Observing Attention

Davide Conigliaro[1,2], Francesco Setti[2], Chiara Bassetti[2], Roberta Ferrario[2], Marco Cristani[1,3], Paolo Rota[4], Nicola Conci[4], and Nicu Sebe[4]

[1] Università degli Studi di Verona, Strada Le Grazie 15, I-37134 Verona, Italy
[2] ISTC–CNR, via alla Cascata 56/C, I-38123 Povo (Trento), Italy
[3] Istituto Italiano di Tecnologia (IIT), via Morego 30, I-16163 Genova, Italy
[4] Università degli Studi di Trento, via Sommarive xx, I-38123 Povo (TN), Italy

Abstract. Understanding whether an event attracted the audience's attention and which moments were mostly enjoyable is a primary goal for sport and show business managers. OZ (Osservare l'attenZione – Observing attention) is an interdisciplinary, mixed-methods project that aims at developing a technology able to automatically detect at run time spectators' attention level, via the integration of microsociological analysis of human behavior into computer vision modeling and techniques. More specifically, we will show how it is possible to distinguish supporters of different teams by automatically detecting their liveliness in different moments of the match, even when they are mingled in the crowd. Moreover, we will show how, only by automatically detecting crowd's motion on the stands, it is possible to detect and annotate the most salient events of the match, like goals, fouls or shots on goal.

Keywords: spectator crowd, crowd analysis, spatio-temporal clustering,sport events, experiential marketing

1 Introduction

A crowd may generally be described as a definable group of people, but merely saying that a crowd is a large number of persons gathered closely together is not enough [1]. Even though such definition is correct and understandable, it is nonetheless oversimplified, as there are various types of crowd, so the notion of crowd is much more complex and requires a more detailed account, which is basically missing in the computer science community. On the other hand, at least two important studies in sociology tried to differentiate among different kinds of crowds based on the purpose of their existence. Momboisse [2] developed a system composed of four types: casual, conventional, expressive and aggressive. Differently, Berlonghi [1] identified eleven types of crowd, categorized according to the primary purpose of their existence. Among these categories, our work focuses on the *spectator crowd*, that is, people "interested in watching something specific that they came to see" [1], and this under a computer vision perspective.

The idea is to observe people while they are watching a public show, as in a sport arena, a movie theater, a classroom, a court, and recording and analyzing their activities. This scenario differs substantially from those analyzed by the typical crowd modeling techniques: due to *territoriality* principles, people are assumed to stay near a fixed location for most of the time, i.e., their seat [3, 4], while what is mainly being monitored in the crowd analysis literature are moving people. In addition, people here are assumed to have a strong relation with the event or contest they are watching[1], that becomes a kind of reference point, where the *focus of attention* [5] of the crowd is located, and around which the space is structured. In this new scenario, diverse techniques and applications can be developed, generalizing the videosurveillance context to the multimedia realms of the entertainment and the edutainment:

- **Spectators segmentation**: finding different groups of people among the spectators, for example the supporters of the opposite teams in a sport match or enthusiastic VS. annoyed spectators;
- **Excitement calculation**: in a given time interval, quantizing the level of excitement of some parts or of the entire crowd; this could be beneficial for example for marketing purposes.
- **Event segmentation**: segmenting diverse activities of the crowd (clapping hands, making a Mexican wave, heckling), and studying how these activities are related with the observed event (i.e. some people clap their hands when the favorite team scores a goal, or get excited when a foul is or is not signaled by the referee);
- **Augmented video summarization**: the spectator feedback, automatically recognized, may help in highlighting exciting or crucial events that should be included in a video summarization of the show;
- **Live show highlights**: the detection of attention and excitement of the spectators at rime time may help in decisions such as when to stop the live show and start a replay of a particularly salient event that just happened;
- **Comparative analysis of spectators**: various factors can be compared, like fans of different teams in the same sport [6], or fans of different sports [7], where spectators are arranged differently etc.;
- **Interpretation of crowd's intentions**: discriminating whether a display of crowd excitement is determined by a rejoicing VS. aggressive attitude, to foresee the subsequent crowd's behavior.

In the following, we will show how the first three aspects discussed above can be dealt with by using standard video surveillance strategies, whose adoption is motivated by social models. This interdisciplinary connection represents one of the most intriguing brand-new perspectives of the so-called Social Signal Processing [8, 9]. In particular, in this work we focus on a sport scenario, where people watch hockey matches.

[1] Such an assumption is commonsensical but, being more precise, one could say that the relation may vary a lot from individual to individual, ranging from very weak to very strong. Moreover, given mutual influences, the strength of such relation could be analyzed at the level of subgroups in the audience, rather than at individual level. Subgroups' relation with the event and its connection with level of attention and excitement are certainly topics that deserve to be studied in the prosecution of the present work.

Our framework has been evaluated on a dataset of 12 videos taken during the 2013 IIHF Ice Hockey U18 World Championship, for a total of 6 hours, showing qualitative and quantitative promising results.

In the rest of the paper we present the related literature in Sec. 2, then we explain our framework in Sec. 3, followed by preliminary results in Sec. 4; Sec. 5 draws some conclusions and future perspectives. In particular, the approach that has been developed will be applied to the hockey matches of the 26^{th} Winter Universiade, to be held in Trentino, December 11-21, 2013, with the purpose of singling out the most exciting matches, the most salient parts of each match and, possibly, to provide augmented video summarization of the video recorded matches.

2 Related literature

Spectator crowd analysis is a novel research area in computer vision, strictly related to the more general crowd analysis.

In computer vision crowd analysis focuses on the modeling of large masses of people, where the single person cannot be finely characterized, due to the small visual resolution and the frequent total occlusions. Therefore, many of the standard computer vision technologies as person detection, multi-target tracking, action recognition, re-identification, cannot be considered in their classical form. As a consequence, crowd modeling has grown with its own set of peculiar techniques (as multiresolution histograms [10], spatiotemporal cuboids [11], appearance or motion descriptors [12], spatiotemporal volumes [13], dynamic textures [14]), calculating on top of them flow information. Such information is then employed to learn different dynamics like Lagrangian particle dynamics [9], and in general fluid-dynamic models.

In computer graphics domain, crowd analysis is seen from several points of view, that are summarized by Jacques et al. in their work [15], where they proposed a taxonomy consisting of three important problems in crowd analysis: people counting/density estimation, tracking in crowded scenes, and crowd behavior understanding in a higher-level analysis, like the temporal evolution, main direction, velocity estimation, and detection of unusual situations.

Also, the sociological realm exhibits some relevant studies strictly related to spectator crowd analysis. First of all, the already cited work by Berlonghi [1], where one of the first definitions of *spectator crowd* is given.

Another relevant work is the one of Schweingruber and McPheil [16], where they built a model for characterizing “collective actions-in-common”, i.e., actions performed spontaneously by several people in coordination. This study, although not specifically centered on viewers, but rather on various forms of crowd, singles out seven dimensions for the analysis of crowd behavior: orientation (facing), vocalization (producing sounds other than words with mouth), verbalization (uttering words), vertical locomotion (movement of the body over the same point on the ground), horizontal locomotion (movement of the body from one point on the ground to another), gesticulation (meaningful bodily configuration based on fingers, hands, and arms movements mainly), and manipulation (using hands to applaud or to strike, carry, throw, pull, etc.). Such study is interesting from our point of view as it includes most of the behaviors that we intended to automatically detect in spectators crowds.

Other studies, as [17], have challenged the idea that a crowd can be seen as an undistinguished collection of individuals, highlighting how crowds can rather be segmented in subgroups of different size, composition and organization, based on their previous acquaintance, on common goals etc. Starting from this idea, one of the aims of spectator analysis is to automatically segmant the crowd.

Turning to spectator crowds, some scholars discussed how collective behavior, like applauding, is generated in contexts where a crowd is attending a public event, e.g. public speeches [18]. In Mann et al. [19] a mathematical model of the applause dynamic was developed, and compared to the dynamics underlying the diffusion of an epidemy. Regarding the sociology of spectators in sport events, many works have been produced, but most of them deal with violence in sport, as shown in [20], where the motivations that bring people to watch sports live are also discussed.

3 Our Framework and results

In our proposal, first of all, by leveraging on the existing sociological literature, we tried to extend and to specify with respect to the spectator crowd case the characterization of collective actions provided by Schweingruber and McPheil in [16], focusing in particular on attention and excitement.

By observing and analyzing spectators through the use of video analysis techniques (see [21]), some indicators of attention and involvement can be singled out:

- head/gaze toward the field VS. spectator or downward (e.g. to one's smartphone, camera, purse)
- high VS. low chin
- hands (open palm) or elbows on knees VS. folded arms or idle hands
- torso inclined toward the field or upright torso, straight shoulders and absence of abdominal contraction VS. reclined chest or curved shoulders and abdominal contraction
- both feet on the ground and moving body weight from one to the other gluteus VS. crossed legs
- pointing toward something for the benefit of a fellow spectator on the field VS. outside the field.

In the same way, behaviors indicating excitement, satisfaction and enjoyment may also be captured:

- hopping
- raising arm/s over the head or opening arms
- repeatedly moving arm/s
- applauding

- shaking fan-objects (like flags)
- putting hands in cone (or megaphone of some kind) in front of one's mouth.

When we have to deal not only with spectators, but with spectator crowds such behaviors are expressed in a mutually coordinated fashion. In other terms:

- head and body of the spectators composing the crowd are oriented in the same direction
- they share the torso posture and, more generally, they sit and stand synchronously
- they point together to something on the game field
- they applaud and clap synchronously
- they shake fan-objects in homogeneous direction (e.g. everybody toward right then left), largeness, and rhythm
- they hop homogeneously in highness of the jump and rhythm
- they enter in physical contact: hugs, pats on the back, etc.

In this work we have tried *a*) to capture collective orientation, vertical and horizontal locomotion by measuring flow direction and *b*) to attain synchronous gesticulation and manipulation by calculating entropy. At the same time, by using these criteria, subgroups can be individuated within the crowd.

As a first step, standard motion flow is computed on the image plane, extracting at each pixel direction and intensity. Then, assuming people as static [3, 4] and considering the size of people, flow information can be re-arranged into a grid of squared patches. On each patch n, at each time frame, we extract four measures: the first is the flow intensity $I(n)$, obtained by averaging over the flow intensity values of the patches' pixels; intuitively, this cue encodes how much movement characterizes a patch. The second cue is the flow direction entropy $E_{\mathrm{dir}}(n)$, calculated over the related flow direction values (opportunely quantized), it describes the kind of movement in the patch: high entropy values mean random directions, while low values address homogeneous movement in the patch. The last two measures are the x, y patch centroid coordinates. In other words, at each time step, each patch is described as a 4D point.

Then the patches (i.e., their 4D vectors) are organized into a partition, thanks to a Gaussian clustering with automated model selection, that operates on each single frame independently. In practice, having two patches in the same cluster at a particular frame means that those patches portray for a while a very similar visual activity. This similarity is then extended, accounting for the temporal dimension. Roughly speaking, we build a matrix which contains at position i, j the number of frames the patches i and j were in the same cluster. This pairwise spatio-temporal similarity is then exploited by a hierarchical clustering, to individuate regions of patches with a similar visual evolution.

In this way we obtain the spectator segmentation, which partitions the scene in regions where the behavior of the crowd is similar. Finally the algorithm was extended by adding the temporal information into the analysis, looking for nonrandomic spatio-temporal clusters; for this purpose, the Lempel-Ziv complexity was considered as proposed in [22]. This way, choral activities can better emerge, indicating for example groups of supporters of the same side. After this, with the adoption of entropic measures, the degree of excitement of such groups can be quantified as demonstrated in [23]. For each region r, a *local* level of excitement is estimated by computing the value:

$$Exc(r) = \frac{I(r) \times E_{\text{dir}}(r)}{E_{\text{int}}(r)^2} \tag{1}$$

over a short time interval (in the order of seconds); here, $E_{\text{int}}(r)$ is the entropy of the motion flow *intensities* at a given time step. The rationale of this measure is that we consider as an high excitement for a group of people an intense movement (high $I(r)$), with diverse directions (high $E_{\text{dir}}(r)$), computed in a coordinated fashion for all people belonging to that region (low E_{int}). Finally, the average of $Exc(r)$ over all frames is considered as the excitement cue in a given interval for the region r.

The event segmentation task is meant to highlight events that globally trigger the excitement of the spectator crowd, against periods in which the level of excitement is generally low. To such aim, the intensity and the entropy of all the patches are collected at each frame and averaged, obtaining a single pair of values. Replicating this process for all frames gives a 2D signal which can be quantized in an unsupervised fashion by Mean Shift.

After the quantization, looking at the mean values of each obtained cluster may serve to get insight on the kind of event being modeled. For example, clusters with high intensity and high entropy may be originated by an interesting event happened in the game.

4 Results

In order to test the proposed framework, we built a novel repository which consists of videos taken during the 2013 IIHF Ice Hockey U18 World Championship, partially played in Asiago from the 7th to the 13th of April 2013. In particular, two entire matches were recorded (Italy VS. Norway, Italy VS. Slovenia), each one by two cameras, mounted frontally at a distance of about 25 meters from the spectators' stand. All videos were manually labeled by highlighting the main actions of the game, especially the fouls, shots and goals. Italy VS. Norway ended 1-12, while Italy VS. Slovenia 3-2.

The spectators segmentation and excitement calculation results are shown in Fig. 1; the Norwegian stand is analyzed, in relation to a sequence of 3 minutes extracted from the first time of the Italy-Norway match. As shown in image b), we have 3 regions, one corresponding to the background (region 1), the other two (regions 2 and 3) focusing on the crowd. Looking at the dendrogram, one can see that the crowd regions are closer than the background, which is reasonable; the excitement level is shown as the color of the regions, highlighting region 3 (dark red) of highly excited people, continuously moving, clapping their hands, shaking flags and yelling; while region 2 shows people who are more quite, and in fact the zoomed image in the light red box of image d) shows a sitting spectator only

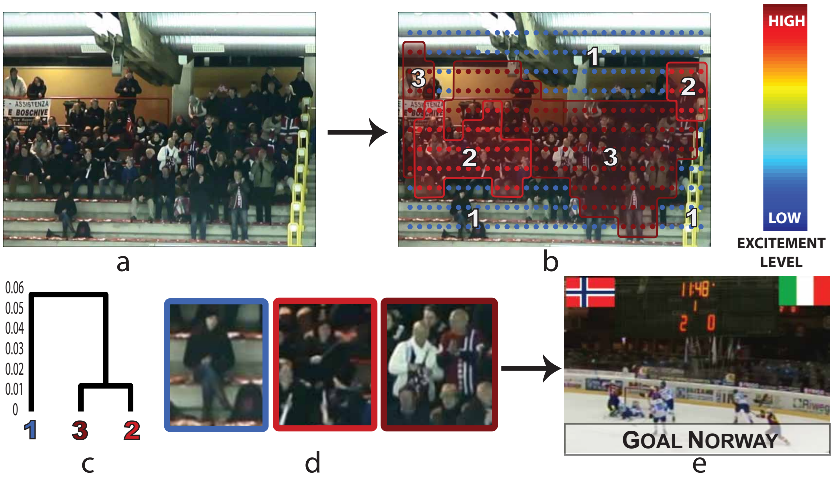

Fig. 1. Spectators segmentation and excitement calculation; a) an example frame of the sequence; b) spectator segmentation result, where the regions are colored considering their mean excitement level; c) dendrogram of the temporal clustering; d) zoomed images, highlighting the behavior of people of the different regions e) a frame of the match played in the considered interval.

shaking the flag. On the focused images, the first one shows a spectator of the background region (blue): this person moves very little for the whole duration of the video and doesn't exult for the goal.

In Fig. 2 we show the spectators segmentation and excitement calculation related to a sequence of 3 minutes extracted from the second time of the Italy-Norway match. In this case, we focus on a different area of the stands, where many Norwegian and some Italian supporters are blended. The sequence reports two goals, one for team. The segmentation gives surprising results, being able to distinguish 5 regions (4 plus the background). Regions 2 and 3 individuate Italian supporters, while regions 4 and 5 show Norwegian fans. The excitement calculation shows that Norwegian supporters are more energetic (at the end of the sequence the score was 5-1 for Norway) than the Italians. Excluding the background, the most quiet region is 2: probably, due to the mixing of the opposite teams, people prefer to be quiet not to offend fans of the other team.

The event segmentation result is shown in Fig. 3. Plot A shows how the two different spectators crowds get excited by different events. Norwegian spectators went crazy at the goal of Norway, while Italians when Italy scored a goal. To be noticed also the yellow box detected for the Italian spectators, in the moments immediately following the Norwegian goal, this is because Italians argued against Norwegian players. Plot B, instead, shows the results calculated on Norwegian spectators over the whole first time. We can see that the 4 goals are well detected as salient events by Mean Shift, but also another event wowed people, a great shot of a Norwegian player. The last yellow box in the strip shows the end of the first time, when the audience gets up and leaves the stand.

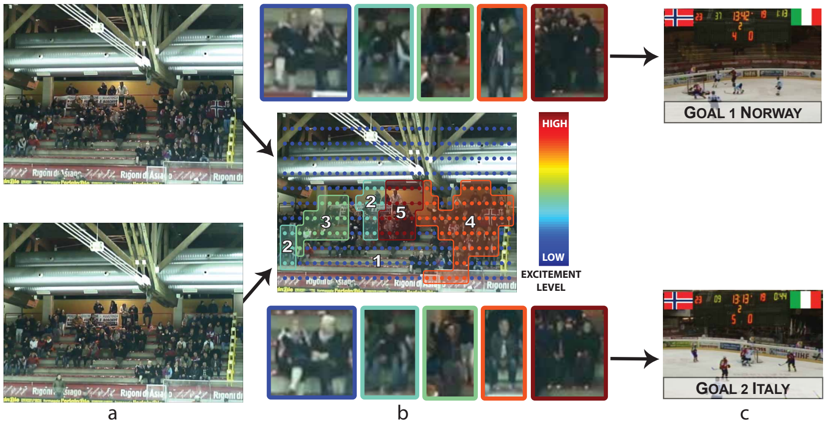

Fig. 2. Spectators segmentation and excitement calculation; a) two different frames of the sequence, the first extracted during the Norwegian goal, the second during the Italian goal. b) in the middle the spectator segmentation result, where the regions are colored considering their mean excitement level. Above and below zoomed images, highlighting the behavior of people in the different regions related with the goals of the different teams (Norwegians on top, Italians on bottom) c) the two goals.

With this preliminary study we showed the possible applications that can be developed in a spectator crowd scenario. In particular, we showed how spectators can be segmented on the basis of their behavior, how their excitement level can be inferred, and how the observed show can be segmented, by looking exclusively at the crowd activity.

5 Conclusions

The study of spectators crowd dynamics offers new perspectives in the crowd modeling field. In this paper we have performed a preliminary study, first of all reasoning on the possible applications that can be developed in such a scenario, and presenting effective implementations for some of them; in particular, we showed how spectators can be segmented on the basis of their behavior, how their excitement level can be inferred, and how the observed show can be segmented, by looking exclusively at the crowd activity. Much more can be done, by employing more sophisticated models: dynamic Bayesian networks may embed spatial and temporal reasoning in a unique model; gesture recognition, face detection and expression recognition may provide detailed cues to better understand the nature of the spectators activities, allowing the discrimination between supporting, heckling or just watching, absent in the present work. Further developments may be achieved by adopting different sensors, like microphones – recording also vocal and verbal signals - infrared and pan-tilt-zoom cameras.

An important theme to be inquired is the establishment of the ground truth for such kinds of scenarios. In this paper we have adopted a sort of "expert

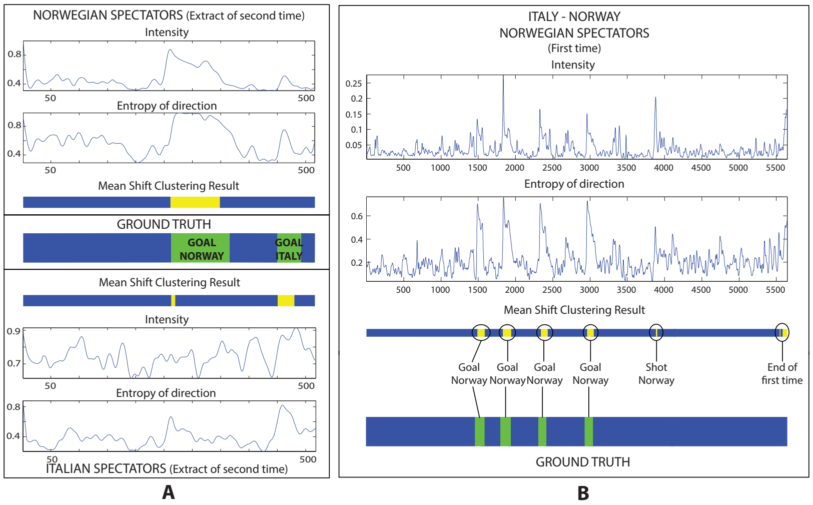

Fig. 3. Salient events detection. Here, the extracted flow intensity and entropy of flow direction of both Norwegian and Italian spectators are shown. The small bars show the results of Mean Shift clustering (the yellow boxes represent detections of salient events). These bars are compared to the ground truth (the bigger bar in the middle) where goals are indicated (green bars). Plot B shows the same results over the first time of the Italy - Norway match, by filming Norwegian spectators.

based ground truth", in that we have compared our findings with what had been explained in sociological theories, complemented with a preliminary ethnomethodologically oriented videoanalysis (see [21]). Alternatively, a more complete approach of this kind (expert) would be based on an ethnographic study: in that case the ground truth would be built on the basis of participant observation carried out by several ethnographers (team ethnography), doing fieldwork on the stands of an arena, stadium, amphitheater, etc. A completely different approach to ground truth would be to found it in a more "bottom-up" way, by asking directly to those belonging to the crowd, either exactly the crowd that was attending the recorded event, or, more generically, people that can report about an experience of participation to a public event as a viewer. Even in this case, there are various ways to implement such approach, ranging from structured questionnaires to in-depth interviews, to the collection of physiological data on a subset of the spectators through sensors measuring skin conductance, heart rate, temperature etc.

Notwithstanding all that have already been mentioned, of course privacy and ethical issues should also be taken more seriously into account in the nearest future developments of this study.

Bibliography

[1] Berlonghi, A.: Undestanding and planning for different spectator crowds. Safety Science **18** (1995) 239–247

[2] Momboisse, R.M.: Riots, revolts, and insurrections. C. C. Thomas (1967)

[3] Guyot, G.W., Byrd, G.R., Caudle, R.: Classroom setting: An expression of situational territoriality in humans. Small Group Behavior **11** (1980) 120–128

[4] Kaya, N., Burgess, B.: Territoriality. seat preferences in different types of classroom arrangements. Environment and Behavior **39**(6) (2007) 859–876

[5] Goffman, E.: Behaviour in Public Places. Free Press of Glencloe. Notes on the Social Organization of Gatherings (1963)

[6] Roadburg, A.: Factors precipitating fan violence: a comparison of professional soccer in Britain and North America. Brit. J. of Sociology **31**(2) (1980) 265–276

[7] Goldstein, J., Arms, R.: Effects of observing athletic contests on ostility. Sociometry **34**(1) (1971) 83–90

[8] Cristani, M., Murino, V., Vinciarelli, A.: Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space. In: CVPRW. (2010) 51–58

[9] Raghavendra, R., Del Bue, A., Cristani, M., Murino, V.: Abnormal crowd behavior detection by social force optimization. In: HBU. (2011) 134–145

[10] Zhong, H., Shi, J., Visontai, M.: Detecting unusual activity in video. In: CVPR. (2004) 819–826

[11] Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: CVPR. (2009) 1446–1453

[12] Andrade, E.L., Blunsden, S., Fisher, R.B.: Modelling crowd scenes for event detection. In: ICPR. (2006) 175–178

[13] Laptev, I.: On space-time interest points. Int. J. Comput. Vision **64**(2-3) (2005) 107–123

[14] Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: CVPR. (2010) 1975–1981

[15] Jacques, Junior, J., Raupp Musse, S., Jung, C.: Crowd analysis using computer vision techniques. IEEE Signal Processing Magazine **27** (2010) 66–77

[16] Schweingruber, D., MacPheil, C.: A method for systematically observing and recording collective action. Sociological Methods Research **27**(4) (1999) 451–498

[17] McPhail, C.: From clusters to arcs and rings: Elementary forms of sociation in temporary gatherings. Research in Community Sociology **suppl. 1** (1994) 35–57

[18] Atkinson, J.M.: Public speaking and audience responses: some techniques for inviting audience applause. In: Structures of Social Action. Cambridge University Press (1984) 370–407

[19] Mann, R.P., Faria, J., Sumpter, D.J.T., Krause, J.: The dynamics of audience applause. Journal of The Royal Society Interface **10**(85) (2013) 1–7

[20] McDonald, M.A., Milne, G.R., Hong, J.: Motivational factors for evaluating sport spectator and participant markets. Sport Marketing Quarterly **11** (2002) 100–113

[21] Heath, C., Hindmarsh, J., Luff, P.: Video in Qualitative Research. Analysing Social Interaction in Everyday Life. Sage, London (2010)

[22] Conigliaro, D., Setti, F., Bassetti, C., Ferrario, R., Cristani, M.: ATTENTO: Attention observed for automated spectator crowd analysis. In: HBU (in ACM–MM). Volume 8212 of LNCS. (2013) 102–111

[23] Conigliaro, D., Setti, F., Bassetti, C., Ferrario, R., Cristani, M.: Viewing the viewers: A novel challenge for automated crowd analysis. In: ICIAP Workshop. Volume 8158 of LNCS., Springer (2013) 517–526