fully-automatic system that monitors TV, radio and text material and allows selective retrieval of news stories based on spoken queries.

The objective of our work is to use natural language understanding and image processing to index and query American Football tapes. We have developed a model for representing spatio-temporal characteristics of multiple objects in dynamic scenes in this domain. Our representation combines several types of knowledge: expert knowledge, domain knowledge (game rules), spatial knowledge (player relationships) and temporal knowledge (action sequencing in individual plays).

2. Play Model Description

The play model description we propose to represent an American Football play has several components which describe the play action in terms of player significance, player relationships, player movements, player actions, player positions and the temporal sequencing of the player actions and movements. The significant characteristic of the play model representation is that it is mainly made up of symbolic data. The play model has the following features:

The play class - there are three classes of plays in American Football that are defined by the main action that takes place in the play. The play classes are defined by set A where: A = {offensive, defensive, special}.

The significant players and actions in the play - each play involves significant players performing predefined tasks. We define as significant the players whose actions are very likely to have a major impact on the main action and outcome of the play. The player significance is defined by set B where: B = {low, average, high, very-high}. The possible actions are defined by set C where: C = {pass, toss, give, hand, catch, get, receive, fake, tackle, block, push, grab, intercept}.

The player combination - each play requires a specific combination of players.

The temporal sequencing of the actions in the play - in each play certain actions take place at specific times and in a certain order. The temporal representation of the play model is based on a set of simple movements combined with a play action. The play action is the main action in the play already defined in the first part of the play model. The movement describes the type of movement of a given player at an instance in time. A player can either move to a position or make a turn relative to the direction in which the football player is facing. The possible movements of a player are defined by set D where: D = {move, turn-left, turn-right}. The order and timing of the combined move-
The movement of individual players in the play - players generally have specific movement patterns in specific plays. The movement of the player over the entire play period is described in terms of shape and distance. The shape of the movement indicates whether the movement is straight or contains turns. The occurrence of a turn is signalled by a change in the direction of the player’s movement which is detected over adjacent frames in the video sequence. The distance covered by the player during the play can be either short or long. To determine the type of distance, the system uses the player’s starting and finishing positions and the position of the defensive line. If the player started behind or in the same line with the defensive line and finished in-front of the defensive line then the distance covered is considered to be long. Otherwise the distance is short. The possible overall movements are defined by the set $E$: $E = \{ \text{short-and-straight, short-and-turn, short-and-many-turns, long-and-straight, long-and-turn, long-and-many-turns} \}$.

The relationships between a player and the other teammates at different instances throughout the duration of the play are described symbolically and derived from four reference points which are assumed to be known at all times. These are the left boundary, the right boundary, the opponent’s goal and the team’s own goal line. The references are dependent on which way the attacking team is facing. The possible relationships are defined by the set $G$: $G = \{ \text{in-front, behind, to-the-left, to-the-right, in-line-with-horizontal, in-line-with-vertical} \}$.

To build the play models we have created a database of plays information which includes data about the player combinations, play actions, player significance and player movement over the duration of a “by the book” play. The information about the plays has been gathered from several sources which include Gridiron play manuals, video tapes and Gridiron computer games. The system collects information, forms the database about a specific play requested and generates the play model description.

3. Extraction of Information from Text and Video

The system developed processes text derived from natural language and geometric information from images to determine the type of play contained in the query. The natural language commentary plays a crucial role in the recognition process because it is an expert description of the action in the video and therefore contains much information about the play. The commentary text provides three clues about the play in the video: the type of action, the name and type of the players involved and game statistics. The first is the type of action that takes place in the game, e.g. “the ball was passed”. The second clue gives the name and the type of the players that have been involved in the play action. The text “Bono throws the ball high to receiver Vanover” describes a play “pass” in which the position of one player is that of a receiver by the name of Vanover. The third clue is the game statistics, e.g. score of the game, the time remaining in the game. This information describes the status of the game and by and large dictates the course of action that is taken by the team in possession of the ball.

The geometric information extracted from the video is made up of a list of unlabelled player coordinates from the frames. This information is used to determine the direction of movement and acceleration of players and the initial and final player setups and positions. The information from the video is assumed to be incomplete and can possibly contain noisy data about the player coordinates at different frames in the video.

4. Play Model Recognition Algorithm and Results

The recognition routine was built around the play model representation and consists of three stages which check whether the query play is similar to any of the known plays (that is if the system already has a model representing the play stored in the memory) and, if similar, then it determines the degree of similarity.

Once the system has processed the football commentary and the video data to extract information (outlined in the previous section), it proceeds to search the solution lattice for potential matches and outputs a list of plays identified as potential matches in ascending order (best match first). Each of the stages in the recognition routine are described in this section.

The process involves pruning a lattice which is used to classify all the plays known by the system. At the start, all plays are considered to be a match to the query. As the system descends the first three levels in the lattice, it eliminates (or prunes) potential play candidates that do not agree with the clues provided from the different input sources. The candidates remaining when the system reaches the fourth level in the lattice have their similarity score updated after each level until the system reaches the leaf nodes. The information from the frames developed in the text analysis stage is used to conduct three tests in sequential order to test for the play class, action type and players involved.

Player Position Labelling. In the second stage of the recognition routine, for each play model in the solution lattice, the system compares the initial setup of the players in the play model and in the query. This is done for the purpose of labelling the players in the query play.
To achieve this we search for the best possible match between the positions of the unlabeled players in the query play and positions of the labeled players in the play model.

**Player Relationship Comparison.** Finally, we test the relationships between the players at each instance in the play action. For each instance in the query play we take one player and develop a symbolic representation of his relationship with his teammates. We then compare the relationships with those of his corresponding player in the known play model.

The third stage deals with the temporal occurrence and sequencing of the actions of the players in the query play. The system projects the end points of the time intervals in the query and play model onto a single axis.

Then for each subinterval projection if the action in the query matches the action in the play model, the evidence score is incremented, otherwise the evidence score is unchanged. The process is repeated for each player. At the end of the temporal processing, the plays are arranged in order of descending evidence score and the one with the highest score is chosen.

Of 20 tests carried out on real data, the system performed correctly 18 times out of 20. It failed when the text was irrelevant to the game. If only video analysis is used, it classifies only 12 of the 20 plays which demonstrates that by combining NL with video processing the accuracy of the classification is improved (for a detailed set of results see [16]).

### 5. Conclusions

We have developed a model for representing multiple labelled objects in a dynamic scene. Our representation combines several types of knowledge: expert knowledge (game tactics), domain knowledge (game rules), spatial knowledge (player relationships) and temporal knowledge (action sequencing in individual plays). This is used in a system that is able to recognise American Football plays. The recognition system has the advantages that: (1) it is not dependent on accurate low level data, (2) its search for a match for a potential query is driven by expert input and it cross checks its input from the several sources it has available.

### References