

The S-HOCK Dataset: Analyzing Crowds at the Stadium

Davide Conigliaro¹, Paolo Rota², Francesco Setti³, Chiara Bassetti³, Nicola Conci⁴, Nicu Sebe⁴, and Marco Cristani¹

¹University of Verona , ²Vienna Institute of Technology , ³ISTC–CNR (Trento) , ⁴University of Trento

Abstract

The topic of crowd modeling in computer vision assumes a single generic typology of crowd, which is very simplistic. In this paper we adopt a widely accepted taxonomy for crowds, focusing on a particular category, the spectator crowd, which is formed by people “interested in watching something specific that they came to see” [5]. This can be found at the stadiums, amphitheatres, cinema, etc. In particular, we propose a novel dataset, the Spectators Hockey (S-HOCK), which deals with hockey matches during an international tournament. The dataset considers 4 hockey matches, where hundreds of spectators are individually annotated, capturing fine grained actions such as hands on hips, clapping hands, watching the cellphone etc., for a total of more than 100 millions of annotations. Analyzing people at the stadium addresses different computer vision tasks, some of them are classic (crowd counting), while other are brand new (as the spectator categorization). For this reason, S-HOCK comes also with a set of protocols for dealing with all of them, and a set of baselines and novel approaches that define the best scores on all the tasks. Anyway, the performances are far from being errorless, and this witnesses the difficulty of the problem and that much can be done in the future.

1. Introduction

Capturing and understanding crowd dynamics is a problem which is important *per se*, under diverse perspectives. From sociology to public safety management, modeling and predicting the crowd presence and its dynamics, possibly preventing dangerous activities, is absolutely crucial.

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

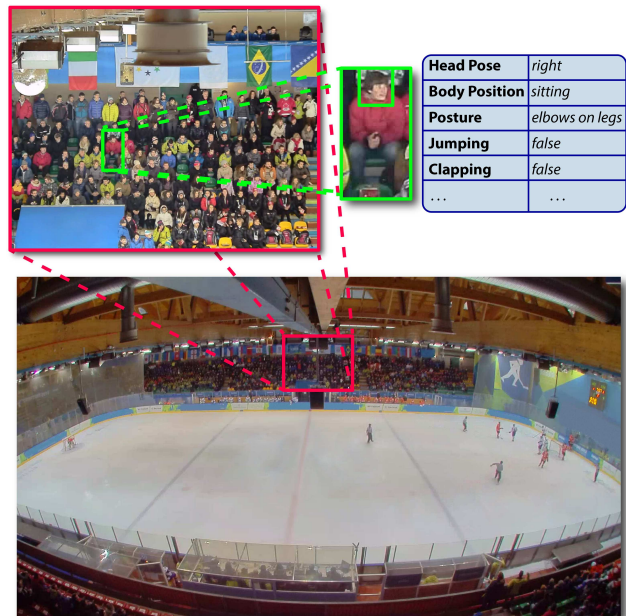


Figure 1. Example of images collected for both the spectators and the rink, plus the annotations.

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 as multiresolution histograms [44], spatio-temporal cuboids [23], appearance or motion descriptors [3], spatio-temporal volumes [27], dynamic textures [29], calculating on top of the flow information. Such information is then employed to learn different dynamics like Lagrangian particle dynamics [35], and in general fluid-dynamic models. The most important applications of crowd analysis are abnormal behavior detection [29], detecting/tracking individuals in crowds [24], counting people in crowds [7], identifying different regions of motion and segmentation [39].

All of these approaches assume a general and unique kind of crowd, while a thorough analysis of the sociological literature offers a taxonomy which could be very interesting

for computer vision. In particular, crowds – better defined as large gatherings [16, 17, 30] – can be divided into four broad categories:

1. *prosaic* [30] or *casual* [6, 18] crowds, where members have little in common except their spatio-temporal location (e.g., a queue at the airport check-in counter);
2. *demonstration/protest* [30] or *acting* [6, 18] crowd, a collection of people who gather for specific protest events (e.g. mob/riot/sit-in/march participants);
3. *spectator* [30] or *conventional* [6, 18] crowds, a collection of people who gather for specific social events (e.g. cinema/theatre/sport spectators), and within which one may find
4. *expressive* crowds [6, 18], a collection of people who gather for specific social events and want to be member of the crowd, to participate in “crowd action” (e.g. flash-mob dancers, mass participants, sport supporters).

Considering this taxonomy, we can certainly say that all the approaches in the computer vision literature focus primarily on casual [24, 35, 7], and protest crowds [25], with hundreds of approaches and ten of datasets, while none of them deals with the spectator crowd and its expressive segments.

This is a critical point: from a recent statistics of 2014 conducted by UK Home Office¹, disorders at stadiums caused 2,273 arrests only considering the FA 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 world².

These statistics motivated in several countries emergency plans for ensuring a better safety and order management, and it is here where computer vision may consistently help. This paper goes in this direction, focusing on the analysis of the spectator crowd, and offering the first dataset on the subject, S-HOCK.

S-HOCK focuses on an international hockey competition (12 countries from all around the world have been invited) which has been held in Trento (Italy) during the 26th Winter Universiade, focusing on the final 4 matches of the tournament.

The dataset has many cues that make it unique in the crowd literature, and in general in the surveillance realm. The dataset analyzes the crowd under different levels of details, offering labeling for each one of them: at the lowest level, it gives the number and position of all the spectators. At the medium level, it gives a fine grained specification of all the actions performed by each single person, such as bending, applauding, the head orientation etc. At the higher

level, it models the network of social connections among the public (who knows whom in the neighborhood), what is the supported team, what has been the best action in the match, etc. This information is summarized by an impressive number of annotations, collected over a year of work: more than 100 millions of double checked annotations. This permits potentially to deal with hundreds of tasks, few of them are documented here, all of them aimed at understanding and predicting the crowd behavior.

Other than this, the dataset is multidimensional, in the sense that offers not only the view of the crowd (at different resolutions, with 4 cameras) but also on the matches. This multiplies the number of possible applications that could be assessed, investigating the reactions of the crowd to the actions of the game, opening up to applications of summarization and content analysis. Besides these figures, S-HOCK is significantly different from all the other crowd datasets, since the crowd as a whole is mostly static and the motion of each spectator is constrained within a limited spatial surrounding of his position.

Together with the annotations, in this paper we discuss about some tasks which focus on the low and high level of details of the crowd analysis, namely, the people detection and the head pose estimation for the low level analysis, and the fan identification for the high level analysis. Fan identification is a kind of crowd segmentation, where the goal is to find the team supported by each one of the spectator. This task is intuitively useful to segregate the different supporter teams, and individuates “hot” zones in which the two teams are mixed. For all of these applications, we define the experimental protocols, so that future researches could easily and fairly compare with us.

From the experiments we conducted, we show how standard methods for crowd analysis, which work well on state-of-the-art datasets, do not fit the type of data we are dealing with, thus requiring us to face the problem from a different perspective. For this reason, together with the baseline approaches, we also propose customized approaches for the spectator crowd, which fit better the scenario at hand, defining new upper bounds.

Summarizing, the contributions of this work 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 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 define the state of the art;
- A sociologically motivated taxonomy of crowds, which individuates four different crowd types, two of

¹Football-related arrests and football banning order statistics, Season 2013–14, available online at <http://goo.gl/j9yYYQ>.

²See at <http://goo.gl/xMU2Zf>.

Annotation	Typical Values
People detection	full body bounding box [$x, y, \text{width}, \text{height}$]
Head detection	head bounding box [$x, y, \text{width}, \text{height}$]
Head pose*	left, frontal, 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 pelvis 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)

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 in the supplementary material. The meaning of the head pose attributes will be explained later in the paper. [*] For the experiments in Sec. 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 rink.

which have never been investigated in computer vision, the spectator crowd and its expressive segments.

The rest of the paper is organized as follows: The details of the data collection and labeling are reported in Sec. 2; the tasks of people detection, head pose estimation, and fan identification are introduced in Sec. 3, focusing on contextualizing the problem, discussing the related state of the art, presenting the considered baselines and our approach, and discussing the results obtained. Finally, in Sec. 4, other applications worth investigating are briefly discussed, which will further promote our dataset as a future benchmark for spectator crowd analysis.

2. Data Collection & 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 among athletes and spectators. The data collection campaign focused on the last 4 matches (those with more spectators) held in the same ice-stadium, and was conducted by a team of 6 people, 4 of them collecting questionnaires and the remaining at the

cameras: in particular we used 5 cameras: a full HD camera (1920×1080, 30 fps, focal length 4mm) for the ice rink and another one for a panoramic view of all the bleachers, and 3 high resolution cameras (1280×1024, 30 fps, focal length 12mm) focusing on different parts of the spectator crowd. In total, 20 hours of recordings have been collected, with inter-camera synchronization: this brought to the interesting feature of having the crowd synchronized with the game on the rink.

After the match, we asked to a percentage of uniformly distributed spectators (30%) to fill a simple questionnaire with three questions (whose significance will be clear later in the paper):

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

In this dataset we propose game segments from different hockey matches in order to stress the generalization capability of the algorithms discussed in this paper, since in each match we have different people and illumination conditions. In particular, from each match we selected a pool of sequences highlighting different situations inside the rink (goals, saves, timeouts, etc.), with each video 31 seconds long (930 frames), for a total of 75 sequences: 15 for each one of the 4 cameras on the crowd and 15 sequences on the game. These sequences were chosen 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). [Paolo: The full annotation reported in Tab. 1 has been performed on one among the three close-field cameras, whereas in the other two are annotated only with the survey information. The forth 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. The 5 cameras are synchronized.]

Each sequence has been annotated frame by frame, spectator by spectator, by a first annotator, using the ViPER format [10]³. Such annotator had to perform three different macro tasks: detection (localizing the body and the head), posture and action annotation, respectively. [Paolo: This amounted to deal with a set of 50 labels, listed in Table 1⁴].

Among the whole set of possible features that can characterize the human dynamics, we selected the annotated *elementary forms of action* [30] as strictly connected with the analysis of social interaction, and related to our specific setting, i.e. sport spectator crowd. In particular, we drew from available literature on (a) social interaction, with particular attention to non-verbal conduct (proxemics, bodily posture,

³The toolkit is available at <http://viper-toolkit.sourceforge.net/>

⁴The database is available at <http://mmlab.disi.unitn.it/SHock>

Method	Simple Search			Grid 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 [38]	0.365	0.642	0.465	0.357	0.685	0.469
ACF [11]	0.491	0.622	0.548	0.524	0.649	0.580
DPM [14]	0.502	0.429	0.463	0.423	0.618	0.502
CUBD [13]	0.840	0.303	0.444	0.613	0.553	0.581

Table 2. Quantitative results of people detection methods, with and without the grid-arrangement prior. [Davide: QUI CHIEDONO LA ROC]

gesture, etc.), especially in public places; and (b) the so-called *crowd behavior*, i.e. social interaction in large gatherings [16, 17, 30], in particular sport spectator gatherings.

Each annotator had two weeks to annotate 930 frames, and was asked to do it in a lab, in order to monitor him/her and ensure a good annotation quality. After all the sequences have been processed, producing a total amount of more than 100 millions of annotations [Paolo: Forse ci sono da rivedere sti numeri alla luce di quanto scritto 3 paragrafi sopra.], a second round of annotations started, with the “second annotators” that had to find bugs in the first-round annotation phase. The whole work involved 15 annotators and lasted almost 1 year, with all annotators paid for their duties.

3. Evaluation

In this section we present a set of possible applications and analysis that can be conducted on the proposed dataset. In particular we focus on two classical applications, such as people detection and head pose estimation, and one more interesting application from the social point of view, such as crowd segmentation based on the behavior of its members. For each one of the mentioned topics we briefly present the state of the art taking into account only the methods applicable in this particular scenario and some preliminary experiments conducted on our dataset. We also propose some ways to improve the performance by exploiting the specific structure of the crowd and the relation between the crowd behavior and taking into account what is happening in the hockey rink.

3.1. People Detection

People detection is a standard and still open research topic in computer vision, with the HOG features [9] and the Deformable Part Models [14] as workhorses, and plenty of alternative algorithms [12]. Unfortunately, most of the methods in the state of the art are not directly usable in our scenario, mostly for two reasons: low resolution – a person has an average dimension of 70×110 pixels – and occlusions – usually only the upper body is visible, rarely the entire body and sometimes only the face.

Recently, some works studied how to improve the performance of detectors by means of an explicit model of the

visual scene. Specifically, focusing on people detection in crowded scenes, Barina et al. [4] used the Hough transform to overcome the non-maxima suppression stage for detecting multiple instances of the same object, while San Biagio et al. [38] proposed a new descriptor able to treat complex structural information in a compact way. To overcome occlusion issues, Wu and Nevatia [43] used a number of weak part detectors based on edgelet features and Eichner et al. [13] fused DPM [14] and Viola-Jones [33] detectors to identify upper bodies. Finally, Rodriguez et al. [36] proposed to optimize a joint energy function combining crowd density estimation and the localization of individual people.

In this work we provide 5 different baselines for people detection, from the simplest algorithms to the state of the art method for object detection, showing how in this scenario the simplest method get very high scores due to the problems listed above.

The first method we consider is a simple detector based on HOG [9] features (cell size of 8×8 pixels) and a linear SVM classifier (HOG+SVM). Similarly, the second method only differs in the descriptor we use, which in this case is the Heterogeneous Auto-Similarities of Characteristics (HASC) descriptor [38]. We use the same sliding window as in the previous case to generate the map and the detections. We will refer to this method as HASC+SVM.

We also test 3 state-of-the-art methods for people detection: (1) the Aggregate Channel Features (ACF) detector [11] uses the Viola-Jones framework to compute integral images (and Haar wavelets) over the color channels, fusing then together; (2) the Deformable Part Model (DPM) [14] combines part’s templates arranged in a deformable configuration fed into a latent SVM classifier; and (3) the Calvin Upper Body Detector (CUBD) [13], a combination of the DPM framework trained on near-frontal upper-bodies – i.e. head and upper half of the torso of the person – and the Viola-Jones face detector.

On top of all these methods, we propose an extension based on the strong prior we have in our kind of crowd, i.e. the people are “constrained” by the environment to arrange in a grid – the seats on the bleachers. Assuming a regular grid (considering the camera perpendicular at the plane of the bleachers and ignoring distortion effects) and considering the fact that since people are more likely to locate on the same rows and columns, we can just add to the detection confidence map the average of the map over the rows and the columns. Consider $D(x, y)$ the output of the detector for the patch (x, y) , the modified output for a target location (\hat{x}, \hat{y}) is:

$$\tilde{D}(\hat{x}, \hat{y}) = D(\hat{x}, \hat{y}) + \sum_i D(x_i, \hat{y}) + \sum_j D(\hat{x}, y_j)$$

In the case there is a distortion due to the camera point of view, this could be easily recovered by using Hough trans-


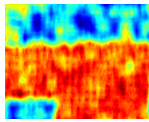
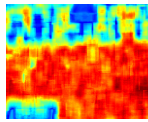
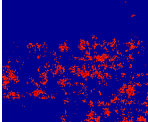
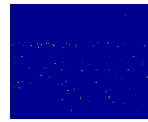
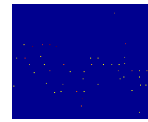
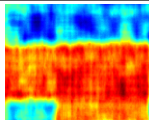
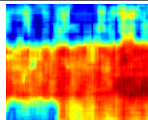
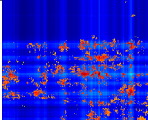
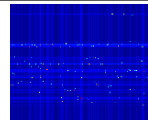
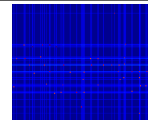
	Image	HOG+SVM	HASC+SVM	ACF	DPM	CUBD
No prior						
With prior						

Figure 2. Qualitative results for people detection algorithms. Detection confidence map for each method with and without imposing the grid-arrangement prior. (best viewed in color)

form for detecting the “principal directions” and summing over these new lines.

As experimental protocol, we use two videos from a single game for training and two from different games for validation, leaving the 11 sequences from the final for testing. A set of 1,000 individuals randomly selected from the training videos are used as positive samples, while a background image is used to generate the negative samples for training. Then, 20 random frames from the validation videos are used to tune the best parameters for minimum detection score threshold and the non-maxima suppression parameters. A subsampling of 1 frame every 10 for each video is used for testing, resulting in 1,000 images and 150,000 individuals. While ACF, DPM and CUBD have their own searching algorithms to generate candidate bounding boxes, for HOG+SVM and HASC+SVM we consider a sliding window of 72×112 px with a step of 8px, generating a detection confidence map of 160×118 patches. A threshold on the minimum detection score and a non-maxima suppression stage have been applied to generate the predicted detections.

We consider an individual as correctly identified if the [Davide: intersection area between the predicted and annotated bounding boxes is more than 50% of the union of the two rectangles by the formula

$$B_p \cap B_{gt} > \frac{(B_p \cup B_{gt})}{2} \quad (1)$$

] where $B_p \cap B_{gt}$ denotes the intersection of the predicted and ground truth bounding boxes and $B_p \cup B_{gt}$ their union. As performance measures we use precision, recall and F_1 scores.

A qualitative evaluation of the baselines and the grid arrangement prior contribution is in Fig. 2, while quantitative results are in Table 2. We can notice how the best performing method is the HOG+SVM, while part based frameworks (i.e. DPM and CUBD) perform poorly in their standard version; this is probably due to the low resolution of the person bounding boxes which makes it very difficult to detect single parts like arms and legs. By introducing our

proposed prior, we can see how all the methods increase their performances, and in particular CUBD increases of about 10%, becoming one of the best detectors for this kind of scenario. As a result of the introduction of our grid-arrangement prior, an average improvement of about 5% in F_1 score is achieved.

3.2. Head Pose Estimation

Once the body has been detected, and the head has been localized, a consequent operation to be carried out is the head pose estimation. It represents another low-level operation, essential for many medium and high level tasks, for example capturing the focus of attention of the spectators, correlating it with the action in the ice rink.

The literature on head pose estimation is large and heterogeneous as for the scenarios taken into account; most of the approaches assume that the face is prominent in the image, as in a multimodal interaction scenario, and rely on the detection of landmark points [47, 22, 8]. Here these solutions are inapplicable since the faces are too small (50x40 pixels on average). In a low resolution domain the work proposed by Orozco et al. [32] seems to fit better, relying on the computation of the mean image for each orientation class. Distances w.r.t. the mean images are used as descriptors and fed into SVMs. In [42], the authors exploit an array of covariance matrices in a boosting framework. The image of the head is divided into different patches, that are weighted depending on their description capability. On S-HOCK these methods are performing roughly the same in terms of classification accuracy, with a huge time consumption (see Tab. 3).

In order to overcome this issue, we propose two novel approaches based on Deep Learning, with comparable results but much faster. The choice of Deep Learning is motivated by the large number of effective approaches in the object recognition literature, witnessing its versatility in many scenarios [28, 21, 26, 41, 40].

In particular, we evaluate the performance of the Convolutional Neural Network (CNN) and the Stacked Auto-encoder Neural Network (SAE) architecture. In both meth-

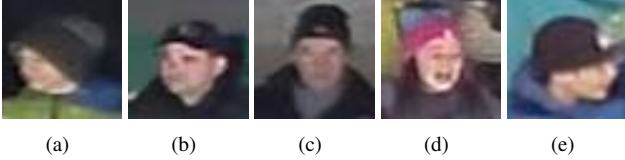


Figure 3. Examples of the five head poses considered for the experiments in Sec. 3.2; in order (a) to (e): *far left, left, frontal, right, far right*.

Method	AVG Accuracy	Training time [sec]	Testing time [sec]
Orozco et al. [32]	0.368	105303	6263
WARCo [42]	0.376	186888	87557
CNN	0.346	16106	68
SAE	0.348	9384	3
CNN + EACH	0.354	16106	68
SAE + EACH	0.363	9384	3

Table 3. 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.

ods we feed the Neural Network with the original image, resized to a standard size of 50x50 pixels, so as to have uniform descriptors. The CNN is composed by 5 layers: 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 iterations. The SAE architecture is depicted in Fig. 4 (b), the input images are fed into an auto-encoder with hidden layers of size $h = 200$, trained separately. A second training phase is performed on the neural network initialized with the weights learned in the previous stage. Both training procedures are refined in 100 epochs.

As for the experimental protocol is the same as in the previous case, except for the fact that there is no validation set; all the training sequences are employed to extract a total of 107299 heads while the testing set is composed by 34949 heads from the testing sequences. In this experiment, we take as input the head locations coming from the ground truth, in order to derive a sort of upper bound on the estimation performances. In this respect, faces are annotated as *frontal, left, right, far left* and *far right*. 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*. This has been detailed to the annotators during the data labeling (see Fig. 3).

In Tab. 3 shows the results of the current state of art methods compared with the two proposed approaches. The overall accuracy spans within a range of 3% for Orozco et

al. [32], WARCo [42], CNN and SAE but in neural networks approaches the computation workload is much less. This speed up in classification time for both training and testing phases makes our method more suitable for real life applications where a quick response and an imminent decision is required. As a further remark, we trained WARCo by randomly sampling 5000 samples among all those available for training, this has been necessary for the huge computation time required to learn the model in case of using the whole set of data.

In case of large sport events the spectator crowd attention tends to be attracted by the location of the action. This observation can be exploited to benefit the final classification. For this reason we propose an additional experiment named EACH (Event Attention Catch). In order to accomplish this task we consider the ice rink as our universe of locations where the puck can be. We are not interested in the pitch information of the head so we reduce the rink to a monodimensional space. We model the position of the puck such as a Gaussian distribution over all the possible locations and we consider it as a prior probability in order to refine the final head pose estimation. This probability $P_A^{(c)}$ is formalized in Eq. (2)

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

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.

$$c = \arg \max_c (\alpha P_A^{(c)} + (1 - \alpha) P_N^{(c)}) \quad (3)$$

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

We observe that this model is much more beneficial when players are playing with respect to when the game is paused by a foul. This particular aspect suggests us to tune the α parameter according to the game phase. The results reported in Tab. 3 are computed using $\sigma = 15$ and $\alpha = 0.3$. The ice rink information increases the accuracy by approximately 2% on both CNN and SAE frameworks.

3.3. Spectators Categorization

In our dataset the *spectators categorization* task consists in finding different groups of people among the spectators. The result of this segmentation is strictly related to the behavior of the people and thus we are able to cluster people supporting different teams by considering their reaction during specific game actions, e.g. goals, saves, etc.

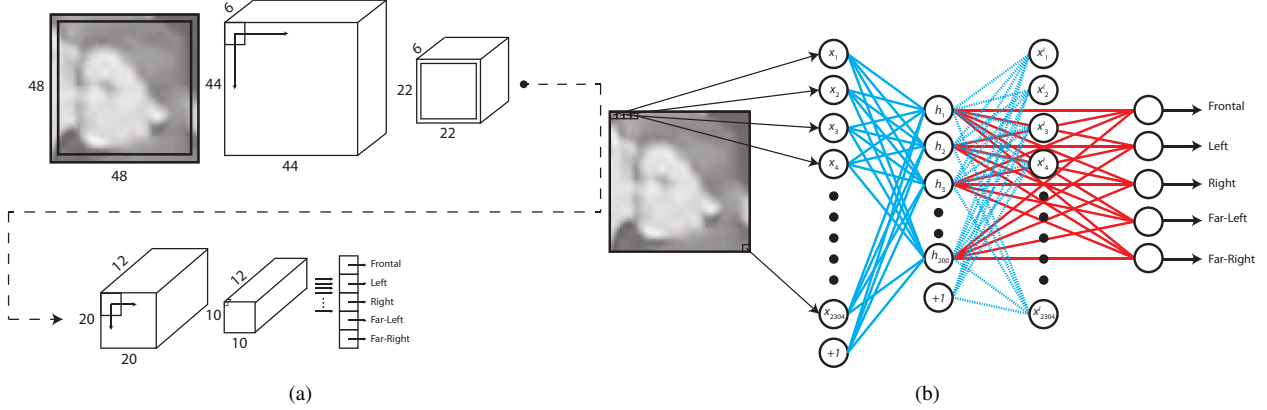


Figure 4. (a)Architecture of CNN. (b) SAE architecture: in cyan are pictured the interconnections between the auto-encoder that must be trained separately, in red instead there are the interconnections of the final NN.

Method	Accuracy
AS2007 [2]	0.592
MMS2010 [31]	0.559
Our	0.621

Table 4. Spectators categorization accuracy obtained from the normalized confusion matrix.

Spectators categorization can be considered a subtask of the crowd behavioral analysis, which is generally associated with human activity analysis [19, 34, 1]. As stated by Jaques et al. [20], in computer vision there are two main approaches for crowd behavior analysis: the *object-based* where the crowd is considered as a collection of individuals, and the *holistic* approach which treats the crowd as a single entity. This second approach is the one that best suits the spectator crowd analysis because it directly tackles the problem of dense occluded crowds. The holistic approaches usually start from optical flow to extract global information from the crowd. In [2], the authors use Lagrangian particle dynamics to segment the flow; here the notion of a flow segment is equivalent to a group of people that perform a coherent motion.

More recently, Mehran et al. [31] propose a streakline representation of flow to address various computer vision problems, including crowd flow. They use a simple watershed segmentation of streaklines to cluster regions of coherent motion.

Both these works and several datasets proposed in the literature focus on pedestrian crowds [37, 2], instead with S-HOCK we propose a crowd with different dynamics and behavior, where the people are assumed to stay near a fixed location for most of the time and their movements are limited to their position. For this reason, the works listed above require some adjustments in order to be applied to the spectators categorization task.

In this paper we also present a new method for specta-

tors categorization, whose framework can be extended to previous methods. As most of the holistic approaches, our method also starts from dense optical flow computation. Then we decompose the flow map into a set of overlapping patches and we describe each patch with five features: x and y coordinates of the patch’s centroid, the average flow intensity I (over all the pixels belonging to the patch), the entropy of flow intensity E_I and directions E_D (both directions and intensities are quantized to compute the entropy).

These feature vectors are then passed to a Gaussian clustering with automatic model selection [15], obtaining an instantaneous spatial clustering of the patches for each frame. Following, we perform a temporal segmentation based on the similarity between patches: we will call it Patch Similarity History (PSH).

Let consider the matrix H_τ^f where each entry $H_\tau^f(i, j)$ measures the similarity between patches p_i and p_j considering the evolution of patches’ labels until frame f of the video. $H_\tau^f(i, j)$ can be computed as:

$$H_\tau^f(i, j) = \begin{cases} H_\tau^{f-1}(i, j) + \tau, & \text{if } \Psi^f(i, j) = 1 \\ \max(0, H_\tau^{f-1}(i, j) - \delta), & \text{otherwise} \end{cases}$$

where τ decides the temporal extent of the similarity in term of frames duration, δ is the decay parameter and $\Psi^f(i, j)$ is an update function defined as:

$$\Psi^f(i, j) = \begin{cases} 1, & \text{if } Lab_i^f = Lab_j^f \\ 0 & \text{otherwise} \end{cases}$$

Lab_i^f and Lab_j^f indicating the labels associated to patches p_i and p_j at the same frame f , during the previous spatial clustering. In such way PSH represents a similarity matrix since it describes how much two patches are similar over time, depending on the spatial cluster to which they belong. By computing the reciprocal of PSH we can obtain a distance matrix and use it to perform a complete linkage hier-

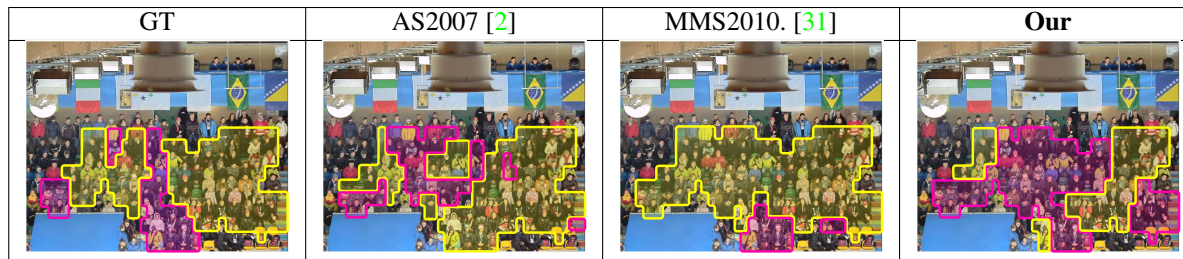


Figure 5. Qualitative results for spectators categorization. The colored areas represent the two groups of spectators supporting different team.

archical clustering. The result is a dendrogram where clusters are defined by cutting branches off the tree. Depending on the height of the cut we can obtain diverse spectators categorization where the clustered patches could represent the whole crowd or a subset of it.

In order to set up the same test protocol for all the methods we divide the scene into overlapping patches. We created a grid of $N_p=585$ patches with size 64×128 px and half a patch size of overlap. Each patch is associated with a ground truth label of the person’s bounding box with the highest overlapping area (if any). The main difference between our method and those in the literature, lies in the fact that the outputs of these methods are based on a frame-by-frame pixels-wise segmentation. So in order to adapt them to our test protocol, we assign to each patch a predicted label corresponding to the most voted within the patch area.

Each method was tested using the standard setting given by the authors. The parameters of the PSH τ and δ have been set respectively to 30 and 1. For the methods that use optical flow it was computed every 10 frames. Table 4 shows the accuracies resulting from the spectators categorization tasks. Instead Figure 5 shows the qualitative evaluation. The results show that the proposed method is able to categorize the spectator better than the other methods with an accuracy of 62.1%. Since the temporal segmentation was the same for all methods, the best result obtained by our method is probably due to the features extracted from each patch. In fact we are able to describe the behavior of the people from the patches, considering how much they move (with the intensity I) and describing the kind of movement (with the flow entropy E_D and E_I). Further results regarding the confusion matrices of the experiments and other qualitative results could be found in supplementary materials.

4. Conclusions

This paper has introduced S-HOCK, a novel dataset focusing on a brand-new issue, the modeling of the spectator crowd. Our main goal is to promote the potentialities of our benchmark, whose features have been barely exploited in the applications we have taken into account here: actually,

we have focused on some low-level, traditional tasks (people detection and head pose estimation) and a novel, high-level challenge, the spectator categorization. This choice has been motivated from the fact that on one side we wanted to show the impact that a similar scenario has on the realm of already existent classification algorithms; on the other side, we wanted to disclose one of the many new challenges that a spectator crowd scenario does offer, as the spectator categorization. At the same time, we have shown that many are the ways with which the performances of modern algorithms can be improved, and that novel challenges request novel solutions, making the spectator crowd an exciting problem to be faced. Many other are the open issues: at a medium level of detail, capturing actions as hugging, clapping hands etc. would be difficult due to the dimension and the dense distribution of the spectators; at a high level of details, understanding groups of people that know themselves will be certainly hard for the classical approaches of group estimation; in facts, they are usually based on proxemics principles, here not usable due to the fixed positions of the people. [Davide: Further high level analysis could be on excitement detection (detecting the peak of excitement of the crowd), crowd/people action forecasting, which is also intriguing since in this case we may consider the behavior of a person as being influenced by the neighbors and by the game, etc. Such applications make S-HOCK richer compared to all other crowd datasets, where usually only the position of the people is annotated (or in some cases estimated, as in [46]), without ground truth obtained also from the people in the crowd itself, and where only tasks of counting, tracking and event detection can be assessed, as in [45].] Therefore, we are confident that S-HOCK may trigger the design of novel and effective approaches for the analysis of the human behavior.

[Davide:

Acknowledgments

F. Setti and D. Conigliaro are supported by the VIS-COSO project grant, financed by the Autonomous Province of Trento through the “Team 2011” funding programme.]

References

- [1] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, 2011. 7
- [2] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*, 2007. 7, 8
- [3] E. L. Andrade, S. Blunsden, and R. B. Fisher. Modelling crowd scenes for event detection. In *ICPR*, 2006. 1
- [4] O. Barinova, V. Lempitsky, and P. Kholi. On detection of multiple object instances using hough transforms. *PAMI*, 34(9):1773–1784, 2012. 4
- [5] A. Berlonghi. Understanding and planning for different spectator crowds. *Safety Science*, 18:239–247, 1995. 1
- [6] H. Blumer. Collective behavior. in: A. McClung Lee (ed.) principles of sociology. Barnes & Noble. In *New Outline of the Principle of Sociology*. 1951. 2
- [7] A. B. Chan and N. Vasconcelos. Bayesian poisson regression for crowd counting. In *ICCV*, 2009. 1, 2
- [8] S.-C. Chen, C.-H. Wu, S.-Y. Lin, and Y.-P. Hung. 2d face alignment and pose estimation based on 3d facial models. In *ICME*, 2012. 5
- [9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 4
- [10] D. Doermann and D. Mihalcik. Tools and techniques for video performance evaluation. 2000. 3
- [11] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *PAMI*, 36(8):1532–1545, 2014. 4
- [12] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, 2009. 4
- [13] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. Articulated human pose estimation and search in (almost) unconstrained still images. Technical report, ETH Zurich, 2010. 4
- [14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010. 4
- [15] M. A. T. Figueiredo and A. Jain. Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):381–396, 2002. 7
- [16] E. Goffman. *Encounters: Two studies in the sociology of interaction*. 1961. 2, 4
- [17] E. Goffman. *Behaviour in Public Places*. 1963. 2, 4
- [18] E. Goode. *Collective behavior*. 1992. 2
- [19] D. Gowsikhaa, S. Abirami, and R. Baskaran. Automated human behavior analysis from surveillance videos: a survey. *Artificial Intelligence Review*, 42(4):747–765, 2014. 7
- [20] J. Jacques, Junior, S. Raupp Musse, and C. Jung. Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine*, 27(5):66–77, 2010. 7
- [21] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *PAMI*, 35(1):221–231, 2013. 5
- [22] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *PAMI*, 33(2):394–405, 2011. 5
- [23] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, 2009. 1
- [24] L. Kratz and K. Nishino. Tracking with local spatio-temporal motion patterns in extremely crowded scenes. In *CVPR*, 2010. 1, 2
- [25] B. Krausz and C. Bauckhage. Loveparade 2010: Automatic video analysis of a crowd disaster. *CVIU*, 116(3):307–319, 2012. 2
- [26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012. 5
- [27] I. Laptev. On space-time interest points. *IJCV*, 64(2-3):107–123, 2005. 1
- [28] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, 2009. 5
- [29] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, 2010. 1
- [30] C. McPhail. *The myth of the madding crowd*. 1991. 2, 3, 4
- [31] R. Mehran, B. Moore, and M. Shah. A streakline representation of flow in crowded scenes. In *ECCV*. 2010. 7, 8
- [32] J. Orozco, S. Gong, and T. Xiang. Head pose classification in crowded scenes. In *BMVC*, 2009. 5, 6
- [33] M. J. P. Viola. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 4
- [34] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010. 7
- [35] R. Raghavendra, A. Del Bue, M. Cristani, and V. Murino. Abnormal crowd behavior detection by social force optimization. In *HBU*, pages 383–411. Springer, 2011. 1, 2
- [36] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert. Density-aware person detection and tracking in crowds. In *ICCV*, 2011. 4
- [37] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert. Data-driven crowd analysis in videos. In *ICCV*, 2011. 7
- [38] M. San Biagio, M. Crocco, M. Cristani, and V. Martelli, S. and Murino. Heterogeneous auto-similarities of characteristics (hasc): Exploiting relational information for classification. In *ICCV*, 2013. 4
- [39] P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. *IJCV*, 80(1):72–91, 2008. 1
- [40] J. M. Susskind, G. E. Hinton, J. R. Movellan, and A. K. Anderson. Generating facial expressions with deep belief nets. *Affective Computing*, pages 421–440, 2008. 5
- [41] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 5
- [42] D. Tosato, M. Spera, M. Cristani, and V. Murino. Characterizing humans on riemannian manifolds. *PAMI*, 35(8):1972–1984, 2013. 5, 6
- [43] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *IJCV*, 75(2):247–266, 2007. 4

- [44] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *CVPR*, 2004. [1](#)
- [45] B. Zhou, X. Tang, H. Zhang, and X. Wang. Measuring crowd collectiveness. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(8):1586–1599, Aug 2014. [8](#)
- [46] B. Zhou, X. Wang, and X. Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2871–2878, June 2012. [8](#)
- [47] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. [5](#)