Computer Vision and Image Understanding 159 (2017) 47–58

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

The S-Hock dataset: A new benchmark for spectator crowd analysis

Francesco Setti[a,b,*], Davide Conigliaro[a,b], Paolo Rota[c], Chiara Bassetti[b], Nicola Conci[d], Nicu Sebe[d], Marco Cristani[a,b]

[a] *University of Verona, Strada Le Grazie 15, I-37134 Verona, Italy*
[b] *ISTC–CNR, via alla Cascata 56/C, I-38123 Povo (TN), Italy*
[c] *Vienna University of Technology, Favoritenstr. 9/183-2, A-1040 Vienna, Austria*
[d] *University of Trento, via Sommarive 9, I-38123 Povo (TN), Italy*

ARTICLE INFO

Article history:
Received 4 April 2016
Revised 16 December 2016
Accepted 2 January 2017
Available online 11 January 2017

Keywords:
Spectator monitoring
Crowd analysis
Spectator crowd

ABSTRACT

Although crowd analysis is a classical and extensively studied problem for the computer vision community, the vast majority of the works in the literature assume a single type of crowd, while the sociological literature classifies a number of different typologies, each one with their own distinctive traits. In this paper we focus on a particular kind of crowd referred in sociology as *spectator crowd*, which consists a number of people that are “interested in watching something specific that they came to see” Berlonghi (1995). This is the typical social formation that attends entertainment events like sport matches, concerts, movies, etc. In this work we present a novel dataset, the Spectators Hockey (S-Hock), containing almost 30 hours of videos recorded at an ice hockey rink during the Winter Universiade “Trentino2013”. On these data we provide a massive annotation, focusing on the spectators at different levels of detail: from high level features describing which team a person supports and if he/she knows his/her neighbors; to a lower level, where we consider standard pose information as well as atomic actions like applauding, jumping, etc. We also provide annotations for the game field, which allows us to analyze the relationship between the crowd behavior and the events of the match. Eventually we provide more than 100 million of annotations, that can be used for many different tasks spanning from standard applications, like people counting and head pose estimation, to higher level tasks, like excitement estimation and automatic summarization. We provide protocols and baseline results for all of these applications, encouraging further research in these field.

1. Introduction

Crowds are more and more a feature of our urban life. Modeling and predicting their dynamics is therefore pivotal from several points of view, ranging from organizing and planning to surveillance and public safety management. Capturing and understanding crowd dynamics, indeed, may help preventing and/or managing critical situations. Given the complexity of the task, however, we believe this has to be an interdisciplinary endeavor.

In computer vision, a crowd is defined as an entity that is identified when “the density of the people is sufficiently large to disable individual and group identification” (Jacques Júnior et al., 2010). From a sociological perspective, this is a quite general definition, encompassing different kinds of crowds, whose members behave in very different ways accordingly. As we are about to see, a thorough review of the sociological literature allows to distinguish four kinds of crowd. Further, it is important to consider that crowds, better defined as large gatherings (Goffman, 1961; 1963; McPhail, 1991), are not homogeneous (not all members are the same), nor unanimous (not all have the same motive/s), nor mutually inclusive (not all behave the same), nor continuous (mutually inclusive behavior, when present, is not uninterrupted). Large gatherings are characterized instead by alternating individual and collective actions, whereby both vary in quality, and collective action present varied proportions of co-present people engaged in any particular action. Therefore, the “crowd mind” is a myth (McPhail, 1991), and crowds can encompass smaller groups with which members do identify (just think to a family at the stadium).

The four categories into which crowds can be divided are the following[1]:

[*] Corresponding author.
E-mail address: francesco.setti@univr.it (F. Setti).

[1] see Bassetti (2016) for more details on this taxonomy.

http://dx.doi.org/10.1016/j.cviu.2017.01.003
1077-3142/

1. *prosaic* (McPhail, 1991) or *casual* (Blumer, 1951; Goode, 1992) crowds consist of large collections of individuals sharing no more than a spatio-temporal location, that is, they are co-present by chance and they do not share a single focus of attention and action (unfocused interaction (Goffman, 1961; 1963; 1981)). People in line at airport security checkpoints, or pedestrians in the streets are a couple of examples;
2. *spectator* (Berlonghi, 1995; McPhail, 1991) or *conventional* (Blumer, 1951; Goode, 1992) crowds are ensembles of people gathering for specific social events, such as theatrical performances or sport matches precisely. Being basically audiences, people in spectator crowds have a common focus of attention and action (common-focused interaction (Goffman, 1961; Kendon, 1988));
3. *expressive* crowd (Blumer, 1951; Goode, 1992) are collections of individuals who gather for a social event and who intend to act as fully active members, that is to participate in collective action. Examples range from flash-mob dancers, to Mass participants, to sport supporters (not just attendees) that get together to dance, to ritually pray, to cheer. Action is concerted rather than just common, and the focus of attention is jointly shared among participants (jointly-focused interaction (Goffman, 1961; Kendon, 1988));
4. *demonstration/protest* (McPhail, 1991) or *acting* (Blumer, 1951; Goode, 1992) crowds are collections of people gathered for events such as mobs, riots, sit-ins or marches who intend to participate in collective action. Therefore, action is concerted and interaction is jointly-focused.

Under this taxonomy, we can say that most of the extant computer vision approaches focus primarily on casual (Chan and Vasconcelos, 2009; Kratz and Nishino, 2010; Raghavendra et al., 2011), and protest crowds (Krausz and Bauckhage, 2012), with hundreds of techniques and various datasets, whereas very few (if any) deal with spectator crowds and their expressive segments (e.g., sport match attendants and groups of supporters within). In computer vision, crowd analysis focuses on modeling large masses, where a single person cannot be finely characterized, due to the low resolution, frequent occlusions and the particular dynamics of the scene. Therefore, many state-of-the-art algorithms for person detection and re-identification, multi-target tracking, and action recognition cannot be directly applied in this context. As a consequence, crowd modeling has developed its own techniques such as multi-resolution histograms (Zhong et al., 2004), spatio-temporal cuboids (Kratz and Nishino, 2009), appearance or motion descriptors (Andrade et al., 2006), spatio-temporal volumes (Laptev, 2005), dynamic textures (Mahadevan et al., 2010), computed on top of the flow information. The extracted information is then employed to learn different dynamics like Lagrangian particle dynamics (Raghavendra et al., 2011), and in general fluid-dynamic models. The most important applications of crowd analysis are abnormal behavior detection (Mahadevan et al., 2010), detecting/tracking individuals in crowds (Kratz and Nishino, 2010), counting people in crowds (Chan and Vasconcelos, 2009), identifying different regions of motion and segmentation (Sand and Teller, 2008). Only recently, Navarathna et al. (2014) started working on spectators of movies, trying to infer movie ratings from their behavior during the show. In this paper the authors present a regression method to estimate the rating of a film by using motion history features, both on the individual and group level, and support vector regression. They tested their algorithm on a testbed environment that contains a maximum of 10 people per session. In our case, the amount of people considered is much higher (from 150 to 500 people).

The above mentioned lack of attention to spectator crowds and their specific dynamics is threatening. From a recent study, conducted in 2014 by the UK Home Office,[2] disorders at stadiums caused 2273 arrests only considering the FA (Football Association) competitions in the last year. Moreover, in the last 60 years, 1447 people died and at least 4600 were injured at the stadiums during major events around the world.[3] These statistics motivated in several countries the implementation of emergency plans to ensure safety and a better management of critical situations. This is here where computer vision may consistently help. Stadiums can often be targets of violence. Hence, protecting event goers, stadium staff, event performers and athletes becomes a priority. This need grew in importance in these last years; as one of the most evident signals of this trend, many international summits are occurring and will occur in the next months. For example, the European City and Public Security Summit in 2017 in London will bring together policy makers and leading experts from the private and public sector, including former and current police and counter terrorism services and the heads of Security of some of Europe's largest sports, leisure, retail and public attractions.[4] As another example, the annual UEFA/EU conference takes place every year at the start of the season. The conference brings together national associations' security officers, stadium safety managers, club safety officers and police representatives from all European clubs that have qualified for the next season of the UEFA Champions League and UEFA Europa League. Over 350 delegates review the past season, exchange good practices and discuss arrangements for the upcoming matches. The high attendance reflects the increased scope of stadium and security affairs, and the importance attached to it by authorities and the football family across Europe. The latest conference took place in Bucharest in September 2016.[5] The ESSMA Stadium Summit gather stadium experts, club and league representatives to discuss various aspects of stadium management, including, but not limited to, fan entertainment, safety and security, commercial exploitation and pitch management.[6]

The present article is an initial attempt to address this topic, and offers the first dataset on the subject, the Spectators Hockey (S-Hock). It regards an international hockey competition (12 countries from all around the world have been invited) held in Trento (Italy) during the 26th Winter Universiade, and focuses on the final 4 matches of the tournament. The dataset is unique in the crowd literature, and in general in the surveillance realm, since the crowd as a whole is mostly static and the motion of each spectator is constrained within a limited space in his/her surroundings.

S-Hock captures the crowd using 4 cameras, at different resolutions and different levels of detail (Fig. 1). At the highest level, it models the network of social relations among the public (who knows whom in the proximity), what is the supported team, and what has been the best action in the match; these data have been gathered through structured questionnaires conducted at the stadium for each match. At a medium level, spectators are localized, and information regarding the pose of their heads and bodies is given. Finally, at the lowest level, a fine grained specification of all the actions performed by each single person is available. This information is summarized by a large number of annotations, collected over a year of work: more than 100 million of double checked annotations. By far, this is a consistent step ahead in the field of the fine grained activity recognition. Indeed, S-Hock potentially allows to deal with hundreds of tasks, some of which are documented in the following sections.

[2] Football-related arrests and football banning order statistics, Season 2013–14, available online at http://goo.gl/j9yYYQ.

[3] http://goo.gl/xMU2Zf.

[4] http://www.citysecuritysummit.com/.

[5] http://www.uefa.org/protecting-the-game/security/.

[6] http://www.ecaeurope.com/news/essma-summit-2016/.

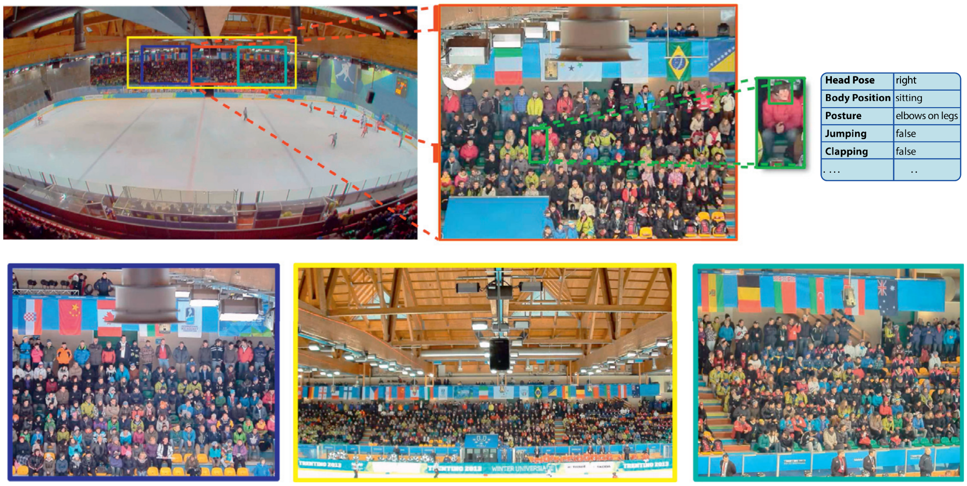


Fig. 1. Sample images forming the S-Hock dataset. On top left, the wider scene with the game field and the stands. In the blue, green and cyan frames are the high resolution spectator scenes; in the yellow frame is the low resolution spectator scene. On top right, a schematic representation of the annotation at the individual scale. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Another crucial cue of S-Hock is that it has two main facets, one focused on the crowd, while the other is spent on the game field. In this sense, the dataset is multidimensional, where the two dimensions consist of data temporally synchronized. Each annotation in the game field has a time stamp which follows that of the crowd. This obviously enlarges the number of possible applications that could be carried out, investigating the reactions of the crowd to the actions of the game, opening up to applications of summarization, content analysis, retrieval, etc. In particular, as we shall see in the closing section, many new applications can be designed for the domain of public entertainment.

In this article we discuss issues related to low and high level detail of the crowd analysis, namely, people detection and head pose estimation for the low level analysis, spectator categorization and automatic highlights generation for the high level analysis. Spectator categorization is a kind of crowd segmentation, where the goal is to cluster different groups of supporters and describe these groups with a set of features like the team they support or the average excitement of the group. For all of these applications, we define the experimental protocols, thereby promoting future comparisons. From the experiments we conducted, we show how standard methods for crowd analysis, which work well on state-of-the-art datasets, are not fully suited to the data we are dealing with, thus requiring us to face the problem from a different perspective. Therefore, besides baselines, we also propose customized approaches specifically targeted at the spectator crowd, thus defining new upper bounds.

In brief, the main contributions of this paper are:

- A novel dataset for spectator crowd, which describes at different levels of detail the crowd behavior with millions of ground truth annotations, synchronized with the game being played in the field. Crowd and game are captured with different cameras, ensuring multiple points of view;
- A set of applicative tasks for analyzing the spectator crowd, some of them are brand new;
- A set of baselines for some of these tasks, with novel approaches which definitely overcome the standard crowd analysis algorithms.

The rest of the paper is organized as follows: The details of the data collection and labeling are reported in Section 2; the tasks of people detection, head pose estimation, and spectator categorization are introduced in Section 3, focusing on contextualizing the problem, discussing the related state of the art (if any), presenting the considered baselines and our approaches, and discussing the results obtained. Finally, in Section 4, other applications worth investigating are briefly discussed, promoting further research on this new topic.

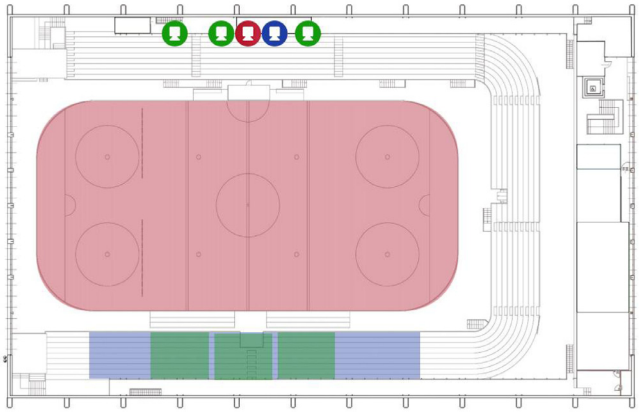


Fig. 2. Schematic representation of the data acquisition settings. The spectators were forced by the logistic of the ice-stadium to seat on the south bleachers (bottom in the map), while the north bleachers were restricted to organization. The red and blue cameras are full HD cameras with wide lens for the acquisition of the ice-rink and the whole spectator crowd, while the green ones are pointing specific areas of the spectators. See Fig. 1 for a sample image from each camera. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Data collection and annotation

The 26th Winter Universiade was held in Trento (Italy) from 11 to 21 of December 2013, attracting about 100,000 people from all over the world, both among athletes and spectators. The data collection campaign focused on the last 4 matches (those with more spectators) of the men's ice hockey tournament, held in the same ice-stadium of Canazei: here we set up 5 cameras arranged in the configuration showed in Fig. 2. Two full HD cameras (1920 × 1080 pixels, 30 fps, focal length 4 mm) were employed: one was pointed on the ice rink to record the match events (the red one in Fig. 2), and another one for a panoramic view of all the bleachers (the blue one in Fig. 2). Moreover, we used 3 HD cameras (1280 ×

Table 1
The annotations provided for each person and each frame of the videos. These are only typical values that each annotation can have, a detailed description of the annotations is provided with the dataset. The meaning of the head pose attributes will be explained later in the paper. For the experiments in Section 3.2, *away* class has been further divided in *far-left* and *far-right* to discriminate the head pose even when a spectator is not looking toward the ice rink.

Annotation	Typical values
People detection	Full body bounding box [*x*, *y*, width, height]
Head detection	Head bounding box [*x*, *y*, width, height]
Head pose	Far left, left, frontal, right, far right, away, down
Body position	Sitting, standing, (locomotion)
Posture	Crossed arms, arms alongside body, elbows on legs, hands on hips, hands in pocket, hands on legs, joined hands, hands not visible, crossed legs, parallel legs, legs not visible
Locomotion	Walking, jumping (each jump), rising body slightly up
Action / interaction	Waving arms, pointing toward game, pointing outside game, rising arms, waving flag, hands a cone, whistling, positive gesture, negative gesture, applauding, clapping (each clap), using device, using binoculars, using megaphone, patting somebody, call for attention, hugging somebody, kissing somebody, passing object, hit for fun, hit for real, opening arms, hands to forehead, hitting hands (once), none
Supported team	The team supported in this game (according to the survey)
Best action	The most exciting action of the game (according to the survey)
Social relation	If he/she did know the person seated at his/her right (according to the survey)

1024 pixels, 30 fps, focal length 12 mm) focusing on different parts of the spectator crowd (the green ones in Fig. 2). In total, about 30 h of recordings have been collected, with inter-camera synchronization: this brought the interesting feature of having the crowd synchronized with the game on the rink.

After the match, we asked to the spectators to fill a simple questionnaire with three questions (whose significance will be made clear later in the paper):

- Which team did you support in this match?
- Did you know at the beginning of the match who was sitting next to you?
- Which has been the most exciting action in this game?

On average 30% of the spectators filled the from, with peaks of 80% of them during the final match on the central part of the standings.

In S-Hock we focus on game segments from different hockey matches in order to stress the generalization capability of the considered algorithms, since in different matches we have different people and illumination conditions. In particular, from each match we selected a pool of sequences in order to represent a wide, uniform and representative spectrum of situations, e.g. tens of instances of goals, shots on goal, saves, faults, timeouts (each sequence has more than one event). Each video is 31 s long (930 frames), for a total of 75 sequences, namely 15 for each camera. The annotations reported in Table 1 have been performed on one of the three close-field cameras, whereas the videos recorded with the other two cameras were annotated only with the survey information. The fourth view is a wide-field view of the previous three views and the fifth is oriented toward the ice rink in order to record the game events.

Each sequence has been annotated frame by frame, spectator by spectator, by a first annotator, using the ViPER-GT format (Doermann and Mihalcik, 2000).[7] The annotator had to perform three different macro tasks: detection (localizing the body and the head), posture and action annotation, respectively. This amounted to deal with a set of 50 labels, listed in Table 1.

From the whole set of possible features that can characterize the human action and interaction, we selected the annotated *elementary forms of action* (McPhail, 1991) as they are strictly connected with those more relevant for the analysis of social interaction, and those most related to our specific setting, i.e. sport spectator crowd (e.g. bodily posture or proxemics, and actions such as waving arms or shaking “fan objects”). More specifically, we considered the available microsociological literature on behavior in public and social interaction (Garfinkel, 1967; Goffman, 1961; 1963; McPhail, 1991), with particular attention to non-verbal conduct (proxemics, bodily posture, gesture, etc.). We took into particular consideration also the literature on social interaction in large gatherings (McPhail, 1991),that is, literature on what is commonly referred to as “crowd behavior”. In doing so, we focused in particular sport spectator gatherings (Bassetti, 2016).

On the other hand, the collected video recordings constituted the database onto which microsociological analysis have been conducted, prior to video annotation, in order to select what we call the *atomic components of action-in-interaction* of the considered setting – that is, the elementary actions to be annotated on the dataset. The empirical analysis have been conducted accordingly to the principles and procedure of Ethnomethodology (Garfinkel, 1967) and Conversation Analysis (Psathas, 1995; Sacks et al., 1995), the so-called EM/CA approach. Ethnomethodological video-analysis (see Heath et al. (2010)) plays particular attention: to the sequentiality of interaction, which is regarded as an unfolding process; to the perspective of the participants (what is available to their knowledge and perception) at any point of such a sequence, rather than the perspective of the human analyst who knows what happens next; to the context of the interaction as simultaneously constitutive of, and constituted by the actions people perform in it. These characteristics make the approach particularly well-suited to be integrated into computer vision techniques.

The EM/CA analysis of the video-set has identified in a first phase two main activities enacted by sport spectators:

- reading the field, that is, game-actions’ projection;
- performing the stands, which entails both
  - doing [attending the game], that is, displaying attention to the game (e.g., pointing to or looking at the game field), and
  - doing [supporting the team], that is, actively cheering, displaying support (e.g., standing, jumping, clapping)

Consequently, the subsequent analytical phase was devoted to identifying markers of:

- (dis)attention and (dis)engagement with the game-field activities;
- game-actions projection, with consequent increase in attention/engagement (i.e. excitement);
- enjoyment/annoyance and (dis)satisfaction with respect to, respectively, game-actions and their outcomes;

[7] The toolkit is available at http://viper-toolkit.sourceforge.net/.

Table 2
The game situations annotated for the 4 matches (considering only the second half of each game). Here the number of persons is an approximation of the number of people during the entire video. The last row indicates the number of sequences in which there is a specific game situation.

	Persons	Goals	Saves	Shots	Fouls	Timeouts	Play
Match 1	500	2	21	43	2	–	–
Match 2	250	3	16	23	5	1	–
Match 3	315	–	13	22	5	–	–
Match 4	150	1	15	28	8	–	–
# Seq.	–	3	7	12	1	1	4

- mutual coordination in doing [attending the game], and in doing [supporting the team] (that is, in displaying enjoyment/annoyance and (dis)satisfaction with particular game-actions or their outcomes.

Each annotator had two weeks to annotate 930 frames, and was asked to do it in a specific lab, in order to monitor him/her and ensure a good annotation quality. After that all the sequences have been processed, producing a total amount of more than 100 million of annotations, a second round of annotations started, with the “second annotators” that were in charge of correcting the errors from the first-round annotation phase. The whole process involved 15 annotators, all paid for their work, and lasted almost 1 year.

Together with this fully labeled reduced version of the dataset, we also release the complete dataset (about 30 h of video stream) with high level annotations in terms of events happening on the ice rink (e.g. goals, shots, saves, etc.). Statistics about the dataset content are reported in Table 2. All the data are available online at http://vips.sci.univr.it/dataset/shock, and are free to use for research purposes.

3. Applications

In this section we propose a number of applications for which S-Hock can represent a valuable resource in terms of algorithms' testing and benchmarking. We will focus on two classical tasks (people detection and head pose estimation), and two more specific ones, related to social aspects (spectators categorization) and multimedia content generation (automatic summarization) respectively. For each task, we briefly present the state-of-the-art, taking into account only those methods that can be applied to our scenario, and some preliminary experiments conducted on our dataset. We also propose some improvements of the standard methods that exploit the specific features of a spectator crowd and the relationship between the crowd behavior and what is happening in the game.

3.1. People detection

People detection is a long running problem in the computer vision community where a number of different methods and algorithms have been presented over the last 30 years (Dollár et al., 2009; Enzweiler and Gavrila, 2009). While many different approaches are available in the literature, such as wavelet-based AdaBoost cascade, NN/LRF and combined shape-texture detection, the most popular approaches in the recent years are classification schemes based on HOG features (Dalal and Triggs, 2005) and the Deformable Part Model (DPM) (Felzenszwalb et al., 2010).

Unfortunately, most of the state-of-the-art methods are not suitable for our scenario in their original version. This is mostly due to two reasons: first, images are extremely low resoluted – the bounding box of a person is on average 70 × 110 pixels –, and second, there are many occlusions – usually only the upper body is visible, rarely the entire body but sometimes only the face.

Mainly to overcome these problems, some recent works studied how to embed an explicit model of the visual scene into the detection algorithms. Barinova et al. (2012) proposed to use Hough transform as an alternative to the non-maxima suppression stage, allowing them to handle multiple instances of the same object class in a very dense scenario. San Biagio et al. (2013) proposed a new image descriptor, called HASC, that encodes linear and nonlinear relationships between heterogeneous dense feature maps through information-theoretic measures; this makes it able to treat complex structural information in a compact and robust way. On the other hand, in order to overcome the occlusions issue, Wu and Nevatia (2007) proposed to use part detectors learned by boosting a number of weak classifiers based on edgelet features, and then to combine the responses of part detectors to form a joint likelihood model including the analysis of possible occlusions. Eichner et al. (2012) fused DPM (Felzenszwalb et al., 2010) and Viola and Jones (2001) detectors to identify upper bodies, i.e. people standing (or seated) upright and seen from the front or the back (yet not from the side). Finally, Rodriguez et al. (2011) proposed to resolve all detections jointly by optimizing a joint energy function that combines crowd density estimation and the localization of individual people.

In this article we provide 5 different baselines for people detection, 2 classic approaches and 3 state-of-the-art methods. Both classic approaches are based on linear SVM classifiers and differ from each other only in terms of the descriptor used; the first is based on Histograms of Oriented Gradients (HOG) (Dalal and Triggs, 2005) (cell size of 8 × 8 pixels) and is dubbed in the following *HOG+SVM*, while the second is the Heterogeneous Auto-Similarities of Characteristics (HASC) descriptor (San Biagio et al., 2013) and is dubbed in the following *HASC+SVM*. As for the state-of-the-art methods, we tested: the Aggregate Channel Features (*ACF*) detector (Dollár et al., 2014), which works on the color channels by computing integral images and Haar wavelets inside a (Viola and Jones, 2001) framework, fusing them together; the Deformable Part Model (*DPM*) (Felzenszwalb et al., 2010), that feeds a latent SVM classifier with a combination of templates representing parts of the whole object to be arranged in a deformable configuration; and the Calvin Upper Body Detector (*CUBD*) (Eichner et al., 2012), a combination of the DPM framework trained on near-frontal upper-bodies and the Viola-Jones face detector.

To further investigate the specificity of the S-Hock dataset, we propose to use, on top of all these methods, a strong prior information, i.e. the people are “forced” by the environment to arrange in a grid – the seats on the bleachers. Thus, assuming people are actually seated or anyway distributed according to seats (e.g. some people is standing most of the time in front of a seat where they put their belongings), we can generate a prior probability map by assigning a higher probability to the locations next to the seats. Since we do not know in advance the camera calibration (i.e. the relative orientation between the camera and the stands), we prefer to adopt a post-processing strategy where we add to the detection confidence map the average of the map over the rows and the columns; in this way we only assume that the camera is roughly perpendicular to the spectators, which is reasonable since the camera is quite far from the stands, but the method is very robust to small movements of the camera that can result in a translation of pixels. Consider D the detection confidence map, being $D(x, y)$ the probability that the patch centered in (x, y) contains a person, the modified output $\tilde{D}$ for a target location $(\hat{x}, \hat{y})$ is given by:

$$\tilde{D}(\hat{x}, \hat{y}) = D(\hat{x}, \hat{y}) + \sum_{i} D(x_i, \hat{y}) + \sum_{j} D(\hat{x}, y_j) \quad (1)$$

We adopted a standard experimental protocol based on training, validation and testing. We used all the 11 sequences of the final match as testing set, while 2 sequences of the same semi-final

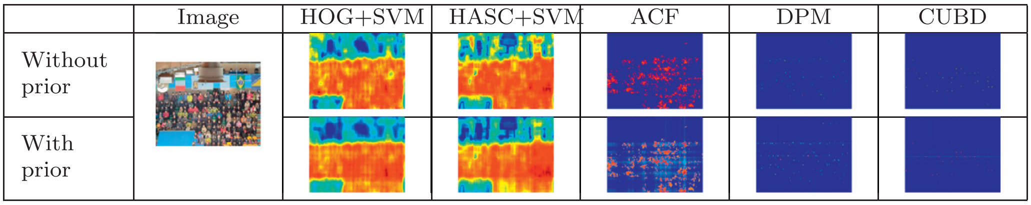


Fig. 3. Qualitative results of the people detection algorithms. The detection confidence map for each method is reported both with and without the application of the grid-arrangement prior. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
People detection results in terms of precision, recall and F_1 score, with and without the contribution of the grid arrangement prior. Best performances reported in bold.

Method	No prior			With prior		
	Prec.	Rec.	F_1	Prec.	Rec.	F_1
HOG + SVM	0.743	0.561	**0.639**	**0.662**	**0.709**	**0.684**
HASC+SVM (San Biagio et al., 2013)	0.365	**0.642**	0.465	0.357	0.685	0.469
ACF (Dollár et al., 2014)	0.491	0.622	0.548	0.524	0.649	0.580
DPM (Felzenszwalb et al., 2010)	0.502	0.429	0.463	0.423	0.618	0.502
CUBD (Eichner et al., 2012)	**0.840**	0.303	0.444	0.613	0.553	0.581

match have been used as training set, and the last 2 sequences (of 2 different matches) as validation set to tune some parameters (specifically threshold for the minimum detection score and the parameters for the non-maxima suppression stage). For the training phase, we randomly selected 1000 individuals to be used as positives, while a background image (i.e. an image of empty bleachers) has been used to randomly generate negatives For the testing phase, we downsampled the videos by processing 1 frame every 10, resulting in 93 frames and about 13,600 individuals per video (about 1000 frames and 150,000 individuals in total). For HOG+SVM and HASC+SVM we adopted a simple sliding window strategy for the generation of the candidates. We used patches of fixed size of 72 × 112px and a movement step of 8px, generating a detection confidence map with dimension 160 × 118. As for ACF, DPM and CUBD, the generation of the candidates is part of the algorithms.

Following the evaluation protocol of Everingham et al. (2010), we consider an individual as correctly identified if the intersection over union of the predicted and annotated bounding boxes is higher than 50%. As performance measures we use precision, recall and F_1 scores.

Qualitative results of the baselines and the grid arrangement prior contribution is shown in Fig. 3, while quantitative results are in Table 3. Surprisingly, the best performing method is also the simplest one (HOG+SVM), while the frameworks based on deformable part models (DPM and CUBD) perform very poorly in the standard version. The main reason we find is that the extremely low resolution of the images make the detection of the parts even more difficult than the detection of the person as a whole. Numeric results also prove that the grid arrangement prior consistently improves the performances of all the methods in terms of F_1 score.

3.2. Head pose estimation

The scenario described by S-Hock can be generalized in most of the visual surveillance environments where cameras are deployed in a big public area in order to maximize the coverage. In these contexts, one interesting task is the analysis of the posture of each individual with the goal of tracking their focus of attention along time. In this section we focus on the automatic pose estimation of the head. In this context, despite the good resolution of the cameras, the distance between the camera and the filmed subjects is rather high, which is necessary in order to cover the whole side of the gallery in the ice stadium. In our particular case, the appearance of each spectator's head can be contained in a bounding box of rather small dimensions (50 × 40 pixels on average). In this scenario, most of the traditional and best performing methods (Chen et al., 2012; Kemelmacher-Shlizerman and Basri, 2011; Zhu and Ramanan, 2012) are inapplicable due to the impossibility in finding landmark points on each subject's face. Considering the low-resolution scenario, viable methods are few. Among them, we selected two algorithms which are suitable for such an application. The first has been proposed by Orozco et al. (2009), it relies on the computation of a descriptor based on the distance between the test image and the mean image for each orientation. An SVM classifier has been used to perform the final decision. The second viable approach has been proposed by Tosato et al. (2013), in this work the authors exploit an array of covariance matrices in a boosting framework. The image of the head has been divided in patches which have been weighted according to their description capability. The application of those methods on S-Hock lead to performance similar to their application to other dataset. However, a considerable amount of time is needed for testing and much more for training the model (see Table 4). In order to overcome these issues we propose two approaches based on neural networks, which have recently produced state of art results in many computer vision branches (Girshick, 2015; Kontschieder et al., 2015; Szegedy et al., 2015).

Table 4
Classification accuracy for state-of-the-art methods averaged on the five classes and the computation time. The time used to refine the prediction through EACH is negligible comparing to the one used to train and test the neural network. The testing and training reported in the table, is considering the whole amount of seconds used to process the entire set of images on our non-GPU powered desktop machine.

Method	Avg. accuracy	Training time [sec]	Testing time [sec]
Orozco et al. (2009)	0.368	105,303	6263
WArCo (Tosato et al., 2013)	0.376	186,888	87,557
CNN	0.346	16,106	68
AE	0.348	9384	3
CNN + EACH	0.354	16,106	68
AE + EACH	0.363	9384	3

In this work we compare the performance of two architectures: namely, Convolutional Neural Networks (CNN) and Auto-encoders (AE). The input image has been resized 50 × 50 pixels and then

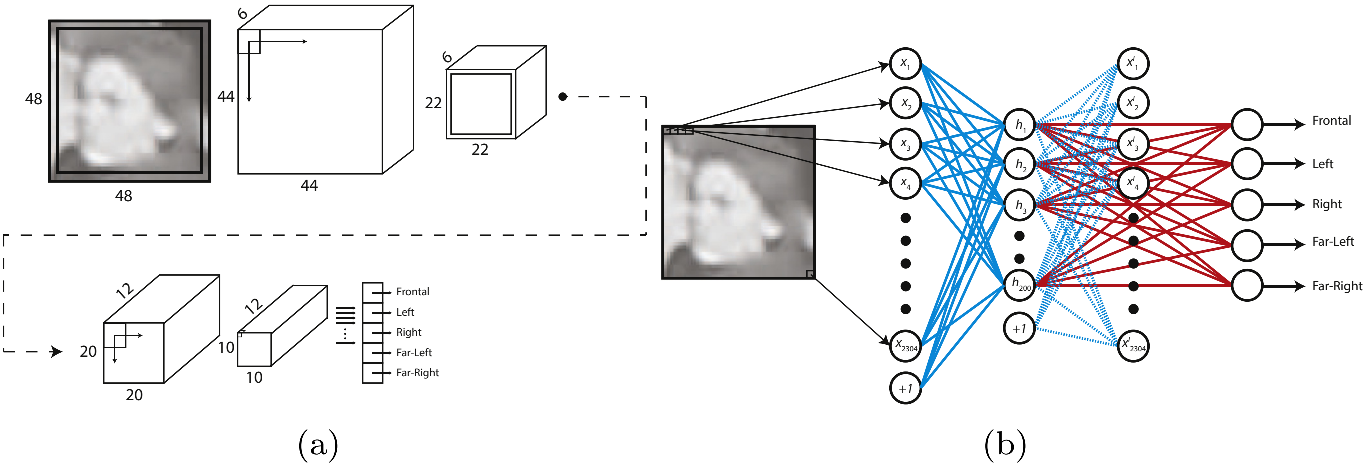


Fig. 4. (a) Architecture of CNN. (b) AE architecture: in cyan are pictured the interconnections between the auto-encoder that must be trained separately, in red instead there are the interconnections of the final NN. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

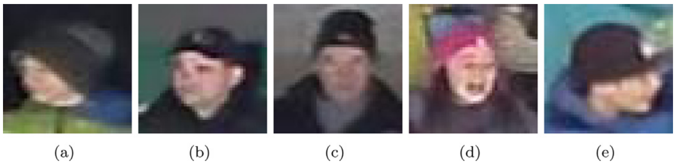


Fig. 5. Examples of the five head poses considered for the experiments in Section 3.2; in order (a) to (e): *far left, left, frontal, right, far right*.

normalized in order to be given as input to the two networks. The CNN model is a deep architecture similar to the one proposed by LeCun for handwritten digit recognition (LeCun et al., 1989): an input layer followed by 2 sets of convolution-pooling layers (see Fig. 4(a)). Both kernels in the convolutional layers are 5 × 5 pixels, the scaling factor of the pooling layer is 2, and the training has been performed over 50 epochs. The AE net is showed in Fig. 4(b) and is performed in two phases. In the first unsupervised phase the weights are learned automatically, in the second phase those weights will be used as initialization of a supervised traditional neural network that will output the final inference on the five classes. In this architecture the only hidden layer has size $h = 200$. Both training procedures (supervised and unsupervised) are refined in 100 epochs.

This task consists in classifying different head poses considering the following partition: *frontal, left, right, far left* and *far right*. These classes allow us to segment three zones of the ice rink and two situations in which the spectator's attention is addressed outside of it. The classes *down* and *away* have been ignored since t hey are not populated as much as the others. In a more quantitative fashion, frontal faces are considered roughly in the range between $-10°$ and $10°$, left and right spans from $-10°$ to $-80°$ and $10°$ to $80°$ respectively. The heads exceeding those angles in both directions are considered as *far left* and *far right* (see Fig. 5). This has been detailed to the annotators during the data labeling. The resulting dataset is then is composed by 107,299 and 34,949 images for training and test respectively.[8] The head location is feeded to the neural network using the ground truth position in order to derive a sort of upper bound in terms of performance.

The results proposed in Table 4 show that neural networks are giving results similar to the baselines, but the training and especially the testing phases are performed much faster in the proposed methods. Also consider that these tests have been performed on the same machine without the use of any GPU that would consistently improve the training speed. This speed up in classification time for both training and testing phases makes our method more suitable for real life applications where quick response and imminent decision are required. As a further remark, we trained WArCo by randomly sampling 5000 samples among all those available for training; this has been necessary in order to perform it in a reasonable amount of time.

A further experiment has been performed on this data scenario; The main intuition at its basis stems from the fact that people attention during a sport match is mainly given by the location of the action on the game field. For this reason we introduced an additional step named EACH (Event Attention CatcH). In this experiment the position of the puck on the ice has been modeled as a one dimensional Gaussian distribution centered on the puck itself. This model allow us to have a rough estimation of the area that is likely to attract spectators' attention. This information is used as a prior probability in order to refine the final head pose estimation. This probability $P_A^{(c)}$ is formalized in Eq. (2)

$$P_A^{(c)} = \sum_{i=L^{(c)}}^{U^{(c)}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - m^{(c)})}{2\sigma}} \tag{2}$$

where $L^{(c)}$ and $U^{(c)}$ are the lower and the upper boundaries of the rink for the specific class c respectively, $m^{(c)}$ is the position of the puck.

$$c = \arg\max_{c} \left(\alpha P_A^{(c)} + (1-\alpha) P_N^{(c)}\right) \tag{3}$$

The final decision is taken according to Eq. (3), where α is a weighting parameter, $P_N^{(c)}$ is the probability of the head pose to be assigned to class c computed by the Neural Network.

We observe that this model is much more beneficial when athletes are playing than when the game is paused by the referee's intervention. This particular aspect suggests us to tune the α parameter according to the game phase. The results reported in Table 4 are computed using $\sigma = 15$ and $\alpha = 0.3$. The ice rink information increases the accuracy by approximately 2% on both CNN and AE frameworks.

3.3. Spectator categorization

The *spectator categorization* task consists in spatially segmenting the spectator crowd on the basis of motion attributes and temporal regularization, and in associating to each segment a set of

[8] The sets are provided along with the dataset in order to ease future comparisons.

high level features. In this paper we propose two different high level features: the supported team, based on the fact that the majority of people in that segment support that particular team, and the average excitement level over the whole sequence. Notice that these are just two possible features, other ones can be proposed and implemented in order to characterize the segment. Moreover, whereas some features are strongly related to the specific scenario at hand, others are of general applicability. For instance, the supported team is meaningless if we analyze the spectator crowd at a music concert, while the excitement level is still relevant.

Spectator categorization is a subtask of the more general topic of crowd modeling and crowd behavior analysis, and in turn very connected to human activity analysis (Aggarwal and Ryoo, 2011; Gowsikhaa et al., 2014; Poppe, 2010). Jacques Júnior et al. (2010) stated that computer vision approaches for the behavioral analysis of crowds can be distinguished into two main typologies: the *object-based* approaches treat the crowd as a collection of persons and thus the analysis relies on the detection of individuals; while the *holistic* approaches treat the crowd as a single complex entity. Despite the sociologically-founded preference for conceptualizing crowds, from a theoretical point of view, as collections of individuals and groups rather than a single entity (or "mind"), from a technical point of view the choice between the two approaches depends on the specific scenario under analysis; in dense scenes, where it is very difficult to detect and track individuals, the holistic approach is more appropriate (see Jacques Júnior et al., 2010, p. 72). This is also the case of spectator crowds, where other than dense, the scene also presents a huge number of occlusions.

Holistic approaches usually collect global information about the crowd (e.g. crowd flows), ignoring local information (e.g. people detection and tracking). This is typically achieved by means of optical flow techniques.

In Ali and Shah (2007), the authors propose to use Lagrangian particle dynamics to segment the flows of a crowd; in this work the notion of a flow segment is equivalent to a group of people that perform a coherent motion. The motion of the crowd is captured by optical flow and a velocity field is generated; subsequently, particles are inserted into the velocity field by means of a numerical integration method, and their movements are used to construct a flow that reveals coherent structures. We will refer to this method with *LPD*.

Mehran et al. (2010) proposed to apply streakline representation of flow to a number of computer vision problems, and in particular they focus on crowd analysis. They use streaklines to transport information about a scene by repeatedly initializing a fixed grid of particles at each frame, then moving both current and past particles according to the optical flow results; this leads to a very accurate representation of the flow that allows to detect both spatial and temporal changes. Finally, streaklines are passed to a watershed segmentation scheme to cluster regions characterized by coherent motion. We will refer to this method with *Streaklines*.

As a third method, we consider the one proposed by Conigliaro et al. (2013a), that combines instantaneous segmentation based on motion features and temporal regularization. The image is divided into a set of overlapping patches, each one described by a 5-dim feature vector containing the position of the patch's centroid, the average intensity of the optical flow and the entropy of flow intensity and directions. Gaussian clustering with automatic model selection (Figueiredo and Jain, 2002) is used to compute the instantaneous segmentation of the scene. Hierarchical clustering is then exploited to group together patches that, over all the frames, consistently belong to the same instantaneous segments. This is achieved by the Patch Similarity History matrix. We will refer to this method with *PSH*.

To ensure a fair comparison, we defined a shared test protocol. We subdivided the scene into a set of patches (size 64 × 128 px) forming a regular grid with 50% of overlap. To generate ground truth, each patch is associated to the individual's bounding box with the highest overlapping area (if any); than each patch is labeled based on the team the associated individual supports. The rationale behind this is we want to segment different supporters groups, and we observed that, especially in concomitance with some context-specific events (e.g. a goal, a good save, a foul), supporters of the same team behave very similarly, and instead very differently than the supporters of the opposite team. Each method was tested using the standard settings provided by the authors of the original papers. To speed up the process and make the optical flow computation more robust, for our experiments we downsampled the original videos taking into account 1 frame every 10.

Fig. 6 shows both qualitative and quantitative results. The best performing algorithm is PSH, with a clustering accuracy of about 62%. This let us think that the personal behavior is well described by the motion flow features, and in particular by the flow entropy computed over both intensity and directionality. This consideration gives us a good starting point for the extraction of other high level features, such as the excitement level of each segment end of the crowd as a whole.

Once the crowd is segmented, it is possible to proceed with higher level analysis. As already showed in Conigliaro et al. (2013a,b), it is possible to estimate the excitement level either of a single segment or of the entire crowd by using motion flow features. This kind of analysis can be very beneficial for many applicative fields like marketing (e.g. to automatically detect the best moments to run advertisements), security (e.g. a very excited crowd is more likely to start a brawl), entertainment (e.g. crowd excitement can be related with the quality of the proposed show), etc.

Navarathna et al. (2014) proposed to use a motion history image to represent the long-time behavior of a single individual; then, the global behavior of the crowd is represented as the entropy of a matrix, that accounts for the pair-wise similarity among all the single behaviors. This approach seems to work very well in the testing environment, but it appears to be sensitive to the crowd size, since the dimension of the similarity matrix is quadratic with the number of people.

For this reason, we present in this paper a holistic method which is independent from the crowd dimension. The excitement calculation process relies on three motion flow features: the average flow intensity, and the entropy of flow in terms of intensity and directions. Given a generic patch k, the *flow intensity* $I(k)$ is the average of the flow intensity of each pixel in the patch; intuitively, this cue encodes how much movement characterizes the patch k. The last two features are based on the definition of entropy of a quantized physical quantity inside the patch k, given by:

$$E(k) = -\sum_{i=1}^{d} p(k_i) \cdot \log p(k_i) \tag{4}$$

where d is the total number of possible values assumed by the physical quantity, and $p(k_i)$ is the realization probability of the specific value i. In this analysis we are interested in the entropy of flow directions E_D and flow intensity E_I. Broadly speaking, E_D describes the kind of movement in the patch: high entropy values mean random directions, while low values address homogeneous movements in the patch (a similar use of this entropic descriptor has been exploited in Cristani et al. (2012)).

After that, considering each segment r computed as explained above, a *local* level of excitement is estimated by computing the value:

$$Exc(r) = \frac{I(r) \times E_D(r)}{E_I(r)^2} \tag{5}$$

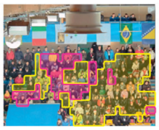
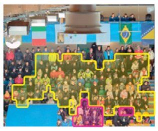
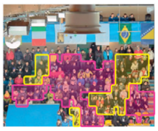
	GT	LPD [51]	Streaklines [52]	PSH [53]
				
Accuracy	–	0.592	0.559	**0.621**

Fig. 6. Spectator crowd segmentation: qualitative and quantitative results. The violet and yellow areas represent the two segments, corresponding with supporters of the different teams. A third segment is ignored in this pictures since it is all the rest of the image and it represents the background. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

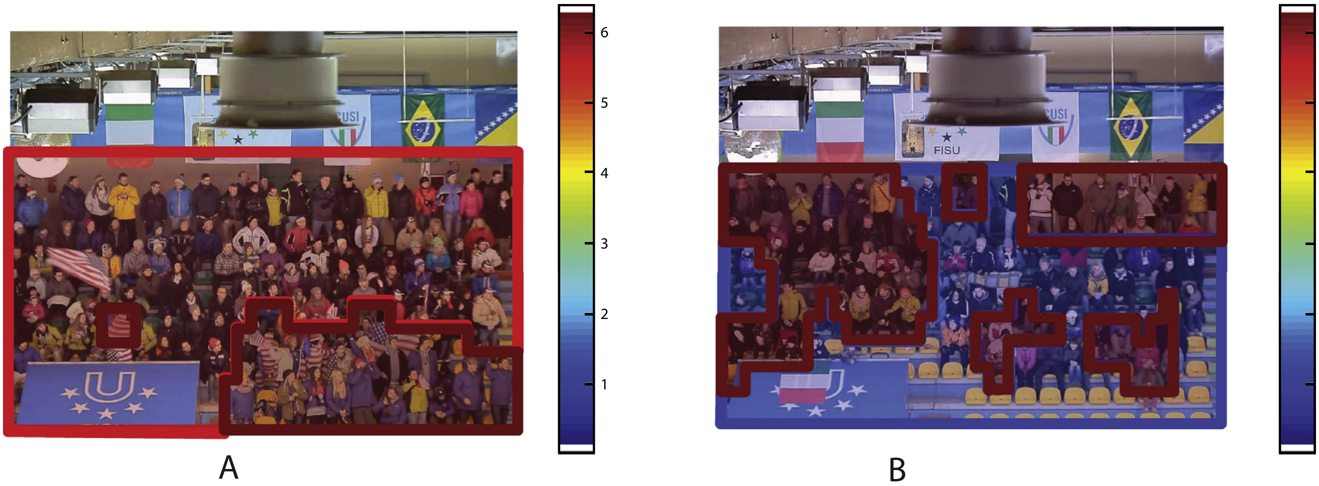


Fig. 7. Spectator segmentation and excitement level on two videos from two different matches. Image A shows the spectator crowd during a goal event, while during the game-play captured by image B, there are no salient events happening. The respective color-bars on the right of the images indicate the excitement level. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

over a short time interval (in the order of few seconds). In this equation $I(r)$, $E_I(r)$ and $E_D(r)$ are the average over all the patches belonging to the segment. The idea behind this equation is that we consider as a high level of excitement for a group of people an intense movement (high $I(r)$), with diverse directions (high $E_D(r)$), computed in a coordinated fashion for all people belonging to that region (low E_I). Finally, we can compute the average of $Exc(r)$ over time to globally characterize the segment.

3.4. Automatic summarization

Following the direction drawn in the spectator categorization task, we present an application that has the main goal to detect events taking place in the game-field, that globally trigger the excitement of the spectator crowd. This is the starting point for automatic video summarization, since we assume that events that generates reactions in the spectator crowd are the ones that people at home would be interested to see. Thus, the spectator feedback, automatically recognized, helps in highlighting exciting or crucial events that should be included in a video summarization of the show.

The highlights detection method is based on the same flow features used for the excitement estimation (i.e. I, E_D and E_I), and the excitement level computed as in Eq. (5). All these features are computed separately for each frame and for each crowd segment. Replicating this process for all frames gives a 4D signal which can be quantized in an unsupervised fashion by Mean Shift. In this case we preferred to use mean shift instead of Gaussian clustering, because pooling together the signal values of an entire sequence leads to highly irregular distributions, that are better handled by non-parametric algorithms.

After the quantization, looking at the mean values of each obtained cluster may serve to get insight on the kind of event being modeled and happening on the game-field. For example, clusters with high excitement may be originated by an interesting event happened in the game that should be highlighted (Fig. 7).

We conducted the experiments for highlights detection on the entire duration of a game period to identify the salient moments for the audience. All the videos are analyzed by considering a time window of 10 s with 5 s of overlap. The bandwidth parameter of mean shift was obtained experimentally, and is the same for each test. Depending on the choice of bandwidth, different actions of the game can be detected, such as goals or shots and saves.

In Fig. 8 we present the qualitative result of a full game period during the final match of the competition, which is Canada against Kazakhstan. During this game period, the Canadian players scored two goals. The upper box in the image shows the spectator categorization results both in terms of global (i.e. averaged over time) and instantaneous excitement level. Here the colored regions on the image represent two different groups of supporters, one for each team. The most excited region is the red one (Canadian supporters), with an average excitement level of 0.47, instead the light-blue region (Kazakhstani supporters) shows an average excitement level of 0.23. The two plots R1 and R2 show the temporal evolution of the excitement level respectively of Kazakhstani (R1) and Canadian (R2) supporters. From this spectator categorization, we identify three events that globally triggered the excitement of the spectator crowd (lower part of Fig. 8). Two of them correspond to the goals scored by the Canadian team, respectively at time 8:17 and 22:43. But the interesting thing is the detection of an event at time 2:16, which correspond to a shot on goal by a Canadian player that was alone in front of the goaltender: this was a great

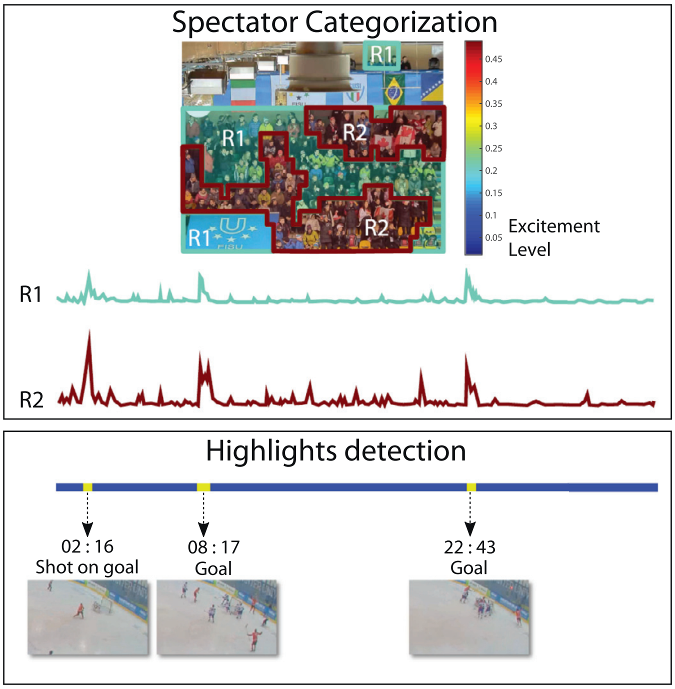


Fig. 8. Spectator categorization and highlights detection of a full game period (about 32 min). The picture in the upper box shows the spectator categorization results of the full video, where R1 and R2 represent the two segments related to supporters of different teams. The color associated to each segment indicates its average excitement (*i.e.* the average over all the individuals and the time span). Below, the temporal evolution of the excitement level for both the segments are reported. The lower box reports highlights detection results by means of mean shift approach: the yellow boxes represent salient events. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

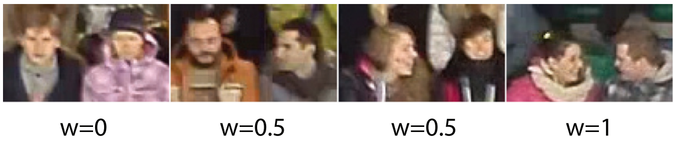


Fig. 9. Examples of relative head poses and associated scores of the weighting function (6).

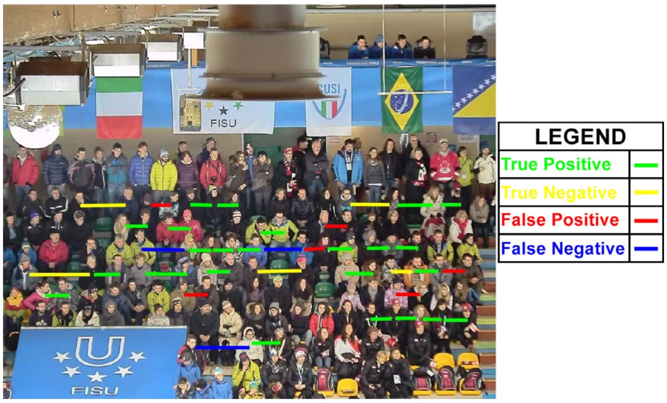


Fig. 10. Qualitative results of the group detection baseline. True positives (green) are pair of individuals that are predicted as a group and claimed in the survey they actually are; true negatives (yellow) are pairs of individuals that claimed they do not know each other and predicted as no-group; false positives (red) are unrelated persons predicted as groups; and false negatives (blue) vice-versa. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

opportunity to score for the Canadian team. Furthermore, the plot R2 shows that this shot on goal, was the most exciting moment for the Canadian supporters. To be noticed also that R1 plot shows peaks in correspondence with salient events detected for the Canadian supporters (R2), but the evolution of the excitement level is completely different for the two groups.

3.5. Group detection

In the recent years, many different works about automatic detection of groups of interacting people has been presented (Choi et al., 2014; Cristani et al., 2011; Hung and Kröse, 2011; Setti et al., 2013a; 2013b; 2015; Tran et al., 2013; Vascon et al., 2016). Unfortunately, most of the state of the art methods rely on the sociological concept of F-formation (Kendon, 1988), which defines a group as the spatial co-presence of two or more individuals, sharing common spaces with specific functions. In the case of spectator crowds, this spatial arrangement is strictly forced by the configuration of the standings and the ice-rink, thus enforcing specific people position and orientation, unrelated to their social behavior. For this reason, the state of the art methods for group detection are not applicable to our scenario.

To overcome the limitations given by the structural constraints, we present here a baseline method that relies on a set of heuristic rules based on the observation of human behavior in our dataset.

The general idea behind the baseline is that two persons, when they are interacting, tend to stay close and look at each other, independently from what is happening around them. Following this intuition, we compute an interaction score that is the average over all the frames of a weighting function that accounts for the view frustum intersection of two individuals. Broadly speaking, the interaction between two persons is more likely to occur when they look at each other, while it is unlikely to occur when they look in opposite directions. Mathematically, the interaction score for individuals p_l and p_r is defined from:

$$\mathcal{I}(p_l, p_r) = \frac{1}{T}\sum_{t=1}^{T} w(\theta_l(t), \theta_r(t)) \tag{6}$$

where p_l and p_r are two individuals seated next to each other respectively on the left and right side (on the image plane), T is the total number of frames, t is the current frame under analysis, $\theta_x(t)$ is the head orientation of p_x, $x \in [l, r]$, at time t with the conventions presented in Section 3.2 where we only consider 4 classes (A=away, L=left, F=front, and R=right), and $w(\cdot, \cdot)$ is a weighting function defined by the following look-up table:

		θ_r			
		A	L	F	R
θ_l	A	0	0	0	0
	L	0	0	0	**1**
	F	0	**0.5**	0	0
	R	0	0	**0.5**	0

Fig. 9 shows some examples of relative head orientations and the corresponding output of the weighting function.

Experimentally, we estimated the head orientation by means of the CNN algorithm presented in Section 3.2. The network has been re-trained with 4 classes defined as follows: classes *left* and *right* contain all the orientations towards the specified direction (i.e. considering both *left* and *far left* as well as *right* and *far right*), class *front* accounts for all the people looking in the middle of the ice-rink, while class *away* contains all the other cases (e.g. people looking down, at the phone, behind, etc.)

Fig. 10 reports qualitative results of the group detection baseline described above, while quantitative results are in Table 5.

Table 5
Quantitative results for group detection baseline.

	Precision	Recall	F_1	Accuracy
baseline	0.62	0.89	0.73	0.67

4. Conclusions

Based on the proposed, sociologically-founded taxonomy of crowds (Section 1), which represents a first contribution of the article, we tackled the issue of spectator crowd modeling, which is brand-new in computer vision, and which presents specific challenges. To this aim, we created a novel dataset, S-Hock, which the paper has illustrated with the purpose of showing its usefulness for testing many, diverse, and in some cases brand-new applications.

In particular, in the article we have focused on some low-level, traditional tasks –people detection and head pose estimation– and three novel, high-level challenges –spectator categorization, automatic summarization and group detection. In fact, on the one hand, we intended to underline the impact that considering the spectator crowd scenario has on the domain of extant crowd analysis algorithms, whereas, on the other hand, we wanted to offer a foretaste, so to speak, of the numerous novel challenges that such a scenario poses.

Alongside those considered in the article, indeed, there are many other challenges that remain open, and we deem S-Hock as a very good starting point to address them. When considering spectator crowds, for instance, capturing groups of people whose members hold pre-existing relationships (e.g. a family, a group of friends) is certainly a hard task for the classical approaches of group estimation, since they are usually based on proxemics principles, which are not usable when people occupy fixed positions. Similarly, capturing actions such as pointing or clapping hands is difficult due to the large dimension of the crowd and the dense distribution of the spectators – yet these are crucial actions to understand crowd dynamics in the considered scenarios.

S-Hock is a richer crowd dataset than all other state-of-the-art ones, given that the latter usually annotate (or estimate, as in Zhou et al. (2012)) people's position only, do not encompass ground truth obtained from the crowd members, and hence are viable only for tasks such as counting, tracking and event detection, as in Zhou et al. (2014). On the contrary, we are confident that S-Hock may trigger the design of novel and effective approaches for the analysis of human action and interaction in public, crowded settings. Interesting, brand-new applications that can be developed starting from S-Hock, for example, are the following: attention level calculation, that is, detecting peaks of attention in the crowd or in some of its segments; collective action detection and forecasting, which is particularly intriguing in the considered context since people's actions are intertwined both reciprocally – being different if a person knows his/her neighbors or if they are strangers– and with the game actions on the field –how do people react when the team they support scores a goal or loses the game?. Other challenges are still to be thought. While we wish S-Hock would be a stimulus to such a future endeavor, we would like to close by underlining that our own endeavor would have not been possible if not through interdisciplinarity.

Acknowledgments

This research was originally part of the project *OZ (Osservare l'attenZione)*, funded by the University of Trento during the "Winter Universiade Trentino2013". D. Conigliaro, F. Setti and C. Bassetti have been supported by the VisCoSo project grant, financed by the Autonomous Province of Trento through the "Team2011" funding program. P. Rota and N. Sebe have been supported by the MIUR cluster project Active Ageing at Home. The authors would like to thank Mr. Lino Faedda for his valuable work.

References

Aggarwal, J., Ryoo, M., 2011. Human activity analysis: a review. ACM Comput. Surv. 43 (3), 16:1–16:43. doi:10.1145/1922649.1922653.

Ali, S., Shah, M., 2007. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/CVPR.2007.382977.

Andrade, E.L., Blunsden, S., Fisher, R.B., 2006. Modelling crowd scenes for event detection. In: IEEE International Conference on Pattern Recognition (ICPR), 1, pp. 175–178. doi:10.1109/ICPR.2006.806.

Barinova, O., Lempitsky, V., Kholi, P., 2012. On detection of multiple object instances using hough transforms. IEEE Trans. Pattern Anal. Mach. Intell. 34 (9), 1773–1784. doi:10.1109/TPAMI.2012.79.

Bassetti, C., 2016. New Frontiers in the Study of Social Phenomena: Cognition, Complexity, Adaptation. Springer International Publishing, pp. 117–143. doi:10.1007/978-3-319-23938-5_7. Chapter A Novel Interdisciplinary Approach to Socio-Technical Complexity

Berlonghi, A.E., 1995. Understanding and planning for different spectator crowds. Saf. Sci. 18 (4), 239–247. doi:10.1016/0925-7535(94)00033-Y.

Blumer, H., 1951. Collective behavior. In: McClung Lee, A., Park, R. (Eds.), New Outline of the Principle of Sociology. Barnes & Noble.

Chan, A.B., Vasconcelos, N., 2009. Bayesian poisson regression for crowd counting. In: IEEE International Conference on Computer Vision (ICCV) doi:10.1109/ICCV.2009.5459191.

Chen, S.-C., Wu, C.-H., Lin, S.-Y., Hung, Y.-P., 2012. 2d face alignment and pose estimation based on 3d facial models. In: IEEE International Conference on Multimedia and Expo (ICME) doi:10.1109/ICME.2012.60.

Choi, W., Chao, Y.-W., Pantofaru, C., Savarese, S., 2014. Discovering groups of people in images. In: European Conference on Computer Vision.

Conigliaro, D., Setti, F., Bassetti, C., Ferrario, R., Cristani, M., 2013. Attento: Attention observed for automated spectator crowd analysis. International Workshop on Human Behavior Understanding doi:10.1007/978-3-319-02714-2_9.

Conigliaro, D., Setti, F., Bassetti, C., Ferrario, R., Cristani, M., 2013. Viewing the viewers: a novel challenge for automated crowd analysis. In: International Conference on Image Analysis and Processing (ICIAP) doi:10.1007/978-3-642-41190-8_56.

Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., Murino, V., 2011. Social interaction discovery by statistical analysis of f-formations. In: British Machine Vision Conference (BMVC), pp. 23.1–23.12.

Cristani, M., Pesarin, A., Vinciarelli, A., Crocco, M., Murino, V., 2012. Constructing Ambient Intelligence: AML Workshops. Springer Berlin Heidelberg, pp. 72–80. doi:10.1007/978-3-642-31479-7_14. Chapter Look at Who's Talking: Voice Activity Detection by Automated Gesture Analysis

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/CVPR.2005.177.

Doermann, D., Mihalcik, D., 2000. Tools and techniques for video performance evaluation. In: International Conference on Pattern Recognition (ICPR), vol. 4, pp. 167–170. doi:10.1109/ICPR.2000.902888.

Dollár, P., Appel, R., Belongie, S., Perona, P., 2014. Fast feature pyramids for object detection. IEEE Trans. Pattern Anal. Mach. Intell. 36 (8), 1532–1545. doi:10.1109/TPAMI.2014.2300479.

Dollár, P., Wojek, C., Schiele, B., Perona, P., 2009. Pedestrian detection: A benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/CVPR.2009.5206631.

Eichner, M., Marin-Jimenez, M., Zisserman, A., Ferrari, V., 2012. 2D articulated human pose estimation and retrieval in (almost) unconstrained still images. Int. J. Comput. Vis. 99 (2), 190–214. doi:10.1007/s11263-012-0524-9.

Enzweiler, M., Gavrila, D.M., 2009. Monocular pedestrian detection: survey and experiments. IEEE Trans. Pattern Anal. Mach. Intell. 31 (12), 2179–2195. doi:10.1109/TPAMI.2008.260.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. 88 (2), 303–338. doi:10.1007/s11263-009-0275-4.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 32 (9), 1627–1645. doi:10.1109/TPAMI.2009.167.

Figueiredo, M.A.T., Jain, A.K., 2002. Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Mach. Intell. 24 (3), 381–396. doi:10.1109/34.990138.

Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice-Hall.

Girshick, R., 2015. Fast R-CNN. In: IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448.

Goffman, E., 1961. Encounters: Two Studies in the Sociology of Interaction.

Goffman, E., 1963. Behaviour in Public Places.

Goffman, E., 1981. Forms of talk. Conduct and Communication. University of Pennsylvania Press, Incorporated.

Goode, E., 1992. Collective Behavior. Saunders College Pub..

Gowsikhaa, D., Abirami, S., Baskaran, R., 2014. Automated human behavior analysis from surveillance videos: a survey. Artif. Intell. Rev. 42 (4), 747–765. doi:10.1007/s10462-012-9341-3.

Heath, C., Hindmarsh, J., Luff, P., 2010. Video in Qualitative Research. SAGE Publications.

Hung, H., Kröse, B., 2011. Detecting f-formations as dominant sets. In: International Conference on Multimodal Interfaces (ICMI), pp. 231–238.

Jacques Júnior, J.C.S., Musse, S.R., Jung, C.R., 2010. Crowd analysis using computer vision techniques. IEEE Signal Process. Mag. 27 (5), 66–77. doi:10.1109/MSP.2010.937394.

Kemelmacher-Shlizerman, I., Basri, R., 2011. 3d face reconstruction from a single image using a single reference face shape. IEEE Trans. Pattern Anal. Mach. Intell. 33 (2), 394–405. doi:10.1109/TPAMI.2010.63.

Kendon, A., 1988. Goffman's approach to face-to-face interaction. Erving Goffman: Exploring the interaction order 14–40.

Kontschieder, P., Fiterau, M., Criminisi, A., Rota Bulo, S., 2015. Deep neural decision forests. In: IEEE International Conference on Computer Vision (ICCV) doi:10.1109/ICCV.2015.172.

Kratz, L., Nishino, K., 2009. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/CVPR.2009.5206771.

Kratz, L., Nishino, K., 2010. Tracking with local spatio-temporal motion patterns in extremely crowded scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/CVPR.2010.5540149.

Krausz, B., Bauckhage, C., 2012. Loveparade 2010: automatic video analysis of a crowd disaster. Comput. Vis. Image Understanding 116 (3), 307–319. doi:10.1016/j.cviu.2011.08.006.

Laptev, I., 2005. On space-time interest points. Int. J. Comput. Vis. 64 (2), 107–123. doi:10.1007/s11263-005-1838-7.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1 (4), 541–551. doi:10.1162/neco.1989.1.4.541.

Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos., N., 2010. Anomaly detection in crowded scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/CVPR.2010.5539872.

McPhail, C., 1991. The Myth of the Madding Crowd.

Mehran, R., Moore, B.E., Shah, M., 2010. A streakline representation of flow in crowded scenes. In: European Conference on Computer Vision (ECCV) doi:10.1007/978-3-642-15558-1_32.

Navarathna, R., Lucey, P., Carr, P., Carter, E., Sridharan, S., Matthews, I., 2014. Predicting movie ratings from audience behaviors. In: IEEE Winter Conference on Applications of Computer Vision.

Orozco, J., Gong, S., Xiang, T., 2009. Head pose classification in crowded scenes. In: British Machine Vision Conference (BMVC).

Poppe, R., 2010. A survey on vision-based human action recognition. Image Vis. Comput. 28 (6), 976–990. doi:10.1016/j.imavis.2009.11.014.

Psathas, G., 1995. Conversation Analysis: The Study of Talk-in-Interaction. SAGE Publications.

Raghavendra, R., Del Bue, A., Cristani, M., Murino, V., 2011. Abnormal crowd behavior detection by social force optimization. In: Human Behavior Understanding, pp. 383–411. doi:10.1007/978-3-642-25446-8_15.

Rodriguez, M., Laptev, I., Sivic, J., Audibert, J.-Y., 2011. Density-aware person detection and tracking in crowds. In: IEEE International Conference on Computer Vision (ICCV) doi:10.1109/ICCV.2011.6126526.

Sacks, H., Jefferson, G., Schegloff, E.A., 1995. Lectures on Conversation. Wiley.

San Biagio, M., Crocco, M., Cristani, M., Martelli, S., Murino, V., 2013. Heterogeneous auto-similarities of characteristics (HASC): Exploiting relational information for classification. In: IEEE International Conference on Computer Vision (ICCV) doi:10.1109/ICCV.2013.105.

Sand, P., Teller, S., 2008. Particle video: long-range motion estimation using point trajectories. Int. J. Comput. Vis. 80 (1), 72–91. doi:10.1007/s11263-008-0136-6.

Setti, F., Hung, H., Cristani, M., 2013. Group detection in still images by f-formation modeling: a comparative study. In: International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), pp. 1–4.

Setti, F., Lanz, O., Ferrario, R., Murino, V., Cristani, M., 2013. Multi-scale f-formation discovery for group detection. In: IEEE International Conference on Image Processing (ICIP).

Setti, F., Russell, C., Bassetti, C., Cristani, M., 2015. F-formation detection: individuating free-standing conversational groups in images. PLoS ONE 10 (5).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/CVPR.2015.7298594.

Tosato, D., Spera, M., Cristani, M., Murino, V., 2013. Characterizing humans on riemannian manifolds. IEEE Trans. Pattern Anal. Mach. Intell. 35 (8), 1972–1984. doi:10.1109/TPAMI.2012.263.

Tran, K.N., Bedagkar-Gala, A., Kakadiaris, I.A., Shah, S.K., 2013. Social cues in group formation and local interactions for collective activity analysis. In: International Conference on Computer Vision Theory and Applications (VISAPP), vol. 1, pp. 539–548.

Vascon, S., Mequanint, E.Z., Cristani, M., Hung, H., Pelillo, M., Murino, V., 2016. Detecting conversational groups in images and sequences: a robust game-theoretic approach. Comput. Vis. Image Understanding 143, 11–24.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/CVPR.2001.990517.

Wu, B., Nevatia, R., 2007. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. Int. J. Comput. Vis. 75 (2), 247–266. doi:10.1007/s11263-006-0027-7.

Zhong, H., Shi, J., Visontai, M., 2004. Detecting unusual activity in video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/CVPR.2004.1315249.

Zhou, B., Tang, X., Zhang, H., Wang, X., 2014. Measuring crowd collectiveness. IEEE Trans. Pattern Anal. Mach. Intell. 36 (8), 1586–1599. doi:10.1109/TPAMI.2014.2300484.

Zhou, B., Wang, X., Tang, X., 2012. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/CVPR.2012.6248013.

Zhu, X., Ramanan, D., 2012. Face detection, pose estimation, and landmark localization in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/CVPR.2012.6248014.