Viewing the Viewers: a Novel Challenge for Automated Crowd Analysis

Davide Conigliaro[1,2], Francesco Setti[2], Chiara Bassetti[2], Roberta Ferrario[2], and Marco Cristani[1,3]

[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

Abstract. We focus on the automated analysis of *spectator crowd*, that is, people watching sport contests alive (in stadiums, amphitheaters etc.), or, more generally, people "watching the activities of an event [...] interested in watching something specific that they came to see" [2]. This scenario differs substantially from the typical crowd analysis setting (e.g. pedestrians): here the dynamics of humans is more constrained, due to the architectural environments in which they are situated; people are expected to stay in a fixed location most of the time, limiting their activities to applaud, support/heckle the players or discuss with the neighbors. In this paper, we start facing this challenge by following a social signal processing approach, which grounds computer vision techniques in social theories. More specifically, leveraging on social theories describing expressive bodily conduct, we will show how, by using computer vision techniques, it is possible to distinguish fan groups belonging to different teams by automatically detecting their liveliness in different moments of the match, even when they are merged in the stands. Moreover, we will show how, only by automatically detecting crowd's motions on the stands, it is possible to single out the most salient events of the match, like goals, fouls or shots on goal.

Keywords: spectator crowd, crowd analysis, spatio-temporal clustering

1 Introduction

Emerged as a Video Surveillance niche, crowd analysis has become in the last 10 years a separate topic of Computer Vision, embracing heterogenous applicative fields like public crowd management, space design, virtual and intelligent environments [19]. 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, the frequent total occlusions and the particular dynamics. 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 [20], spatiotemporal

cuboids [12], appearance or motion descriptors [1], spatiotemporal volumes [14], dynamic textures [15]), calculating on top of them flow information. Such information is then employed to learn different dynamics like Lagrangian particle dynamics [16], and in general fluid-dynamic models. The most important applications of crowd analysis are abnormal behavior detection [15], detecting/tracking individuals in crowds [13], counting people in crowds [3], identifying different regions of motion and segmentation [18].

In this paper, we focus on a novel applicative field for crowd analysis, centered on the modeling of the so called *spectator crowd* [2] (or viewers' crowd, as we have called them in the title). 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 [9,11], 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, that becomes a kind of reference point, where the focus of attention [7] of the crowd is located, and around which the space is structured. In classical crowd modeling no such clear reference point is present.

These two key elements build a context where diverse techniques and applications can be developed, some of which are listed in the following:

Spectators segmentation finding diverse groups of people among the spectators, for example the fans of the opposite teams in a sport match; attentive VS distracted students in a classroom; enthusiastic VS annoyed spectators at a theater play;

Excitement calculation in a given time interval, quantizing the level of excitement of some part or of the entire crowd;

Event segmentation segmenting diverse activities of the crowd (clapping hands, making a 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;

Anomaly detection given an expected crowd behavior, highlighting anomalous activities that may lead to dangerous situations, like fights, mass escapes, etc. in order to prevent them;

Comparative analysis of spectators various factors can be compared, like fans of different teams in the same sport [17], or fans of different sports [8], or the behavior of the same fans in different stadiums, 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 faced using Social Signal Processing methods [4], focusing on a sport scenario, where people watch hockey matches [1]. For the spectators segmentation and the excitement calculation issues, we use local flow information (position, flow intensity and direction), as input of a Gaussian clustering framework operating on the single frame. The spatial segmentations are then joined together along the temporal axis by a hierarchical clustering. The results are impressive, since it becomes possible to distinguish the different fan groups, even when they are merged; regions of activities indicating how much lively some supporters are can also be automatically found. In the event segmentation problem, we calculate global flow measures (intensity, entropy of the flow direction) at each frame, obtaining a 2D signal which is subsequently quantized by Mean Shift segmentation. This way, important events (goals, shots on goal) can be easily discovered.

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 our framework in Sec. 2, followed by preliminary results in Sec. 3; Sec. 4 draws some conclusions and future perspectives.

2 Our framework

In the following, we will detail the methodologies adopted for solving the *spectators segmentation*, the *excitement calculation* and the *event segmentation* issues (see Sec. 1).

2.1 Spectators segmentation and excitement calculation

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 [9,11] and considering the size of people, flow information can be re-arranged into a grid of N squared patches $\{x\}$. On each patch x, at each time frame, we extract four measures: the first is the flow intensity $I(x)$, 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 $E_{\text{dir}}(x)$, , calculated over the related flow direction values (opportunely quantized). The entropy is defined as

$$E_{\text{dir}}(x) = -\sum_{i=1}^{d} p(x_i) \log p(x_i) \tag{1}$$

[1] The last two aspects, in order to be studied seriously, imply the availability of a background behavioral model, which is not the case here; we thus leave such analyses to future studies. Finally, the augmented video summarization application implies multimedia aspects that cannot be dealt with here.

where d is the total number of directions, and $p(x_i)$ is the probability to have the direction i in the patch. The entropy $E_{\text{dir}}(x)$ describes the kind of movement in the patch: high entropy values mean random directions, while low values address homogeneous movement in the patch (a similar use of this entropic descriptor has been exploited in [5]). 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.

The segmentation occurs in two steps: first, a Gaussian clustering with automatic model selection [6] is applied on the values of all the patches in a given time frame. This way, an instantaneous grouping is inferred. This process is replicated for all the T frames.

At this point, a $N \times N$ similarity matrix is built, containing at entry i, j how many times patches i and j have been in the same cluster. Making the similarity matrix as a distance (computing the reciprocal) it is possible to perform single link hierarchical clustering, and to obtain the spectator segmentation, which partitions the scene in regions where the behavior of the crowd is similar (in terms of the measures quoted above).

For each region r, a *local* level of excitement is estimated by computing the value:

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

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.

2.2 Event segmentation

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. In this case Mean Shift has been preferred to the Gaussian clustering, since pooling together the signal values of an entire sequence leads to highly irregular distributions, proper for a non parametric treatment.

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.

3 Experiments

In order to test our 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 by two cameras, mounted frontally at a distance of about 25 meters from the spectators' stand. Each camera was pointing at an half of the whole stand, the zoom being fixed. Therefore, for each match we have two sequences, further divided in 3 as the times of the hockey play. This resulted in 12 videos at 30 fps, with a resolution of 640x480 pixels for a total duration of about 6 hours. 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 experiments have been partitioned in two groups. In the former, we focus on the spectators segmentation and the excitement calculation; in the latter we perform event segmentation (see Sec. 1). In all cases, a grid of rectangular patches of size 40×80 was built, with the patches overlapping for an half of their size, in both dimensions. Flow was computed on the entire scene each 10 frames, so we have 3 processed frames per second; after that, the flow direction was quantized in five values (up, down, left, right, none) where the fifth value corresponded to all those flow vectors whose intensity was inferior to a given threshold $I = 0.8$.

3.1 Spectators segmentation and excitement calculation

The whole footage was analyzed by temporal windows of 3 minutes length, overlapped by 10 seconds. For each window, we computed first the frame-based Gaussian clustering and subsequently the temporal hierarchical clustering. This way, for each window we get interesting spectators segmentations, clearly explaining the occurred events; for longer windows, the segmentation tends to discriminate solely the presence of the crowd against the background. Some segmentation results are shown in Fig. 1 and Fig. 2.

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 Figure 1b), 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; the other region, 2, shows people who are more quite, and in fact the zoomed image in the light red box of Figure 1d) shows a sitting spectator only shaking the flag. Of 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-

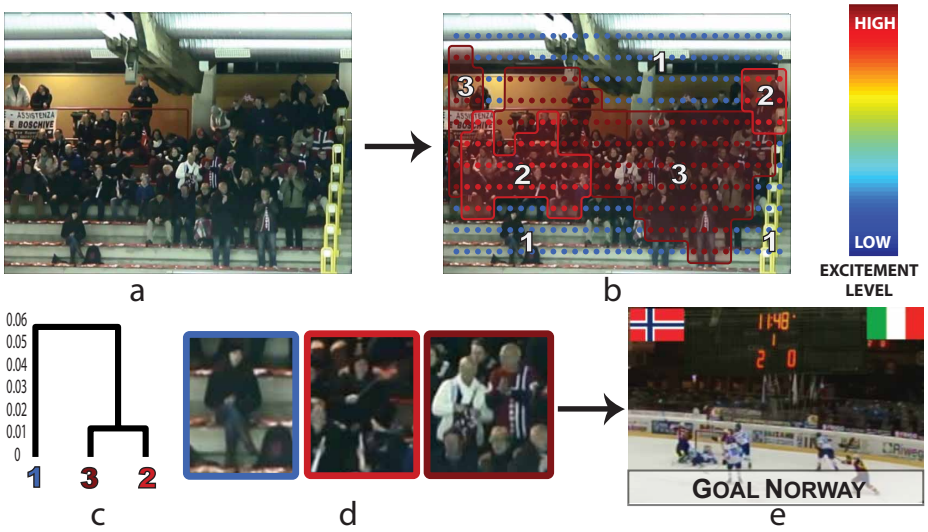

Fig. 1. Spectators segmentation and excitement calculation; a) an example frame of the sequence; b) spectator segmentation result, where the regions are colored considering their mean excitement level; c) dendrogram of the temporal clustering; d) zoomed images, highlighting the behavior of people of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

Norway match. In this case, we focus on a different stands area, 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.

3.2 Event segmentation

For the event segmentation, we analyze the video for the entire duration of a game time to identify the salient moments for the audience. All the 12 videos are analyzed by considering a time window of 2 seconds with 1 second of overlapping. The bandwidth parameter of Mean Shift was obtained experimentally, and is kept the same for each match. Depending on the choice of bandwidth, different actions of the game can be detected, such as goals or shots on. For the Italy–Norway match a bandwidth value of 0.181 was fixed and 0.1464 for Italy–Slovenia.

The obtained results show that in the Italy–Norway match, the most salient events detected for the Norwegian spectators are 16, including 11 goals scored

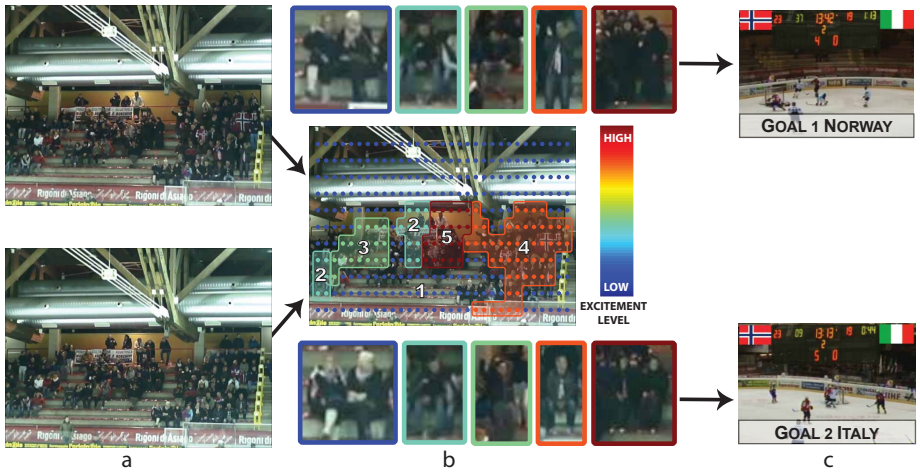

Fig. 2. Spectators segmentation and excitement calculation; a) two different frames of the sequence, the first extracted during the Norwegian goal, the second during the Italian goal. b) in the middle the spectator segmentation result, where the regions are colored considering their mean excitement level. 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.

and 2 spectacular shots, the other 3 are false positives caused by people arriving or leaving from the stands at the start or end of the game time. For the Italian spectators, instead, only 3 salient events were detected, 1 goal of Italy, 1 goal of Norway and 1 is a false positive caused by spectators leaving at the end of the match.

On the other hand, in the Italy–Slovenia match spectators are mainly Italian and two different stands are filmed, but the results are mixed. The salient events detected are 9, including 3 Italian goals and 1 nice Italian shot. The other 5 events are false positives always related to people arriving or leaving.

An example of these results is shown in Figure 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, quite as it was to be expected, 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 argue against Norwegian players.

Plot B, instead, shows the results calculated on Norwegian spectators over the whole the 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.

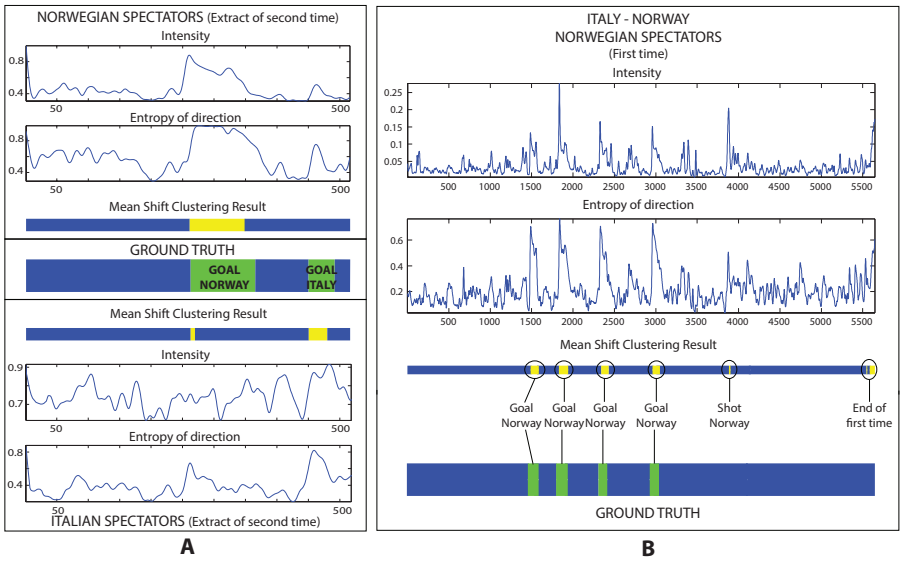

Fig. 3. Salient events detection. Plot A shows the results over the same video considered in Figure 2. Here, the extracted flow intensity and entropy of flow direction of both Norwegian and Italian spectators are shown. The small bars show the results of Mean Shift clustering (the yellow boxes represent 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.

4 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, 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 based ground truth”, in that we have compared our findings with what had

been explained in sociological theories. 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. This, moreover, could be complemented with ethnomethodologically oriented videoanalysis (see [10]). 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.

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.

Acknowledgments

This work is part of the Oz ("Observing the attention") project, financed by the Winter Universiade Trentino 2013 Educational Programme.
D. Conigliaro, F. Setti, C. Bassetti and R. Ferrario are supported by the VisCoSo project grant, financed by the Autonomous Province of Trento through the "Team 2011" funding programme.

References

1. E. L. Andrade, S. Blunsden, and R. B. Fisher. Modelling crowd scenes for event detection. In *ICPR*, pages 175–178, 2006.
2. A.E. Berlonghi. Undestanding and planning for different spectator crowds. *Safety Science*, 18:239–247, 1995.
3. A. B. Chan and N. Vasconcelos. Bayesian poisson regression for crowd counting. In *ICCV*, pages 545–551, 2009.
4. M. Cristani, V. Murino, and A. Vinciarelli. Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space. In *CVPRW*, pages 51–58, 2010.
5. M. Cristani, A. Pesarin, A. Vinciarelli, M. Crocco, and V. Murino. Look at who's talking: Voice activity detection by automated gesture analysis. In *AML Workshops*, pages 72–80, 2011.
6. M.A.T. Figueiredo and A.K. Jain. Unsupervised learning of finite mixture models. *PAMI*, 24(3):381–396, 2002.
7. E. Goffman. *Behaviour in Public Places.* Free Press of Glencloe. Notes on the Social Organization of Gatherings, 1963.
8. J.H. Goldstein and R.L. Arms. Effects of observing athletic contests on ostility. *Sociometry*, 34(1):83–90, 1971.
9. G. W. Guyot, G. R. Byrd, and R. Caudle. Classroom setting: An expression of situational territoriality in humans. *Small Group Behavior*, 11:120–128, 1980.

10. C. Heath, J. Hindmarsh, and P. Luff. *Video in Qualitative Research. Analysing Social Interaction in Everyday Life*. Sage, London, 2010.
11. N. Kaya and B. Burgess. Territoriality. seat preferences in different types of classroom arrangements. *Environment and Behavior*, 39(6):859–876, 2007.
12. L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, pages 1446–1453, 2009.
13. L. Kratz and K. Nishino. Tracking with local spatio-temporal motion patterns in extremely crowded scenes. In *CVPR*, pages 693–700, 2010.
14. I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, 2005.
15. V. Mahadevan, Weixin Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, pages 1975 –1981, 2010.
16. R. Raghavendra, A. Del Bue, M. Cristani, and V. Murino. Abnormal crowd behavior detection by social force optimization. In *HBU*, pages 134–145, 2011.
17. A. Roadburg. Factors precipitating fan violence: a comparison of professional soccer in britain and north america. *The British Journal of Sociology*, 31(2):265–276, 1980.
18. P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. *Int. J. Comput. Vision*, 80(1):72–91, 2008.
19. B. Zhan, D.N. Monekosso, P. Remagnino, S.A. Velastin, and L.-Q. Xu. Crowd analysis: a survey. *Machine Vision Applications*, 19(5-6):345–357, September 2008.
20. H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *CVPR*, pages 819–826, 2004.