# The shape of the game

Danilo B Coimbra*, Tácito Tiburtino†, Alexandru C Telea‡ and Fernando V Paulovich§¶
*Computer Science Department - Federal University of Bahia – UFBA, Salvador, BA, Brazil
†Federal University of Alagoas – UFAL, Arapiraca, AL, Brazil
‡University of Groningen, Groningen, The Netherlands
§Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada ¶ICMC – USP, São Carlos, SP, Brazil

***Abstract*—The development of multimedia and network technologies strongly increase the interest on Internet broadcasting or streaming services, especially for soccer games. An example is the 2014 World Cup soccer tournament that registered record-breaking audiences worldwide, providing attractive alternatives to traditional TV viewing. The confluence of video streaming and computational resources opens up many possibilities for applications such as the online detection of interesting events, strategy analysis, or statistics comparisons. While much research targets algorithms to detect match statistics, strategy, retrieval, and indexing, the problem of presenting such information to users is much less studied. This paper proposes a simple but effective visual metaphor to help users browse and get insight into sports matches, with a focus on soccer games. We extract video segments, based on audio and metadata, identifying the main events of a game. Next, we use such events to define a visual representation that preserves the time-order of the video sequence, highlighting the most important events. Our visual representation enables the quick finding of the main events, allowing users to improve navigation when exploring a match, and also provides a way to evaluate the quality of a game or entire tournaments. We demonstrate our approach by applying it to several matches of 2014 World Cup, analyzing its knockout stage and comparing the final match in six different languages.**

## I. Introduction

The development of multimedia and network technologies provides an increasing presence of video content over broadcasting and streaming services. Recently, the 2014 World Cup soccer tournament registered record-breaking audiences all over the world [1], becoming the most accessible edition of a soccer tournament in history, reaching up to 5.9 billion screens worldwide [2]. In line with this development, there is a steady increase of interest in soccer video analysis [3] involving multimedia information retrieval, video indexing and processing, video semantic analysis, and video visualization.

Different kinds of insights and facts can be found using soccer video analysis [4], [5]. Users who wish to see important or interesting events, also known as *highlights*, are best served by match summarization techniques [6], [7] or tools [8]–[10]. Users who want to see the team strategy by, *e.g.,* tracking the ball and/or players in the field, are best served by tactical-information systems [11]–[13]. Systems that compute match statistics provide a quantitative game analysis including red and yellow cards, ball possession, shots on target, and goals [14], [15]. Users who want to find similar matches or match fragments are best served by search-by-example systems using video retrieval and indexing techniques [16], [17].

All above systems provide efficient algorithms for data mining and analysis. In contrast, much less effort has been invested in the *presentation* of the produced information. For instance, in match summarization content-based methods extract information from many sources (audio, image and/or text) to find key events. Such events are typically presented in a text-only list or table form. Such visual metaphors are not particularly easy to use or appealing to casual users (soccer fans). For instance, they do not help readily answering questions like “Which events are interesting in this match?”; “Are there any polemic or controversial events?”; “Which event is the most important?”; “When do the main events occur in the match?” Similar problems occur, even more prominently, for tasks related to comparing several matches or analyzing entire tournaments.

We propose a simple, efficient, and effective visual metaphor to help soccer fans to answer the above questions and also to browse and get insight into different matches and tournaments. For this, we extract video segments, or *video skims*, based on audio and metadata, so that the main events are found according to the narrator’s emotion. Next, we use the skims to create a visual representation that preserves the video’s temporal sequence but also highlights the key events. The proposed visual representation supports an easy finding of events of interest and allows users to quickly navigate to, or between, them to explore a match. Additionally, our visualization supports exploring matches at coarser scales, such as evaluating the quality of a match or parts thereof and comparing different matches or tournaments. Our visual design let casual users quickly browse and instantly choose which highlight to watch based on its importance level. Another advantage is to change the level of summarization (quantity of highlights), enabling different levels of video browsing.

## II. Related Work

Due to the enormous popularity of soccer, several solutions have been proposed to extract insights from soccer videos. Visualization-wise, the main such approaches target the presentation of statistical [14], [15], [18], [19] and tactical [11]–[13] information. Wongsuphasawat and Gotz explore the performance of Manchester United’s soccer season creating a visualization that shows pathways with good (wins) and bad (losses) outcomes [18]. Improving upon traditional soccer ranking tables, À Table! provides temporal navigation in two different views [19]: dynamic animation over the rows of the ranking table and a transient line-chart of team ranks to visually explore teams performance in a championship. Khacharem *et al.* use animation to show time-dependent data like motion and trajectory in a single display [15]. They also note that non-expert users prefer simpler static visualizations while experts prefer the more complex dynamic ones. SoccerScoop proposes two visualizations (field and player views) to show and compare the dynamics of distinct players, targeting users such as team managers [14]. Based on this tool, managers were able to determine strategic insights, like if a given player plays better on the road or at home. Similar insights and use-cases were presented by Lucey *et al.* who analyze spatiotemporal ball-tracking data from English Premier League matches [11]. They showed how hypotheses such as "win at home and draw away" can be checked based purely on the video data. Spatiotemporal tracking of player positions was used to detect team-formation patterns associated to match events [12]. SoccerStories conveys tactical and statistical facts in a single view, showing player actions and ball shots at the same time of the game and individual player statistics [13]. Although some of these studies analyze image sequences to extract scene characteristics [11], [12], [15], this would require a quite high computational effort, opposing to audio track features that require less computation, memory and can even be processed on a local set-top box [20].

Besides the systems above that provide tactical and statistical facts to expert users, other systems have been designed for casual users for video summarization and content retrieval. In general, video summarization aims to detect typical events of interest, such as foul, goal, shot, corner, offside, as the main events of a match, also called highlights [6]. Additional events like competition intensity or emotional moments can be captured from frame sequences [7], automatically generated from audio and video descriptors [21], or can be identified using social media (Twitter) based on sentiment analysis [5]. To overcome the lack of visual representations of previous works, other authors proposed different visual summarization approaches besides the traditional clip/frame grid. Examples involve, mosaics of highlights based on ball tracking and camera movement compensation [8]; hierarchical representations using a tree-based visualization for non-sequential navigation [22]; friezes describing salient moments based on motion content analysis [23]; and video players with timeline-based visualizations [10], [24], [25]. Apart from the mosaics, which has no interaction, tree-based and timeline-based methods are time-consuming approaches. When time is not the bottleneck, and the user wants a particular event, content retrieval systems also offer personalized events [16], [17].

In summary, while many efficient and effective methods for video summarization and event extraction exist, little attention has been put into *presenting* the derived information in an easy-to-understand and easy-to-use way to the casual watcher. We next present a soccer-video summarization approach based on a simple and easy-to-understand visual strategy. Our target audience is casual fans that watched (or not) a match previously. Additionally, we provide statistical information that can be easily read by either casual and more expert users.

## III. Sports Video Summarization

Sports fans are used to revisit important events of a match even if they had already watched it earlier. Hence, summarization is highly useful, both for revisiting matches' highlights or for a fast preview of the critical events in a new match. We summarize sports videos and provide a metaphor to speed-up video browsing by a two-step approach (see Fig. 1). In the first step, the most important or critical events, the highlights, are found based on (a) the sound extracted from the actual video and on (b) the metadata retrieved from the Internet that contains match statistics such as goals, cards (yellow/red), and substitutions. In the second step, the video is segmented, based on the captured highlights, into several small video segments or video skims, that are used to create the visual depiction of the match. We use video skims to build our visual metaphor because (i) they present audio and motion elements that enhance the expressiveness of information, and (ii) they are more entertaining to watch than a group of keyframes [9]. The skims length is defined a priori, as time duration (detailed further), and the generation process is compatible with excerpt selection (picking-up relevant events), commonly used for event-based approaches. Finally, users can interact with this depiction to focus on the most relevant events, get more insight on particular moments or match phases, and compare several matches. These steps are detailed next.

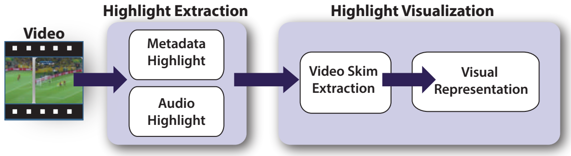

Fig. 1. Visual summarization pipeline. Highlights are extracted from the video representing important events, and the visual representation is arranged by combining video skims of each captured event.

### A. Detection of audio highlights

The soundtrack of the main broadcasted sports, soccer, in particular, consists mainly of foreground comments (which contains the narrator's speech) coexisting with background noise (which includes audience cheering, applause, and the referee whistle). Audio features to analyze such scenario are divided into physical properties, spectral and temporal characteristics, and perceptual properties, describing the perception of sounds by human beings [17]. As we aim to aid casual users, we chose the loudness (sound volume) perceptual feature, since it is a common way to describe highlights [26]. For this, we first divide the audio signal into short sub-segments, called *clips* [27]. A clip consists of a certain number of partially overlapping audio-frames, depending on the sampling frequency. In our work, each clip is a one-second time interval in a mono channel, sampled at $44.1$kHz, resulting in $86$ audio-frames, each having $1,024$ consecutive audio samples with $50\%$ overlap (512 samples) between consecutive frames. The loudness $l_{frame}(k)$ of a frame $k$ is given by the RMS (Root Mean Square) of the audio signal magnitude, *i.e.*

$$l_{frame}(k) = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} y_{k,n}^2}, \tag{1}$$

where $\{y_{k,n}\}_{0\leq n<N}$ is the set of $N = 1,024$ audio samples of frame $k$. For each clip $c$, its audio loudness $l_{clip}(c)$ is computed as the mean value of all frame loudnesses in the clip, *i.e.*

$$l_{clip}(c) = \frac{1}{86}\sum_{k\in c} l_{frame}(k). \tag{2}$$

Based on these data, we next define highlights as those clips with the largest loudness in the video that are not neighbors (given a window of size $2t$) of a clip with a larger loudness. We set $t = 7$ seconds since we empirically observe that it is unusual consecutive highlights to occur in less than 7 seconds in a soccer match. Nevertheless, it is not an issue to have more than one highlight in a 7 seconds window. Since the skim duration is set to the same size of such window, we will have skims with multiple highlights. Not missing highlights. Separately, the number of found highlights is limited by the user according to the display area available or considering a minimum value of loudness (as discussed next in Sec. IV).

### *B. Detection of metadata highlights*

Besides finding highlights by analyzing the video content, we can use alternative information sources, such as metadata. Metadata is a detailed labeled and annotated data together with precise temporal and spatial information and can enhance semantic analysis as it can provide valuable descriptions of match events [16]. In our work, we used metadata from an open-source soccer API describing all 2014 World Cup matches (see Sec. IV-A). This API provides descriptions and timestamps for three different event types (goals, substitutions, and cards). Goal events also include the scoring player's name. Substitution events include the entering and exiting player names. Card events indicate the card type (color) and involved player. One highlight is extracted per event.

We next show how the metadata+sound highlights are used both for segmenting the video and for creating the visual representations of key events during a match.

### *C. Constructing the visual representation*

To show the highlights, we use the video skim abstraction rather than traditional keyframes, due to the higher summarization power that video segments have as opposed to static images since audio and motion elements enhance both the expressiveness and conveyed information [9]. In our work, a video skim is a segment centered on a detected highlight, lasting $2t = 15$ seconds (Sec. III-A). Making a skim's time length equal to $2t$ guarantees that no key highlight is lost and that each highlight is mapped to a single video skim. Frequently, importance level (or interest level) is associated with each highlight [24], [28]. As such, we resize the extracted video skims to reflect the *importance* of the highlights they represent. Below, we first discuss how the importance is computed, based on the audio signal and/or video metadata. Next, we discuss how video skims are resized and assembled to produce the final visualization.

**Audio highlights importance:** The importance $s^A(v_i)$ of a video skim $v_i$ extracted considering the audio highlight corresponding to the clip $c_i$ (Sec. III-A) is computed based on the loudness of $c_i$ as

$$s^A(v_i) = \left[\left(\frac{l_{clip}(c_i) - \rho_{min}}{\rho_{max} - \rho_{min}}\right) + 1\right]^2, \tag{3}$$

where $\rho_{max}$ and $\rho_{min}$ are the maximum and minimum loudness values of the entire video, respectively. The importance $s^A(v_i)$ ranges between 1 and 4, with larger values indicating more important audio events.

**Metadata highlights importance:** As with the audio highlights, an importance $s^M(v_i)$ is calculated for each video skim $v_i$ that is extracted from the metadata highlights. In contrast to audio highlights (Eqn. 3), the importance of metadata highlights is computed as

$$s^M(v_i) = \begin{cases} 1.0 & \text{if} \quad type(v_i) = substitution \\ 2.0 & \text{if} \quad type(v_i) = yellowcard \\ 3.0 & \text{if} \quad type(v_i) = redcard \\ 4.0 & \text{if} \quad type(v_i) = goal, \end{cases} \tag{4}$$

where $type(v_i)$ gives the type of metadata event $v_i$. The importance values in Eqn. 4 were chosen empirically, with goals as the most important and substitutions as least important. Note also that both $s^A$ and $s^M$ take values in the same range $[1, 4]$.

**Combining overlapping video skims:** To compose the final visualization, the video skims corresponding to audio and metadata events are combined. This merges salient event information (such as goals or cards) with information that captures the narrator's excitement and polemic/controversial or even uncommon events (as captured by audio highlights), thereby presenting to the user a more complete match overview. If two video skims $v_i$ and $v_j$, corresponding to events extracted from the audio, respectively metadata, are found to overlap in time, we merge them into a single skim $v$ whose importance $s(v) = (s^A(v_i) + s^M(v_j))/2$ is set to the average of the importance of the merged skims, and whose time extent $v = v_i \cup v_j$ corresponds to the union of the two skims.

**Visual layout:** After creating the combined audio-and-metadata video-skim stream $V = \{v_i\}$, we resize each skim $v_i$ in $V$ to show its importance $s(v_i)$. The main idea is to make skims larger for more important events, so that users see these saliently in the resulting visualization. Additionally, we limit the size of the smallest video skim (corresponding to $\min_{v\in V} s(v)$) to 50 pixels, to ensure its visibility, and the size of the largest video skim (corresponding to $\max_{v\in V} s(v)$) to four times the minimal size, to limit the utilized screen space.

Once skims are resized, we organize them sequentially along the $x$ screen axis, which encodes the video timeline. In detail, skims are vertically centered along this timeline, and horizontally aligned concerning their temporal order, so that consecutive skims $v_i$ and $v_{i+1}$ have a small overlap of few

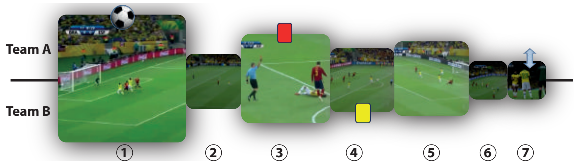

Fig. 2. Visual representation resulting from our technique. The skims $(1, 3, 4, 7)$ represent events detected using metadata. The other skims $(2, 5, 6)$ were detected using audio data only.

pixels. We use this overlap to further emphasize the skims' importance. For this, skims are rendered back-to-front in order of increasing importance, so that more important skims slightly overlap less important adjacent skims. By adding a faint shadow around the borders of each skim, this pseudo-depth effect further emphasizes important skims (large, overlapping, and appearing in front) as opposed to less important skims (small, overlapped, and appearing in the back). Finally, we also map skim importance to the saturation of the respective video skims, thereby making important events more colorful and less important more grayish. Overall, the above four cues (size, overlap, shadows, and saturation) jointly help observers in quickly finding the most important skims, and also locally sorting skims by relative importance.

The three types of information provided by metadata (goals, substitutions, and cards) are mapped as small icons over a video skim.The icon position on the skip (top *vs* bottom) shows if the event relates to team A or B respectively. Figure 2 illustrates our video summarization. Here, event **(1)** represents a goal for team A. Events **(3)** and **(4)** show a red card for team A and a yellow card for team B, respectively. Event **(7)** is a substitution for team A. Events **(2)**, **(5)**, and **(6)** were detected using the audio data only, and represent other types of events.

We use the above skim layout to construct two separate timelines, one for each of the two match halves. For matches having extra-time, we add a third timeline below the two. When all skims do not fit in the screen width, a slider is shown allowing the user to navigate through all skims in any timeline. This makes our layout independent of the screen width.

### D. Computational Complexity

The computational cost of our method follows the visual summarization pipeline (Figure 1). For the highlight extraction, the computational complexity is O($N + L$), where $N$ is the number of clips extracted from the video and $L$ is the number of metadata entries. Since $N \gg L$, the final complexity of this stage is O($N$). For the highlight visualization, the computational complexity is O($M$), where $M$ is the number of detected highlights. In the worst case scenario $M = N$ (every clip is a highlight), therefore, our approach is linear given the video size and the visual representation can be constructed in real-time as the video is received.

## IV. Results

### A. Datasets

Finding publicly available sports datasets which include soccer metadata is not easy. To overcome the locked-nature restrictions of proprietary sports datasets, we based our work on matches from the World Cup 2014 and UEFA Champions League 2015. Statistical metadata was collected manually from the FIFA official website. Corresponding video data was obtained separately from different Internet sources and includes matches recorded at different resolutions and produced by various broadcasters.

We next describe the analysis process performed on these data, starting with the video stream analysis and ending with the usage of our summary visualization.

### B. Visual exploration

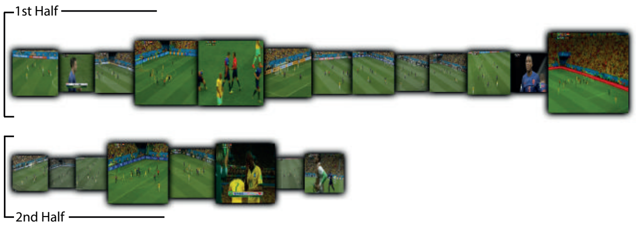

Fig. 3. Visualization constructed from audio highlights only. The match is summarized based on the narrator's emotion and audience excitement.

**Audio-based exploration:** As a first element, our visualization allows users to interpret the match statistics and the narrator's emotion and audience excitement. When constructed only based on highlights extracted from audio data, the visual representation will capture the emotion and excitement as a function of the sound loudness. Figure 3 presents the visual outcome for the Brazil-Netherlands match for the third place in the 2014 World Cup considering audio highlights only. The audio is in the English language. The presented visualization allows one to navigate temporally through the match highlights and easily get the most exciting moments of a match. In the first half of the match, three main events are detected, two refer to goals of Netherlands, and one is a clear chance of a goal for the Brazilian team. In the second half, two main events are detected, a yellow card for Brazil, resulting from a penalty simulation, and another goal for the Netherlands.

To provide more insight, and support answering additional questions, we provide user interaction, as follows. A mouse-over movement on a video skim sets the image used to depict the video skim to the time moment corresponding to the mouse $x$ position. This way, moving the mouse to the left or right plays the video skim forward or backward respectively. A left-button click on the timeline plays the underlying video skim. Finally, if the user wants to see more detail, a right-button click pops the underlying video skim in the screen center and plays it. When playback is ready, the skim pops back to its original space in the visualization.

**Metadata-based exploration:** In contrast to the above, visualizations constructed from metadata only support scenarios where one wants to focus on statistical information, such as goals, cards, substitutions, and to which team do these belong. Figure 4 shows such a metadata-only visualization of the same match as in Figure 3. As explained in Sec. III-C, icons indicate metadata events: a soccer ball shows a goal; yellow and red rectangles show yellow and red cards, respectively;

and colored arrows show player substitutions. The annotation locations (top or bottom of the skims) indicates if the event relates to the home team or away team respectively.

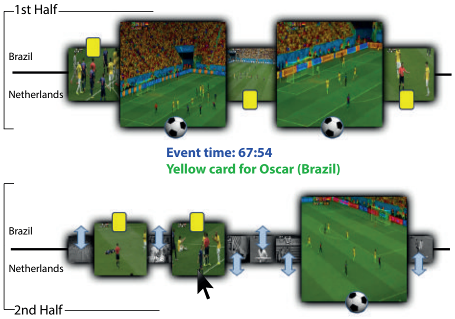

Fig. 4. Visualization constructed from metadata only.

In contrast to the audio-based visualization, frames shown in the video skims correspond now to the precise moments recorded by the respective events – *e.g.,* the two first-half goals, complaining moment of players after the first yellow card, and lastly the three Dutch substitutions in the second-half, where the camera zooms on the player going out (Fig. 4). Mouse-over interaction for metadata-based skims shows additional information about the event as text annotations – *e.g.,* the information about the yellow card received by Oscar (mouse over the second video skim, second half, Fig. 4).

**Combined audio+metadata exploration:** Audio and metadata based highlights can be combine to convey more information and/or provide more accurate answers to questions such as "Which are the most interesting events?" Figure 5 illustrates this multimodal (audio+metadata) approach combining highlights of Figures 3 and 4. The visualization's level-of-detail is controlled by a slider, filtering audio highlights according to a minimum value of loudness (Eqn. 1). All metadata-based skims are kept and multiplexed with audio-based skims, since metadata is typically of high accuracy and importance [4].

Figure 5a shows a situation where the user selected a minimum loudness of 18.9; the resulting visualization shows all metadata highlights and audio highlights louder than the given threshold – 25 highlights in total for the entire match. If the loudness slider is decreased to 18.6, 30 highlights are selected (Fig. 5b). Conversely, increasing the minimal loudness to 19.3 shows only 20 highlights (Fig. 5c)). Note how, irrespective of the selected level-of-detail, all goal events are kept *and* mapped as large (thus, important) video skims.

A separate unexpected finding is the crucial role of the audience's cheering in the background. In Figure 5a-c, the rightmost large video skim in the first half denotes the best chance Brazil had to score in the match. Brazilian audience was by far the largest in the stadium. This explains why some goal video-skims are not as large as expected. Separately, both yellow cards for the Netherlands were vigorously celebrated by the audience in the first half of the match, causing larger video skims than what a metadata-only visualization would have made. The same can be observed for the yellow cards on the second half of the match. The second yellow card has larger importance than the first one since it results from a penalty simulation and its importance is increased due to the audience's cheering. Finally, the use of audio highlighted some shots on goal by Brazil and faults made by the Netherlands, which are not present in the metadata. Overall, the above points emphasize the added-value of the combined audio+metadata visualization as opposed to audio-only or metadata-only approaches. Therefore, this is more a complementary than redundant combination aiming at guaranteeing that the most important events are identified despite cultural, language or narrator style aspects.

**Large-scale analysis:** For further insight, we used our approach to analyze and understand a portion of the 2014 World Cup. We consider the set of matches played by the finalists (Germany and Argentina) in the so-called knockout stage; matches where the losing team is out, resulting in more exciting dynamics. It is precisely this excitement that we want to recover using our proposed visualization.

Figure 6 illustrates the finalists' dynamics in the knockout stage, excluding the final match (discussed separately below). One direct insight is visible in the top-right: Argentina had to face two extra-time matches (*vs* Switzerland and the Netherlands) before the final while Germany only one (*vs* Algeria). Both matches of the round of 16 can be classified as "boring" matches during the regular time, but very exciting on the extra-times, probably because all goals were scored in the respective extra-times, 3 goals in Germany *vs* Algeria and one in Argentina *vs* Switzerland. Quarter-finals were quite balanced; both teams scored only one goal and both at the beginning of the first half, (largest-third video skim, first-half row, Fig. 6), which are the largest skims in both matches. Last but not least, the semi-finals were marked by the largest tournament score in the Brazil *vs* Germany game, easily visible in the respective visualization by the many large skims, present mainly in the first-half when Germany scored 5 out of its 7 goals. Interesting to observe is that the importance of these 5 goals increases during the time, probably indicating that the audience's hope (predominantly composed of Brazilians), given by the little cheering on the first goal, turns into complaints on the last one. An entirely different scenario is visible in the Netherlands *vs* Argentina match, where there were no goals during regular time, but Argentina won in the penalty shootouts. Although such match does not present goals on the regular or extra times, this was a very exciting match, reflected on the number of highlights captured by the sound. As we expect that our method captures the most exciting (and/or key) audio events, it is separately interesting to analyze the specificity of the soundtrack. Clearly, audio variations are expected between different broadcasters, as each one has different styles, commentator voices, and audio post-processing (*e.g.*, smoothing out noise). Still, one interesting question is: Which variations can be visibly measured considering not only different broadcasters but also different languages? Does the same match visualization come over differently?

To get insight in the above, we looked at the 2014 World Cup final match in six different languages/channels

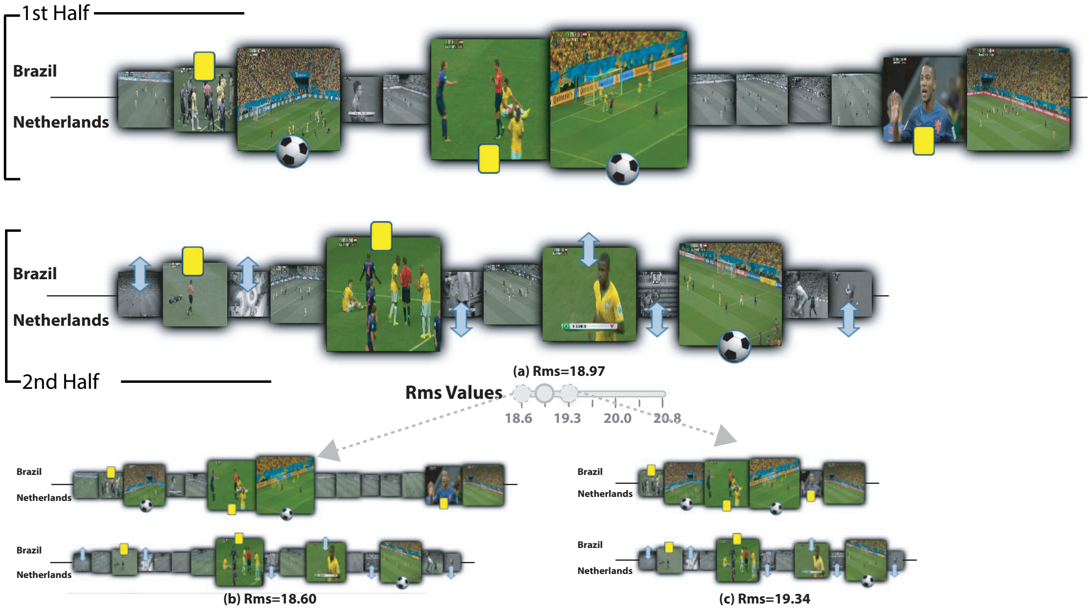

Fig. 5. Visualization using audio and metadata information with different levels of detail. The most important events are captured, some only by the metadata (substitutions) and some only by the audio (the goal chances).

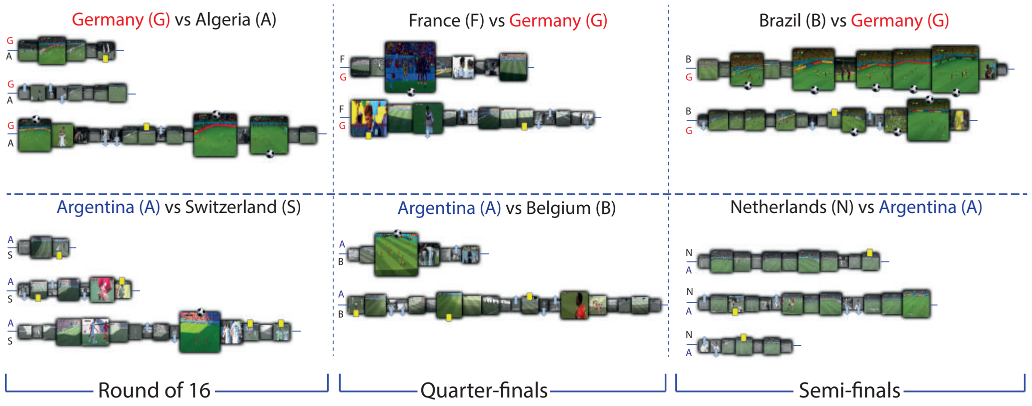
Germany (G) vs Algeria (A)
France (F) vs Germany (G)
Brazil (B) vs Germany (G)
Argentina (A) vs Switzerland (S)
Argentina (A) vs Belgium (B)
Netherlands (N) vs Argentina (A)
Round of 16
Quarter-finals
Semi-finals

Fig. 6. Comparing different matches of a tournament. Dynamics of finalists in the knockout stage of the 2014 World Cup.

(Arabic/Bein; English/BBC; French/TF1; German/Das Erste; Portuguese/Globo; and Spanish/Gol), see Fig. 7. From a quick look, it is evident that the displayed visualizations are very different and do not show the same dynamics, even though they all have the same metadata information. For example, the number of important highlights (large skims) is entirely different; also, given the same loudness threshold, visualizations show quite different numbers of highlights, meaning that the excitement level was considerably different. If we look at the less salient (thus, having smaller skims) half-time, we see that Arabic and Portuguese broadcasters were less excited than other broadcasters. In contrast, the English had a more muted second-half, and the French show fewer highlights in extra-time – notice that the French narrator gives no importance to the German goal in the extra-time. Excluding Arabic and English, the remaining broadcasts show a very excited second-half with many highlights, in particular, French and Portuguese. For the extra-time, Arabic and English prevailed with larger and many highlights. By clicking on the most salient skims in the respective visualizations, and listening to the respective videos, we find another interesting fact: when a goal is scored, Arabic, Portuguese, and Spanish yelled “Gooaal” in their respective languages, while the rest chose to call the scoring team or player name. Separately, Argentina had a disallowed goal event in the first half; the English broadcast was the only one to capture this event. As all those visualizations referred to the same match, the same metadata is mapped for each language/channel, providing a easy way to compare the excitement level for each metadata event in each language. This is an example of our tool being able to support answering questions such as: “Were there some controversial events?”.

## V. Visual Metaphor Comparison

To elucidate our contributions, we briefly discuss the advantages of our approach over existing methods developed

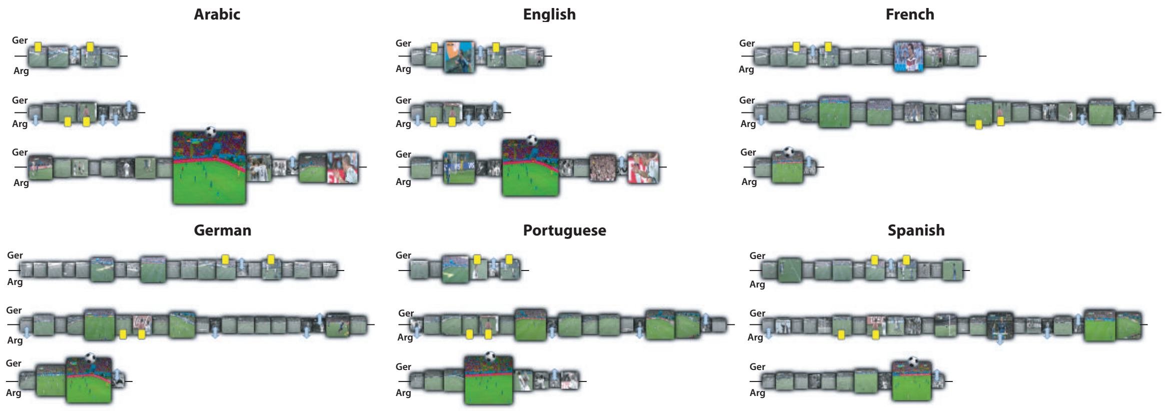

Fig. 7. Visualization of dynamics of 2014 World Cup final match in six different languages. Different cultures define different views of a match. Our visual metaphor can be used to easily verify and analyze cultural aspects.

to summarize soccer matches' videos that contain an explicit visual representation of the match.

Yow *et al.* [8] present a technique for highlight summarization that creates a mosaic-like single static image. Although an interesting approach, it is not easy to distinguish how action flows, especially if there are several players in the visual representation, which can cause occlusions. Besides, users prefer to watch video segments than see static keyframes for highlights [9], the reason we prefer to use video skims.

The hierarchical browsing approach [22] has some advantages, enabling non-sequential navigation and preserving browsing history through horizontal layers. However, the displayed video segments do not represent highlights, but equal length parts of the video, inducing to time-consuming navigation towards finding interesting events. Our approach, in contrast, extracts and presents highlights explicitly, speeding up the exploratory process. Nevertheless, it offers some features that can be used to increase or decrease the number of highlights, enabling different levels of video browsing, similar to the threshold interaction presented in [29].

Approaches that map additional information in the timeline slider of a video player to improve the traditional video browsing have been reported in the literature. Wang *et al.* [24] classify short clips of a video into three different types and color the slider bar according to them: blue for far-view, red for close-up and green for replays. Smits and Hanjalic [25] map pictographic icons and mouse-over text boxes with keywords above the timeline, enabling users to explore videos nonsequentially by jumping into those tagged segments. Chen and Vleeschouwer [10] employ vertical markers to represent in the timeline the exact moment of a highlight, mapping its importance to color saturation. Blanc *et al.* [23] developed friezes with colormaps to identify salient moments from motion content, such as zoom-in/out, zoom changes count, saturation, slow motion detection, and activity score. The main advantage is that timeline navigation is a known method for most users. However, even providing an enhanced timeline slider, they still present limitations; no user feedback about the visited positions is provided, and interactions in long video sequences can be time-consuming if the user does not know where to search [30]. In contrast, our method provides highlights in the exact moment they happen and, although we do not present a timeline slider, it preserves the time-order of the video sequence. Thus, the proposed image annotation, mouse-over metadata interaction, the highlight importance mapped to color saturation and skim size, can help users to decide if it is worthy or not to watch an event.

## VI. Discussion and Limitations

As already outlined in Sec. II, many techniques exist for finding relevant events in video broadcasts. Some of them are more advanced than our metadata+audio analysis, especially for highlight detection, such as the audio highlight detection in [31]. However, as also detailed earlier in Sec. V, such techniques (a) produce data which may be too technical for the typical casual user; and (b) when there is a visual presentation of such data, it is far from the ease-of-use and simplicity required by casual users. As such, our main contribution is to show that it is possible, by combining a descriptive analysis with a simple visualization, to offer non-trivial match-related insights to casual users in an easy-to-use way.

Several technical points can be further refined. Depending on the narrator's excitement, some goal events last more than one minute, adding successive videos skims related to this goal in both multimodal and audio-based visualizations. However, this limitation could depend on the point of view because a user can find such consecutive same-event highlights useful when understanding the entire match and its content. Cross-match language-dependent normalization can help here. Separately, if the match ends at the extra-time second-half in a draw, and penalty shoot-outs follow, our visualization can get biased by the excessive amount of metadata and/or audio intensity for that period. However, on the other hand, if a visualization shows a highly important set of end-match skims, this implies that such events are important ones, possibly dominating other events. A final remark is that our visualization is build to support computer-based devices (or tablets/mobiles), not considering interactive television applications.

Arguably our main limitation is the lack of extensive evaluation of the proposed technique. To gain more confidence regarding ease-of-use, a large-scale study involving casual users is needed. We note that such large-scale evaluations are very expensive to conduct, which is one of the causes why they are not present in the literature describing most video summarization applications.

## VII. Conclusions

We have presented an interactive visualization that allows casual users (sports fans) to browse quickly and easily a soccer match video, or a collection of matches, to find and get insights on events of interest, and also to compare several games, *e.g.*, in a tournament. For this, we analyze both audio and metadata streams associated with a match and reduce these to sequences of important events, annotated by the event type, corresponding video fragments that depict the event. The resulting event stream is next visualized and browsed interactively by users via a straightforward user interface that does not require learning complex visual or interaction metaphors. Thus, casual users can quickly explore a summarization of one or several matches. We show the added-value and way of working of our proposal by analyzing several matches from the 2014 World Cup knock-out phase.

Future work includes extending our visualization tool to handle real-time broadcast transmissions systems via multimedia streaming. This would enable us to capture (and analyze) not only the audio of the real-time broadcast but also the dynamically generated metadata provided by sports websites. As the implementation of our technique is simple and computationally scalable, we aim at applying it to a real-time match, to provide real-time summarization to users.

## Acknowledgment

We acknowledge the financial support of CNPq and Fapesp Brazilian agencies.

## References

[1] *Broadband Tv News*, 2014 (accessed May 01, 2018). [Online]. Available: https://www.broadbandtvnews.com/2014/07/15/fifa-world-cup-final-breaks-records-for-tv-broadcasters/

[2] *Channel World*, 2014 (accessed May 01, 2018). [Online]. Available: http://www.channelworld.in/news/fifa-world-cup-2014-most-accessible-history-ovum

[3] T. D'Orazio and M. Leo, "A review of vision-based systems for soccer video analysis," *Patt Recogn*, vol. 43, no. 8, pp. 2911–2926, 2010.

[4] P. Oskouie, S. Alipour, and A.-M. Eftekhari-Moghadam, "Multimodal feature extraction and fusion for semantic mining of soccer video: a survey," *Artificial Intelligence Review*, vol. 42, no. 2, pp. 173–210, 2014.

[5] S. Jai-Andaloussi, I. E. Mourabit, N. Madrane, S. B. Chaouni, and A. Sekkaki, "Soccer events summarization by using sentiment analysis," in *Int Conf on Comput Science and Comput Intellig*, 2015, pp. 398–403.

[6] M. Tavassolipour, M. Karimian, and S. Kasaei, "Event detection and summarization in soccer videos using bayesian network and copula," *IEEE T Circ Syst Vid*, vol. 24, no. 2, pp. 291–304, 2014.

[7] N. Nguyen and A. Yoshitaka, "Soccer video summarization based on cinematography and motion analysis," in *Proc. IEEE MMSP*, Sept 2014, pp. 1–6.

[8] D. Yow, B.-L. Yeo, M. Yeung, and B. Liu, "Analysis and presentation of soccer highlights from digital video," in *proc. ACCV*, vol. 95, 1995, pp. 499–503.

[9] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans Multimedia Comput Commun Appl*, vol. 3, no. 1, 2007.

[10] F. Chen and C. De Vleeschouwer, "Personalized summarization of broadcasted soccer videos with adaptive fast-forwarding," in *Intelligent Technologies for Interactive Entertainment*. Springer, 2013, pp. 1–11.

[11] P. Lucey, D. Oliver, P. Carr, J. Roth, and I. Matthews, "Assessing team strategy using spatiotemporal data," in *Proc. ACM KDD*, 2013, pp. 1366–1374.

[12] X. Wei, L. Sha, P. Lucey, S. Morgan, and S. Sridharan, "Large-scale analysis of formations in soccer," in *Proc. DICTA*, 2013, pp. 1–8.

[13] C. Perin, R. Vuillemot, and J.-D. Fekete, "Soccerstories: A kick-off for visual soccer analysis," *IEEE TVCG*, vol. 19, no. 12, pp. 2506–2515, 2013.

[14] A. Rusu, D. Stoica, E. Burns, B. Hample, K. McGarry, and R. Russell, "Dynamic visualizations for soccer statistical analysis," in *Proc. IV*, 2010, pp. 207–212.

[15] A. Khacharem, B. Zoudji, S. Kalyuga, and H. Ripoll, "Developing tactical skills through the use of static and dynamic soccer visualizations: An expert-nonexpert differences investigation," *J Appl Sport Psych*, vol. 25, no. 3, pp. 326–340, 2013.

[16] F. Sulser, I. Giangreco, and H. Schuldt, "Crowd-based semantic event detection and video annotation for sports videos," in *Proc. ACM CrowdMM*, 2014, pp. 63–68.

[17] H.-G. Kim, S. Roeber, A. Samour, and T. Sikora, "Detection of goal events in soccer videos," in *Storage and Retrieval Methods and Applications for Multimedia 2005*, vol. 5682. International Society for Optics and Photonics, 2005, pp. 317–326.

[18] K. Wongsuphasawat and D. Gotz, "Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2659–2668, 2012.

[19] C. Perin, R. Vuillemot, and J.-D. Fekete, "À table!: Improving temporal navigation in soccer ranking tables," in *Proc. ACM CHI*, 2014, pp. 887–896.

[20] P. Kathirvel, M. S. Manikandan, and K. Soman, "Automated referee whistle sound detection for extraction of highlights from sports video," *Int Journal of Computer Applications*, vol. 12, no. 11, pp. 16–21, 2011.

[21] A. Raventós, R. Quijada, L. Torres, and F. Tarrés, "Automatic summarization of soccer highlights using audio-visual descriptors," *SpringerPlus*, vol. 4, no. 1, pp. 1–19, 2015.

[22] M. del Fabro, K. Schoeffmann, and L. Böszörmenyi, "Instant video browsing: A tool for fast non-sequential hierarchical video browsing," in *HCI in Work and Learning, Life and Leisure*, G. Leitner, M. Hitz, and A. Holzinger, Eds. Springer Berlin Heidelberg, 2010, pp. 443–446.

[23] K. Blanc, D. Lingrand, and F. Precioso, "Singlets: Multi-resolution motion singularities for soccer video abstraction," in *Conf on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 66–75.

[24] L. Wang, P. Fonseca, and B. Zoetekouw, "User test of soccer highlights application," in *Proceedings of the 22Nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction, Vol 2*, ser. BCS-HCI '08. British Computer Society, 2008, pp. 241–244.

[25] E. Smits and A. Hanjalic, "A system concept for socially enriched access to soccer video collections," *IEEE MultiMedia*, vol. 17, no. 4, pp. 26–35, 2010.

[26] R. Leonardi, P. Migliorati, and M. Prandini, "Semantic indexing of soccer audio-visual sequences: a multimodal approach based on controlled markov chains," *IEEE T Circ Syst Vid*, vol. 14, no. 5, pp. 634–643, 2004.

[27] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis-using both audio and visual clues," *IEEE Signal Processing Magazine*, vol. 17, no. 6, pp. 12–36, 2000.

[28] M. H. Kolekar and S. Sengupta, "Bayesian network-based customized highlight generation for broadcast soccer videos," *IEEE Transactions on Broadcasting*, vol. 61, no. 2, pp. 195–209, 2015.

[29] H. Shubin, A. Divakaran, K. Wittenburg, K. A. Peker, and R. Radhakrishnan, "Assessment of end-user response to sports highlights extraction for personal video recorders," pp. 650 605–650 605–8, 2007.

[30] K. Schoeffmann and L. Boeszoermenyi, "Video browsing using interactive navigation summaries," in *2009 Seventh International Workshop on Content-Based Multimedia Indexing*, 2009, pp. 243–248.

[31] Y. Sun, Z. Ou, W. Hu, and Y. Zhang, "Excited commentator speech detection with unsupervised model adaptation for soccer highlight extraction," in *Audio Language and Image Processing (ICALIP), 2010 International Conference on*. IEEE, 2010, pp. 747–751.