METADATA EXTRACTION FROM TEXT IN SOCCER DOMAIN

Ziya Ozkan Gokturk[1], Nihan Kesim Cicekli[2] , Ilyas Cicekli[3]

[1]SIEMENS EC, METU, Technopolis, Ankara, 06531, Turkey
[2]Department of Computer Engineering, METU, Ankara, 06531, Turkey
[3]Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey

Abstract — Event detection is a crucial part for soccer video searching and querying. The event detection could be done by video content itself or from a structured or semi structured text files gathered from sports web sites. In this paper, we present an approach of metadata extraction from match reports for soccer domain. The UEFA Cup and UEFA Champions League Match Reports are downloaded from the web site of UEFA by a web-crawler. Using regular expressions we annotate these match reports and then extract events from annotated match reports. Extracted events are saved in an MPEG-7 file. We present an interface that is used to query the events in the MPEG-7 match corpus. If an associated match video is available, the video portions that correspond to the found events could be played.

***Keywords* — Semantic querying of video content, MPEG-7, information extraction, video annotation**

I. INTRODUCTION

Sport fans could not watch every live game because of several reasons such as time and region differences, channel availability etc. There will be highlights of games but they are generally prepared by studio professionals and they do not cover the audience's appetite. On the other hand, people want to see events related to certain teams or players.

Unfortunately, not all multimedia have metadata available with them. Content based knowledge extraction from large multimedia repositories is an important research area. For multimedia data without semantic content tags, it is necessary to extract the metadata automatically. Web content is usually inside XML or HTML documents, which contain additional information that can be used to obtain the metadata by applying natural language processing techniques and information extraction algorithms.

In this paper, we present a system that annotates soccer game videos automatically by using the information in match summary texts. The proposed system downloads live match reports from UEFA by a web-crawler. It, then, tags these match reports by regular expressions. All events are extracted from match reports via a hand-written rule set. These extracted events are converted to valid MPEG-7 [7] files and match corpus is generated. A user interface is provided for querying and searching the match corpus. Relevant video segment of the game is displayed for successful search results.

The rest of the paper is organized as follows. The related work on metadata extraction in sports domain is given in Section II. Section III gives brief information about the web-crawler used for downloading minute by minute match reports. Section IV demonstrates annotating text using regular expressions. Section V explains the MPEG-7 ontology and describes how the mapping of football events to that ontology is done. Section VI summarizes the implementation of our system and finally Section VII concludes the paper and gives some comments about future work.

II. RELATED WORK

Information extraction from different kinds of sources is so popular nowadays. Especially for multimedia content annotation, information extraction can be preferred over video analysis if an associated text is available with the multimedia data. For instance, it is easy to find text describing videos in sports domain, specifically soccer videos. Popular sport sites publish match reports in a structured or semi structured format where events of the game are summarized along with their time information. The video segments can be annotated by aligning them with the time information and extracting the metadata from the text in summaries.

There exist many projects that have been developed for metadata extraction for sports events. In [3,4] a framework is proposed for detecting events from live sport videos and also live text analysis. They have four modules that are live text/video capturing, live text analysis, live video analysis and live text/video alignment. Live text analysis module extracts the events from text and then these events will be synchronized with video with the live text/video alignment module. Here the main concern is to increase the precision of their video analysis techniques using text as a second source of

information. Our work is focused more on semantic querying of the matches with a more detailed description of events using the textual descriptions.

Information extraction by template pattern matching is used to summarize football matches in [6]. They use GATE for intermediate analysis results. Their focus is semantic processing in information extraction. They process documents in English and use machine learning.

An automatic audio video summarization tool is presented in [5]. The tool uses content-based metadata which is extracted from match summaries manually. Once an ontological metadata is provided, the system tries to generate summaries of the game.

SOBA [2] is an ontology based approach for metadata extraction from match reports in soccer domain. SOBA automatically downloads match reports from UEFA and FIFA and sends them to a linguistic annotation web service. After that, applying the rule set to the annotated documents, events are extracted. The extracted events are stored in their own format. In our work, on the other hand, we use MPEG7 standard which makes our system more interoperable.

The work in [1] aims to extract events from both tabular match reports which are structured, and from minute by minute match reports which are unstructured. They use video analysis results and combine them with several textual resources. The aim is to discover the relations among six video data detectors and their behavior during a time window that corresponds to an event described in the textual data whereas we are concerned with semantic querying of the game contents in our work.

Compared with previous approaches, the contributions of our approach include the mapping of match events into the MPEG-7 ontology. The usage of MPEG-7 standard makes our corpus interoperable with other systems. Besides that, synchronizing match events and match videos with MPEG-7 standard provides convenience. The search and query operations on MPEG-7 files are managed by XQuery language. Since there are many search options, we obtain a dynamic XQuery generator over MPEG-7 files. If a match video is found as the result of querying, it could be displayed according to the time of the event. The synchronization of video and event is accomplished by the minute information of event that is gathered from the MPEG7 file.

III. WEB CRAWLER

The web-crawler is used for producing the match corpus. The match data is formed from HTML documents of official web site of UEFA[1]. There are two big organizations that UEFA arranges, UEFA Champions League and UEFA Cup. These organizations have their own web sites and match centers. In their match centers, for each match, the data source contains semi-structured data of player names, referees, match result and scorers of the match. Minute by minute match report is also provided at match center. Minute by minute match reports informs people about the events at the game in a textual form. These reports include the events, performers of events and their exact time point. The crawler is able to extract minute by minute match reports from Champions League and UEFA Cup match centers. At the fixtures and results part of each competition, there are links to the match days and in each of those match days, there are links to the match reports. Therefore, match report links could be found by crawling from fixtures and results site. For each match report link, minute by minute reports are extracted and saved to the match corpus.

IV. ANNOTATING TEXT USING REGULAR EXPRESSIONS

In minute by minute match report, an event is described by one sentence which contains the exact time point of the event, performers of the event and teams of the performers. The structure of these sentences is well-defined. Therefore, extracting events from sentences could be done by matching the sentences with a template that consists of labeled match events. Besides event types, time point of event, performers of event, teams of performers could be extracted from minute by minute match report. Before extracting event types, text must be labeled or in another words tagged. Since the text is well-defined, tagging of teams, players and minutes could be performed with regular expressions. In UEFA match reports minute information takes part at the beginning of the sentence and team information follows a player name. Player names are proper nouns and always start with uppercase letters. The team of a player is indicated by a team name in parentheses. Minute is represented by numeric values. However for extra time of normal match, numeric value is followed by a plus character and then another numeric value. For extra time after the match, prefix Ex. is used and it

[1]http://www.uefa.com/

is followed by numeric values. Apart from the annotated parts, other words and punctuations are labeled as token.

After a document containing match summaries is downloaded by the crawler, it is processed by tagging each sentence properly. The minute by minute text is transformed into an XML document where each sentence is represented as a separate element. The XML file contains labeled words and tokens under the sentence element. Converting plain match text into structured XML files ease applying information extraction algorithms on the match corpus. Figure 1 shows a tagged form of the sentence: "87: Crouch (Liverpool) has an effort on goal."

```
<Sentence>
    <Minute>87</Minute>
    <PlayerName>Crouch</PlayerName>
    <Token>(</Token>
    <Team>Liverpool</Team>
    <Token>)</Token>
    <Token>has</Token>
    <Token>an</Token>
    <Token>effort</Token>
    <Token>on</Token>
    <Token>goal</Token>
    <Token>.</Token>
</Sentence>
```

Figure 1*:* A Tagged Sentence

In order to extract event information from the processed texts we prepared a rule set for each soccer event. We have identified all distinct events appearing in match reports and identified different sentence structures for each of these events. For each event type there is a set of hand-written rules. These rules are applied to the data set to extract the events in another XML file for each match. Rule set can be thought as a template for match corpus events and the extracted information on that event. For each sentence in a match report, it compares the patterns of hand-written rules and if there is a match, it will extract it from the rule set and fill it with the specified information in the sentence.

Figure 2 shows an example rule for discerning the corner event. Under the *Rule* element there are two sub elements *Pattern* and *MatchEvent*. *Pattern* is the template for the event. If the tagged sentence is coherent with a pattern of the rule, the corresponding event will be extracted according to the *MatchEvent* element of the rule. Template matching is done by matching the field name of the sentence with the pattern sub-element of rule element first and if the field name is token match the content also. In Figure 2, for corner event, there are three extracted information: minute, player name and team. That information is filled from the tagged sentence by matching the field names. After all sentences are scanned, an *xml* file is generated according to the extracted information from the rule set. The *xml* file contains the events in the format of *MatchEvent* element that is described under the *Rule* element.

```
<Rule>
    <Pattern>
            <Minute></Minute>
            <PlayerName></PlayerName>
            <Token>(</Token>
            <Team></Team>
            <Token>)</Token>
            <Token>delivers</Token>
            <Token>the</Token>
            <Token>corner</Token>
            <Token>.</Token>
    </Pattern>
    <MatchEvent>
        <CornerEvent>
        <Minute></Minute>
        <PlayerName></PlayerName>
          <Team></Team>
        </CornerEvent>
    </MatchEvent>
</Rule>
```

Figure 2: A Sample Rule for Extraction

Table 1 lists all match events and the extracted fields for the events that we can process. For each of these events, all possible sentences are examined and rules that are similar to the one in Figure 2, are created. There are two important points in creating the rules: The first element *Pattern* is the template that will be matched with the sentences in match summaries and the second element *MatchEvent* represents the extracted information. As it is seen in Figure 2, the extracted information for corner event such as *Minute*, *PlayerName* and *Team* has no content. If this rule is matched with a sentence, they will be filled by gathering values for these fields from the sentence.

Match Events	Extracted Information
Cautioned	(Minute, PlayerName,Team)
Corner	(Minute, PlayerName,Team)
Foul	(Minute, FoulCommittedPlayerName, FoulCommittedTeam, FoulSufferedPlayerName, FoulSufferedTeam)
Free-Kick	(Minute, PlayerName,Team)
Goal	(Minute, PlayerName,Team)
FreeKickGoal	(Minute, PlayerName,Team)
PenaltyGoal	(Minute, PlayerName,Team)
OwnGoal	(Minute, PlayerName,Team)
GoalPosition	(Minute, PlayerName,Team)
Offside	(Minute, PlayerName,Team)
PenaltyEvent	(Minute, PlayerName,Team)
PenaltyMiss	(Minute, PlayerName,Team)
Redcard	(Minute, PlayerName,Team)
YellowCard	(Minute, PlayerName,Team)
Substition	(Minute, SubstitionInPlayerName, SubstitionOutPlayerName,Team)
SaveGoal	(Minute, PlayerName,Team)

Table 1: Match Events

V. CREATING MPEG-7 METADATA

It is preferable that we store match events in a standardized manner so that other systems can use the same match corpus for their systems. Besides, we want match events to be synchronized with the football videos. For this purpose, we use MPEG-7 standard to keep the semantic annotations of the games.

MPEG-7 (Multimedia Content Description Interface) developed by MPEG (Moving Picture Experts Group) aims at standardizing the annotation of multimedia content especially for interoperability purposes. XML syntax is used for Description Definition Language (DDL). It allows the creation of MPEG-7 Description Schemes and Descriptors. The use of XML smoothens the ability to work with other metadata standards.

MPEG-7 descriptions collaborated with audiovisual data content may comprise of pictures, graphics, 3D models, audio, speech, video, and composition information about how these elements are joined in a multimedia scenario. MPEG-7 descriptors do not depend on how described content is stored or coded. MPEG-7 description can be created for a picture or an analogue movie in the same way as a digitized content. MPEG-7 permits different granularity in its descriptions to have different level of discernment. It does not depend on the representation of material. If the material has certain relations in time and space, it will be possible to attach descriptions to elements within the scene. Since the descriptions can be attached to time and space relationship, it will be adapted to the context of an application.

The MPEG-7 standard allows querying the metadata and synchronizes it with the audiovisual (AV) content. It involves the descriptors that are associated with AV. Attributes for AV content such as location; time and quality are described in MPEG-7 Description Schemas. The description Schemas allow more complex descriptions by declaring relationships among the description components. In MPEG-7 descriptions are arranged into categories of multimedia, audio and visual domain. Their combination and possibly textual data related to them could be described in content of Description Schemas.

```
<SemanticBase id="Prica_AaB_45+1_1" xsi:type="AgentObjectType">
  <Label>
    <Name />
  </Label>
  <Definition>
    <FreeTextAnnotation />
  </Definition>
  <MediaOccurrence>
    <MediaLocator xsi:type="TemporalSegmentLocatorType">
      <MediaTime>
        <MediaTimePoint>45+1</MediaTimePoint>
      </MediaTime>
    </MediaLocator>
  </MediaOccurrence>
  <Relation type="urn:...:agentOf" target="Cautioned" />
  <Relation type="urn:...:memberOf" target="AAB" />
  <Relation type="urn:...:hasAccompaniertOf" target="ANDERLECHT" />
  <Agent xsi:type="PersonType">
    <Name>
      <GivenName>Prica</GivenName>
      <FamilyName />
    </Name>
  </Agent>
</SemanticBase>
```

Figure 3: Example MPEG-7 Descriptors

In our work, match corpus have an MPEG-7 file for each match that is extracted from minute by minute match reports. In each file, all events and their time points are represented according to MPEG-7 Descriptions Schemas. MPEG-7 standard allows us to define semantic content under the *Semantic Description*. *SemanticBase* element under the *Semantics* is used for performers of events with type

AgentObjectType which let us to describe performer under the *Agent* element. *MediaTimePoint* under the *MediaTime* is used for the minute information of the event. For the team information, event and opponent team information, we used the *Relation* element. For event name, relation type will be *agentOf*; for team information of the performer relation type will be *memberOf* and for the opponent team relation type will be *hasAccompaniertOf*. Player name is stored under the *Agent* element with type *PersonType*.

There are some events that have hierarchical representation such as foul event which is a combination of *FoulCommitted* and *FoulSuffered* event. For these events, two separate events are created and the relationship between these events is guaranteed by another *SemanticBase* element.

VI. IMPLEMENTATION

In our work, we download match reports from UEFA web site by crawling and we extract match events from these reports. We use the NekoHTML for parsing the UEFA web site which is a simple HTML scanner that enables parsing the HTML documents and accessing the information using standard XML interfaces. We transformed the extracted match events into MPEG-7 files for each match.

In our system, there is a user interface for querying the matches. We use XQuery for searching the MPEG-7 files. To include XQuery into our system, we use XQEngine library that is full-text search engine for XML documents, utilizing XQuery as its front-end query language.

There are minute based, player name based, team based, match based and event based search options in our system. User can select one of them or a combination of them. Since there are multiple search options and we are using XQuery, we implemented a module that dynamically generates XQuery according to search options. This module adjusts the necessary joins in an efficient manner.

After the search operation, the user can select one of the results, and if a match video is associated with that match, its video is shown at the top of search results. While playing the video, we use Java Media Player API, a portion of the Java Media Framework (JMF) that enables audio and video within applications.

In Figure 4, Corner events that Inter team has taken against Liverpool are shown. The result has shown the event name, minute of the event, the player who takes the corner and the match result.

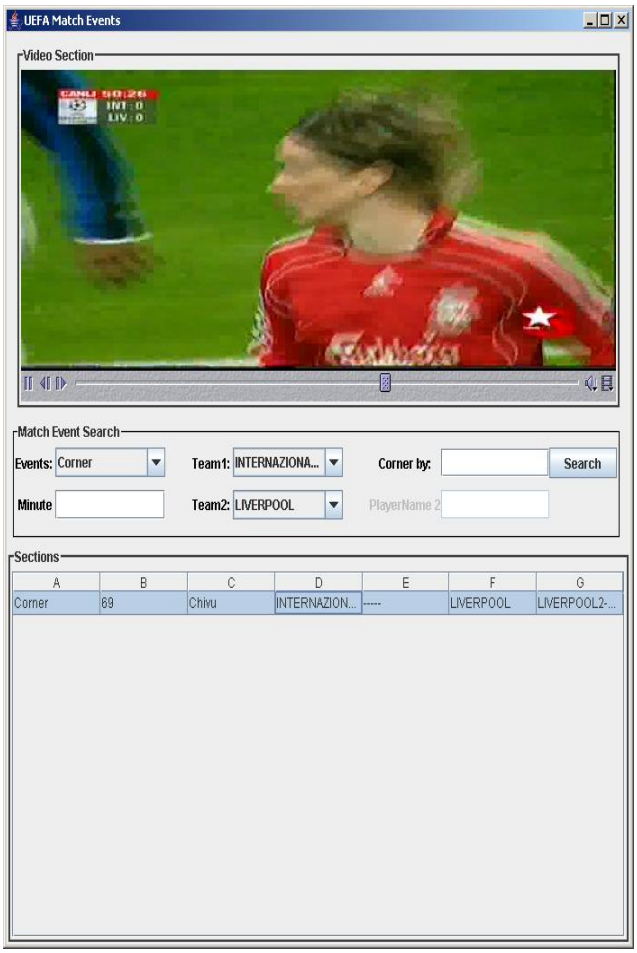

Figure 4: Displaying results for event-based querying

Figure 5 illustrates querying of all events that player Gerrard was involved in games where Liverpool and Internazionale are opponents. Goal Position at 25 minute is selected and that event is shown at the top of search results.

VII. CONCLUSION AND FUTURE WORK

We present an MPEG-7-based approach to information extraction in soccer domain that targets the automatic generation of MPEG-7 files from match reports. We choose the MPEG-7 standard for our match corpus to make the system more interoperable.

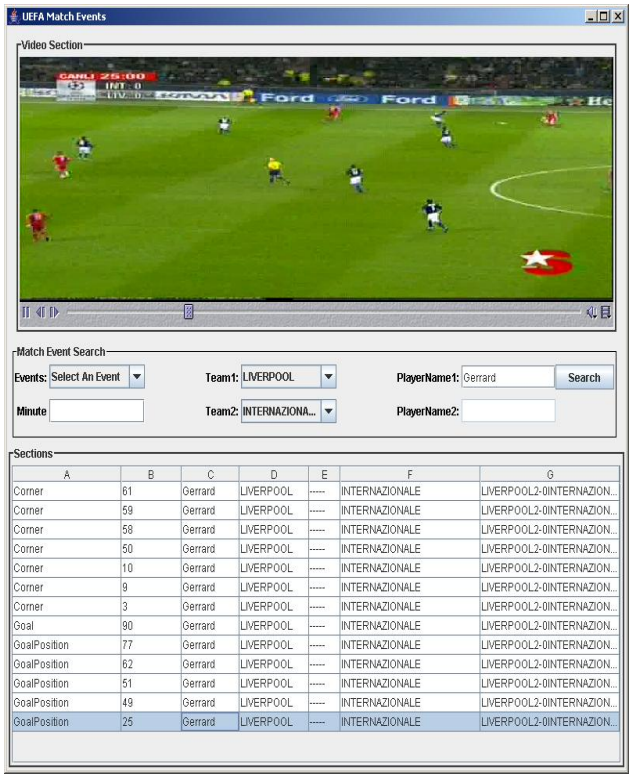

Figure 5: Player-based querying

A web crawler is used for downloading the match reports from UEFA web site. These match reports are annotated using regular expressions. According to the hand-written rule set, match events are extracted and converted to the MPEG-7 standard.

Finally, we adapted an interface for querying match events over MPEG-7 files by using XQuery Language. However since there are several kinds of query types we need a module that dynamically prepares XQueries according to search criteria.

Future work includes the replacement of hand-written rules with machine learning algorithms. We plan to learn the template rules for each soccer event from the match summaries. In this way the system will be more flexible and it will be easy to adapt it to a more variety of game narrations. Since we have a modular architecture, after labeling match reports we associate them with events and run information extraction algorithms on that match reports corpus.

ACKNOWLEDGMENTS

This work is partially supported by The Scientific and Technical Council of Turkey Grant ‘‘TUBITAK EEEAG-107E234.

REFERENCES

[1] Jan Nemrava, Paul Buitelaar, Vojtech Svatek and Thierry Declerck, “Event Alignment for Cross-Media Feature Extraction in the Football Domain” In: Proceedings of WIAMISS'07, Santorini, 2007, IEEE Computer Society.

[2] Paul Buitelaar, Thomas Eigner, Greg Gulrajani, Alexander Schutz, Melanie Siegel and Nicolas Weber, “Generating and Visualizing a Soccer Knowledge Base”, In Proceedings of the EACL’06 Demo Session, Trento, Italy, 2006.

[3] Changsheng Xu, Jinjun Wang and Yifan Zhang, “A Novel Framework for Semantic Annotation and Personalized Retrieval of Sports Video”, IEEE Transactions on Multimedia Vol 10(3) 421-436.

[4] Changsheng Xu, Jinjun Wang, Kongwah Wan, Yiqun Li and Lingyu Duan, “Live Sports Event Detection Based on Broadcast Video and Web-casting Text”, In Proceedings of the 14th annual ACM international conference on Multimedia, 2006, pp.221-230.

[5] Catherine Dolbear and Michael Brady “Soccer Highlights generation using a priori semantic knowledge”, In Proceedings of International Conference on Visual Information Engineering, 2003.

[6] Milena Yankova and Svetla Boytcheva, “Focusing on Scenario Recognition in Information Extraction”, In Proceedings of EACL, 2003, pp.41-48..

[7] http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm