ATTENTO: ATTENTion Observed for Automated Spectator Crowd Analysis

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

[1] Verona University
[2] Institute of Cognitive Sciences and Technologies, CNR
[3] Italian Institute of Technology, IIT

Abstract. We propose a new type of crowd analysis, focused on the *spectator crowd*, that is, people "interested in watching something specific that they came to see" [1]. This scenario applies on stadiums, amphitheaters etc., and shares some aspects with classical crowd monitoring: actually, many people are simultaneously observed, so that per-person analysis is hard; however, here the dynamics of humans is more constrained, due to the architectural environment in which they are situated; specifically, people are expected to stay in a fixed location most of the time, limiting their activities to applaud, support/heckle the players or discuss with the neighbors. In this paper, we start facing this challenge by considering hockey matches, locating a videocamera 25-30 meters far from the bleachers, pointing at the crowd: in this scenario, aggregations of spectators that exhibit similar behavior are detected, and the behavior is classified into a set of predefined classes, highlighting the overall excitement. To these aims, in a first step we focus on individual frames, clustering local flow measures into spatial regions. The clustering is then extended by adding the temporal axis into the analysis, looking for non-randomic spatio-temporal clusters; to this aim, the Lempel-Ziv complexity is considered. This way, choral activities can emerge, indicating for example fan groups belonging to different teams. After this, with the adoption of entropic measures, the degree of excitement of such groups can be quantified.

Keywords: spectator crowd, Lempel-Ziv complexity, spatio-temporal clustering

1 Introduction

Crowd analysis is a videosurveillance topic that focuses on large masses of people, where the single person cannot be finely characterized, due to the small visual resolution, the frequent total occlusions and the complex dynamics. As a consequence, crowd analysis has grown with its own set of peculiar techniques, most of them avoiding to perform standard surveillance operations (as example, people detection, tracking, gesture recognition); instead, motion flow information [2, 3] is usually exploited as ingredient of higher level descriptors (as multiresolution histograms [4], spatiotemporal cuboids [5], appearance or motion descriptors [6], spatiotemporal volumes [7], dynamic textures [8]), which eventually are fed into standard classifiers.

In this paper, we focus on a novel applicative field for crowd analysis, centered on the modeling of the so called *spectator crowd* [1]. The idea is to observe people while they are watching a public show, as in a sport arena, a movie theater, a classroom, a court, and recording and analyzing their activities. This scenario differs substantially from those analyzed by the typical crowd modeling techniques: due to *territoriality* principles, people are assumed to stay near a fixed location for most of the time, i.e., their seat [9, 10], while what is mainly being monitored in the crowd analysis literature are moving people. In addition, people here are assumed to have a strong relation with the event or contest they are watching, that becomes a kind of reference point, where the focus of attention [11] of the crowd is located, and around which the space is structured. In classical crowd modeling no such clear reference point is present.

In this new scenario, diverse techniques and applications can be developed, generalizing the videosurveillance context to the multimedia realms of the entertainment and the edutainment:

- **Spectators segmentation**: finding diverse groups of people among the spectators, for example the fans of the opposite teams in a sport match; attentive VS distracted students in a classroom; enthusiastic VS annoyed spectators at a theater play; in the entertainment and edutainment fields, detecting when the audience is annoyed may trigger reactive mechanisms that for example inform the speaker/teacher that something should be done for rekindle the observers;
- **Excitement calculation**: in a given time interval, quantizing the level of excitement of some part or of the entire crowd; this could be beneficial for example for marketing purposes.
- **Event segmentation**: segmenting diverse activities of the crowd (clapping hands, making a wave, heckling), and studying how these activities are related with the observed event (i.e. some people clap their hands when the favorite team scores a goal, or get excited when a foul is or is not signaled by the referee);
- **Augmented video summarization**: the spectator feedback, automatically recognized, may help in highlighting exciting or crucial events that should be included in a video summarization of the show;
- **Comparative analysis of spectators**: various factors can be compared, like fans of different teams in the same sport [12], or fans of different sports [13], where spectators are arranged differently etc.;
- **Interpretation of crowd's intentions**: discriminating whether a display of crowd excitement is determined by a rejoicing VS aggressive attitude, to foresee the subsequent crowd's behavior.

In the following, we will show how the first two aspects discussed above, i.e., spectators segmentation and excitement calculation, can be faced using Social Signal Processing methods [14–16], where Social Psychology and sociological notions are incorporated into Computer Vision and Pattern Recognition algorithms. In particular, we will focus on a sport scenario, where people watch hockey matches[1]. In this scenario, assuming a single camera capturing the whole crowd, located at 25-30 meters from the bleachers, we divide the acquired scene into squared patches, extracting from each patch local flow information (position, flow intensity and direction). These features are then fed into a per-frame Gaussian clustering framework [17], that groups together patches with similar dynamics. These instantaneous associations are then summed along the temporal axis, creating a spatio-temporal similarity matrix; such matrix is weighted by a factor (the Lempel-Ziv complexity [18]) which indicates how randomic the instantaneous associations have been during the entire sequence. Finally, the weighted spatio-temporal similarity matrix becomes the input of a single-link hierarchical clustering, whose dendrogram models the different fan groups, even when they are spatially merged.

Concerning the excitement calculation, entropic measures based on the optical flow are exploited, to indicate how much lively the fan groups are. The rationale is that the excitement for a group of people is high when we have a strong, various and coordinated dynamic activity.

The remaining of the paper is organized as follows: Sec. 2 explores the related literature, showing how the proposed problem has interesting connections with some studies in sociology that could constitute a foundation for the visual analysis of behavior. Our framework is presented in Sec. 3, followed by preliminary results in Sec. 4; Sec. 5 draws some conclusions and future perspectives.

2 Related Literature

Under a computer vision perspective, not many approaches modeling spectator crowds watching sports are present in the literature. Conversely, the sociological realm exhibits some relevant studies. Schweingruber and McPheil in [19] has built a model for characterizing “collective actions-in-common”, i.e., actions performed spontaneously by several people in coordination. This study, though not specifically centered on viewers, but rather on various forms of crowd, singles out seven dimensions for the analysis of crowd behavior: orientation (facing), vocalization (producing sounds other than words with mouth), verbalization (uttering words), vertical locomotion (movement of the body over the same point on the ground), horizontal locomotion (movement of the body from one point on the ground to another), gesticulation (meaningful bodily configuration based on fingers, hands, and arms movements mainly), and manipulation (using hands to applaud or to strike, carry, throw, pull, etc.). Such study is interesting from our point of view as it includes most of the behaviors we intended to observe

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

in spectators crowds. Other studies, as [20], have challenged the idea that a crowd can be seen as an undistinguished collection of individuals, highlighting how crowds can rather be segmented in subgroups of different size, composition and organization, based on their previous acquaintance, on common goals etc. Starting from this idea, one of the aims of this paper is exactly that of trying to identify how the crowd may be segmented with automatic means. Turning to spectator crowds, some scholars have discussed how collective behavior, like applauding, is generated in contexts where a crowd is attending to a public event, for example in public speeches, as is the case for [21]. Regarding the sociology of sport, many works have been produced, but most of them deal with violence in sport, as shown in [22], where the motivations that bring people to watch sports alive are also discussed. Finally, few works have specifically considered fans of hockey teams (for instance [23]), analyzing how their support can influence the players. In this work we have taken inspiration from the analyses displayed in the works mentioned above and focused on extracting *locomotion* and *gesticulation* features from the video recordings of the spectators.

3 Our framework

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

3.1 Spectators segmentation and excitement calculation

As a first step, standard motion flow is computed on the image plane, extracting at each pixel direction and intensity. Then, assuming people as static [9, 10] and considering the size of people, flow information can be re-arranged into a grid of N squared patches $\{x\}$. On each patch x, at each time frame, we extract four measures: the first is the flow intensity $I(x)$, obtained by averaging over the flow intensity values of the patches' pixels; intuitively, this cue encodes how much movement characterizes a patch. The second cue is the flow direction entropy $E_{\text{dir}}(x)$, calculated over the related flow direction values (opportunely quantized); $E_{\text{dir}}(x)$ describes the kind of movement in the patch: high entropy values mean random directions, while low values address homogeneous movement in the patch. The last two measures are the x, y patch centroid coordinates. In other words, at each time step, each patch is described as a 4D point.

The segmentation occurs in two steps: first, a Gaussian clustering with automatic model selection [17] is applied on the values of all the patches in a given time frame. This way, an instantaneous grouping is inferred. This process is replicated for all the T frames. We define then the $N \times N \times T$ matrix M: at time t, $M(i, j, t) = 1$ if patches i and j are in the same cluster, 0 otherwise. At this point, we want to check the non-randomness of the associations of two given patches i, j: the rationale is that non randomic associations are more plausibly indicating a strong relation among the patches. Therefore, for each couple of patches, we calculate the Lempel-Ziv complexity [18]: high complexity indicates a randomic aspect in the temporal evolution of the association between i and j. After this, we compute the summation over t of the matrix M, obtaining the $N \times N$ matrix M_T. As final spatio-temporal similarity among patches, we thus take

$$S = M_T \bullet \frac{1}{LZ} \tag{1}$$

where LZ indicates the $N \times N$ matrix containing the Lempel-Ziv complexity scores, and "$\bullet$" indicates point to point multiplication. The rationale is that we want to highlight associations that are strong in time (several instantaneous associations) which are also non randomic. Making the similarity matrix S as a distance (computing the point-to-point reciprocal), it is possible to perform single link hierarchical clustering, and to obtain the spectator segmentation, which partitions the scene in regions where the behavior of the crowd is similar (in terms of the measures quoted above).

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

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

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

Finally, the average of $Exc(r)$ over all frames is considered as the excitement cue in a given interval for the region r.

4 Experiments

In order to test our framework, we built a novel repository which consists of videos taken during the 2013 IIHF Ice Hockey U18 World Championship, played in Asiago from the 7th to the 13th of April 2013. In particular, two entire matches were recorded (Italy VS Norway, Italy VS Slovenia), each by two cameras, mounted frontally at a distance of about 25 meters from the spectators' stand. Each camera was pointing at an half of the whole stand, the zoom being fixed. Therefore, for each match we have two sequences, further divided in 3 as the times of the hockey play. This resulted in 12 videos at 30 fps, with a resolution of 640x480 pixels for a total duration of about 6 hours.

From this dataset, we extracted 6 videos, each lasting three minutes, and we focused on the spectators segmentation and the excitement calculation (see Sec. 1). The ground truth labels have been manually set for each frame of the videos, by assigning to the patches one among these classes: background (patches with no people); quiet spectators (patches with people who don't exult); excited spectators (patches with standing up and exulting people).

The videos have been analyzed considering a grid of rectangular patches of size 40×80, with the patches overlapping for an half of their size, in both dimensions. Flow has been computed on the entire scene every 10 frames, so we have

3 processed frames per second; after that, the flow direction has been quantized in five values (up, down, left, right, none) where the fifth value corresponds to all those flow vectors whose intensity is inferior to a given threshold $I = 0.8$.

4.1 Spectators segmentation and excitement calculation

For each video, we first computed the frame-based Gaussian clustering and subsequently the temporal hierarchical clustering considering separately the two matrices M'_T and S', which are the distance matrices (computing the point-to-point reciprocal) of M_T and S, respectively. We consider the matrix M'_T without the Lempev-Ziv complexity to investigate how important is to take into account randomness in the processing. This way, for each window we get interesting spectators segmentations, clearly explaining the occurred events; for longer windows, the segmentation tends to discriminate solely the presence of the crowd against the background. The clustering results obtained by operating directly on M'_T and S' were compared, in order to determine which distance matrix is the most informative.

In order to measure the agreement between the ground truth labels and the two clustering procedures, we computed four different indices of external validity, that is: Rand index, adjusted Rand index, Jaccard index, Fowlkes-Mallows index [24]; furthermore, we calculate an error rate, which is the percentage of error class labels of the clustering solution compared with true labels, adopting the figure of merits suggested in [25]. The results are shown in Fig. 1. All the boxplots of the four indices show that Lempel-Ziv complexity improves the segmentation/clustering results; also the error rate is lower using S'.

The picture on the left in Fig. 2 shows an example of ground truth labels assigned to the patches. The three different colors are a measure of the excitement of each class. Here the background region includes 196 patches which have been assigned the blue color. The spectators are divided into two classes: the *quiet spectators* region (light blue) includes 102 patches, while the region of major excitement (dark red) includes 43 patches with cheering and clapping people.

A first interesting result is that the excitement level, automatically computed for each region by eq. 2, is consistent with what the expert has indicated in the ground truth. In particular, the three clusters corresponding to background, quite spectators and excited spectators generates excitement levels of 0.0033, 0.0075 and 0.0217, respectively.

In Fig. 2, the center and right images show the segmentation obtained by considering the methods with M'_T and S'. At a glance, the colors and the region size of the last image suggest that the method which uses S' is closer to the ground truth. This is confirmed by the results showed in the row V_1 of Table 1, where are shown the error rates for the excitement level computation, calculated as:

$$Err(r) = \frac{|Exc_{\mathrm{gt}}(r) - Exc_{\mathrm{x}}(r)|}{Exc_{\mathrm{gt}}(r)} \tag{3}$$

where $Exc_{\mathrm{gt}}(r)$ is the average excitement level computed for the ground truth region r over all the video and $Exc_{\mathrm{x}}(r)$ is the same value computed for the

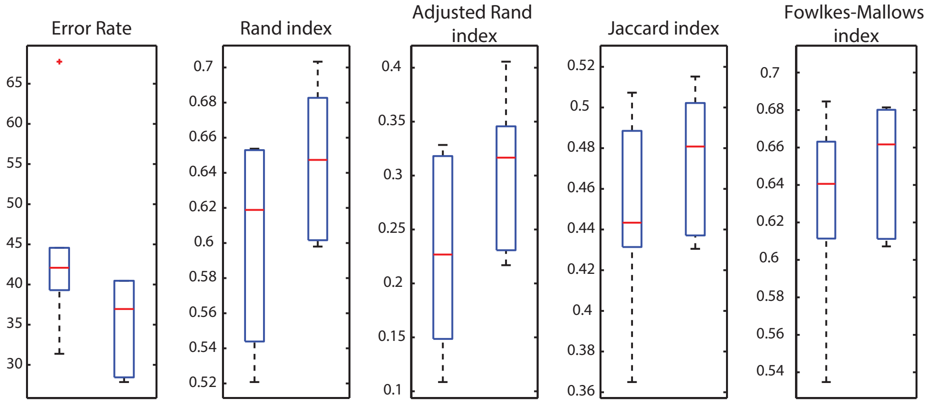

Fig. 1: Boxplots of the five different external validity indices computed by considering all 6 videos. In each plot the first box refers to the clustering results obtained by operating directly on M'_T an the second on S'. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, respectively; the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually as red crosses. Low values in error rate and high values in other indexes indicate better segmentation performance.

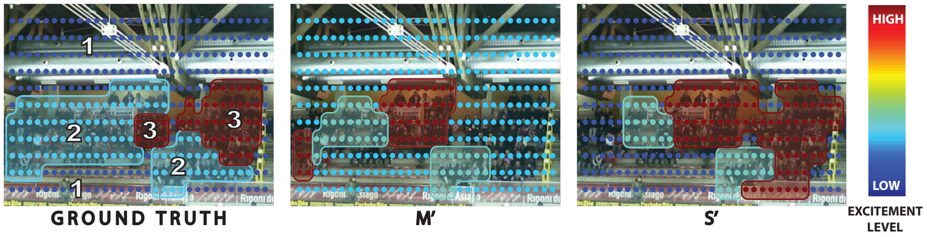

Fig. 2: Spectators segmentation and excitement calculation obtained from the first video. The image on the left shows the ground truth segmentation. The numbers indicate the excitement (in ascending order) estimated by the expert who made the ground truth. Region 1 is the background, region 2 includes quite people and region 3 comprises more excited spectators. The center and right images refer to the segmentation obtained by using M'_T and S', respectively. Each dot in the images is plotted at the center of the respective patch, and its color depends on the mean excitement level of the region to which it belongs.

region r obtained with our two method of segmentation, by using M'_T and S'. The mean error rate for each video and region, highlighted in the last row of the table, shows that the regions obtained by the segmentation on the matrix S' achieve an excitement level more similar to the ground truth, than the method that exploits M'_T.

	M'_T			**S'**		
	R_1	R_2	R_3	R_1	R_2	R_3
V_1	0.578	0.138	0.263	0.122	0.024	0.172
V_2	0.339	0.167	0.119	0.394	0.119	0.038
V_3	0.646	0.372	0.451	0.646	0.650	0.270
V_4	1	0.371	0.543	0.620	0.193	0.354
V_5	0.174	0.113	0.228	0.167	0.201	0.039
V_6	0.527	0.940	0.065	0.331	0.830	0.302
Average	**0.544**	**0.350**	**0.278**	**0.380**	**0.336**	**0.195**

Table 1: Error rates $Err(r)$ for excitement level estimation. M'_T and S' indicates the two different methods of segmentation, ignoring or considering the randomness value given by the Lempev-Ziv complexity, respectively. R_1, R_2 and R_3 are the background and the regions of quite and excited spectators, respectively. $V_1 \ldots V_6$ indicate each tested video. The last row shows the average value of each column.

5 Conclusions

The study of spectators crowd dynamics offers new perspectives in the crowd modeling field. In this paper we have performed a preliminary study, first of all reasoning on the possible applications that can be developed in such a scenario, and presenting effective implementations for some of them; in particular, we have shown how spectators can be segmented on the basis of their behavior, and how their excitement level can be inferred by looking exclusively at the crowd activity. Much more can be done, by employing more sophisticated models: dynamic Bayesian networks may embed spatial and temporal reasoning in a unique model; gesture recognition, face detection and expression recognition may provide detailed cues to better understand the nature of the spectators activities, allowing the discrimination between supporting, heckling or just watching, absent in the present work. Further developments may be achieved by adopting different sensors, like infrared and pan-tilt-zoom cameras. Audio analysis can also be fruitful, especially in the case of sport events: in absence of facial information, capturing whistles instead of shouts of joy may be crucial to perform sentiment analysis of the crowd. From a sociological point of view, proxemics principles can be taken into account [26], in order to assess if social relations can be discovered by analyzing interpersonal distances among seated people, following the same idea of [27].

An important theme to be inquired is the establishment of the ground truth for such kinds of scenarios. In this paper we have adopted a sort of “expert-based ground truth”, in that we have compared our findings with what had been explained in sociological theories. Alternatively, a more complete approach of this kind (expert) would be based on an ethnographic study: in that case the ground truth would be built on the basis of participant observation carried out by several ethnographers (team ethnography), doing fieldwork on the stands of an arena, stadium, amphitheater, etc. This, moreover, could be complemented with ethnomethodologically oriented videoanalysis (see [28]). A completely different approach to ground truth would be to found it in a more "bottom-up" way, by asking directly to those belonging to the crowd, either exactly the crowd that was attending the recorded event, or, more generically, people that can report about an experience of participation to a public event as a viewer. Even in this case, there are various ways to implement such approach, ranging from structured questionnaires to in-depth interviews. Notwithstanding all that have already been mentioned, of course privacy and ethical issues should also be taken more seriously into account in the nearest future developments of this study.

References

1. Berlonghi, A.: Undestanding and planning for different spectator crowds. Safety Science **18** (1995) 239–247
2. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 935–942
3. Raghavendra, R., Del Bue, A., Cristani, M., Murino, V.: Abnormal crowd behavior detection by social force optimization. In: Proceedings of the Second international conference on Human Behavior Unterstanding. HBU'11, Berlin, Heidelberg, Springer-Verlag (2011) 134–145
4. Zhong, H., Shi, J., Visontai, M.: Detecting unusual activity in video. In: CVPR (2)'04. (2004) 819–826
5. Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: CVPR. (2009) 1446–1453
6. Andrade, E.L., Blunsden, S., Fisher, R.B.: Modelling crowd scenes for event detection. In: Proceedings of the 18th International Conference on Pattern Recognition - Volume 01. ICPR '06, Washington, DC, USA, IEEE Computer Society (2006) 175–178
7. Laptev, I.: On space-time interest points. Int. J. Comput. Vision **64**(2-3) (September 2005) 107–123
8. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. (June 2010) 1975 –1981
9. Guyot, G.W., Byrd, G.R., Caudle, R.: Classroom setting: An expression of situational territoriality in humans. Small Group Behavior **11** (1980) 120–128
10. Kaya, N., Burgess, B.: Territoriality. seat preferences in different types of classroom arrangements. Environment and Behavior **39**(6) (2007) 859–876
11. Goffman, E.: Behaviour in Public Places. Free Press of Glencloe. Notes on the Social Organization of Gatherings, New York (1963)
12. Roadburg, A.: Factors precipitating fan violence: a comparison of professional soccer in britain and north america. The British Journal of Sociology **31**(2) (1980) 265–276
13. Goldstein, J.H., Arms, R.L.: Effects of observing athletic contests on ostility. Sociometry **34**(1) (1971) 83–90
14. Vinciarelli, A., Pantic, M., Bourlard, H.: Social Signal Processing: Survey of an emerging domain. Image and Vision Computing Journal **27**(12) (2009) 1743–1759

15. Cristani, M., Murino, V., Vinciarelli, A.: Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2010). (2010) 51–58
16. Cristani, M., Raghavendra, R., Del Bue, A., Murino, V.: Human behavior analysis in video surveillance: A social signal processing perspective. Neurocomputing **100** (January 2013) 86–97
17. Figueiredo, M.A.T., Jain, A.: Unsupervised learning of finite mixture models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **24**(3) (2002) 381–396
18. Kaspar, F., Schuster, H.G.: Easily calculable measure for the complexity of spatiotemporal patterns. Phys. Rev. A **36** (Jul 1987) 842–848
19. Schweingruber, D., MacPheil, C.: A method for Systematically Observing and Recording Collective Action. Sociological Methods Research **27**(4) (1999) 451–498
20. McPhail, C.: From clusters to arcs and rings: Elementary forms of sociation in temporary gatherings. Research in Community Sociology **Supplement 1** (1994) 35–57
21. Atkinson, J.M.: Public speaking and audience responses: some techniques for inviting audience applause. In J. Maxwell Atkinson, J.H., ed.: Structures of Social Action: Studies in Conversation Analysis. Cambridge University Press, Cambridge (1984) 370–407
22. McDonald, M., Milne, G., Hong, J.: Motivational factors for evaluating sport spectator and participant markets. Sport Marketing Quarterly **11**(2) (2002) 100–113
23. Bowker, A., Boekhoven, B., Nolan, A., Bauhaus, S., Glover, P., Powell, T., Taylor, S.: Naturalistic observations of spectator behavior at youth hockey games. Sport Psychologist **23** (2009) 301–316
24. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Information Systems **17**(2-3) (2001) 107–145
25. Wang, K., Wang, B., Peng, L.: Cvap: Validation for cluster analyses. Data Science Journal **8** (2009) 88–93
26. Hall, E.T.: Handbook for proxemic research. Anthropology News **36**(2) (1995) 40–40
27. Cristani, M., Paggetti, G., Vinciarelli, A., Bazzani, L., Menegaz, G., Murino, V.: Towards computational proxemics: Inferring social relations from interpersonal distances. In: SocialCom/PASSAT 2011. (2011) 290–297
28. Heath, C., Hindmarsh, J., Luff, P.: Video in Qualitative Research. Analysing Social Interaction in Everyday Life. Sage, London (2010)