SOCIAL INTERACTION ANALYSIS IN VIDEOS, FROM WIDE TO CLOSE PERSPECTIVE

UNIVERSITY OF TRENTO - Italy

Department of Information Engineering and Computer Science

PAOLO ROTA

Advisors: Prof. Dr. **Nicola Conci** and Prof. Dr. **Nicu Sebe**

February 23rd 2015

CONTENTS

1 INTRODUCTION 5

2 FAR-RANGE ANALYSIS: AN OVERVIEW ON CROWD MONITORING 9

2.1 Crowd Grouping through Particle Social Motion Analysis 10

2.1.1 Particles Mutual Influence 11

2.1.2 Entity Grouping 13

2.1.3 Results 16

2.2 Spectator Crowd Analysis 19

2.2.1 Data Collection & Annotation 21

2.2.2 Evaluation 22

2.3 Discussion 30

3 MID-RANGE APPROACH: PROXEMIC ANALYSIS 33

3.1 Proxemics as Important High-level Feature 35

3.1.1 Methodology 36

3.1.2 Proxemics parameters 36

3.1.3 Feature Ectraction 38

3.1.4 Classification procedure 39

3.1.5 Results 40

3.2 Music Room 42

3.2.1 Related Works 45

3.2.2 Music Room 48

3.2.3 Evaluation 54

3.3 Discussion 60

4 LOCALIZING UNSTRUCTURED SOCIAL INTERACTIONS IN REAL-LIFE SCENARIO 63

4.1 Related Work 66

4.2 Real-life Events - Dyadic Interactions Dataset (Re-DID) 68

4.3 Evaluation Framework 69

4.3.1 Dense Trajectories and Visual Features 69

4.4 Results 72

4.4.1 Fight Detection and Localization 73

4.4.2 Fighting vs Dancing 76

4.5 Conclusion 77

5 CONCLUSION AND FUTURE WORK 79

5.1 Future Work 80

BIBLIOGRAPHY 84

1

INTRODUCTION

In today's digital age, the enhancement of the hardware technology has set new horizons on the computer science universe, asking new questions, proposing new solutions and re-opening some branches that have been temporary closed due to the overwhelming computational complexity. In this sense many algorithms have been proposed but they have never been successfully applied in practice up to now.

In this thesis we are focusing our attention on the *computer vision* scenario where such technological enhancement has lately played a paramount role. This phenomena has also influenced the researchers' way of thinking, pushing forward the limits and opening new frontiers boosting the results in several research areas. For instance, the problem of visual tracking has been tackled for many years as a combination of target's acquisition and chasing. The task itself encloses many problems, from camera positioning, miss detections, false positives and low frame rate processing to name a few. Even though we cannot say that this is a problem completely solved, a huge improvement has been done in managing occlusion, single and multi-camera re-identification and on temporal models even in situation in which the objects to be tracked are in a considerable number, pushing this task closer to real-time applications.

The recent boosting in computer performance has not only affected long date *computer vision* issues as previously mentioned, it has also inspired the researchers to move forward, exploring a wider plot giving birth to brand new topics that have never been considered before by this research community. Semantic scene analysis is a well-fitting example. Semantic is the branch of linguistic that studies the meaning of the words. Recently the growing interests for *Big Data* has incentivized the researchers to combine semantics and advanced machine learning algorithms in order to find connections among multimedia data, increasing the quality of the outcome.

This technological blowout has also fostered the tendency to interdisciplinarity, promoting applications that are connecting computer vision with a broad spectrum of different disciplines. Social behavioral analysis indeed, stems from this multi-disciplinary tendency, borrowing inputs not only from traditional computer vision

algorithms but also from different concepts taken from other disciplines, including sociology and psychology.

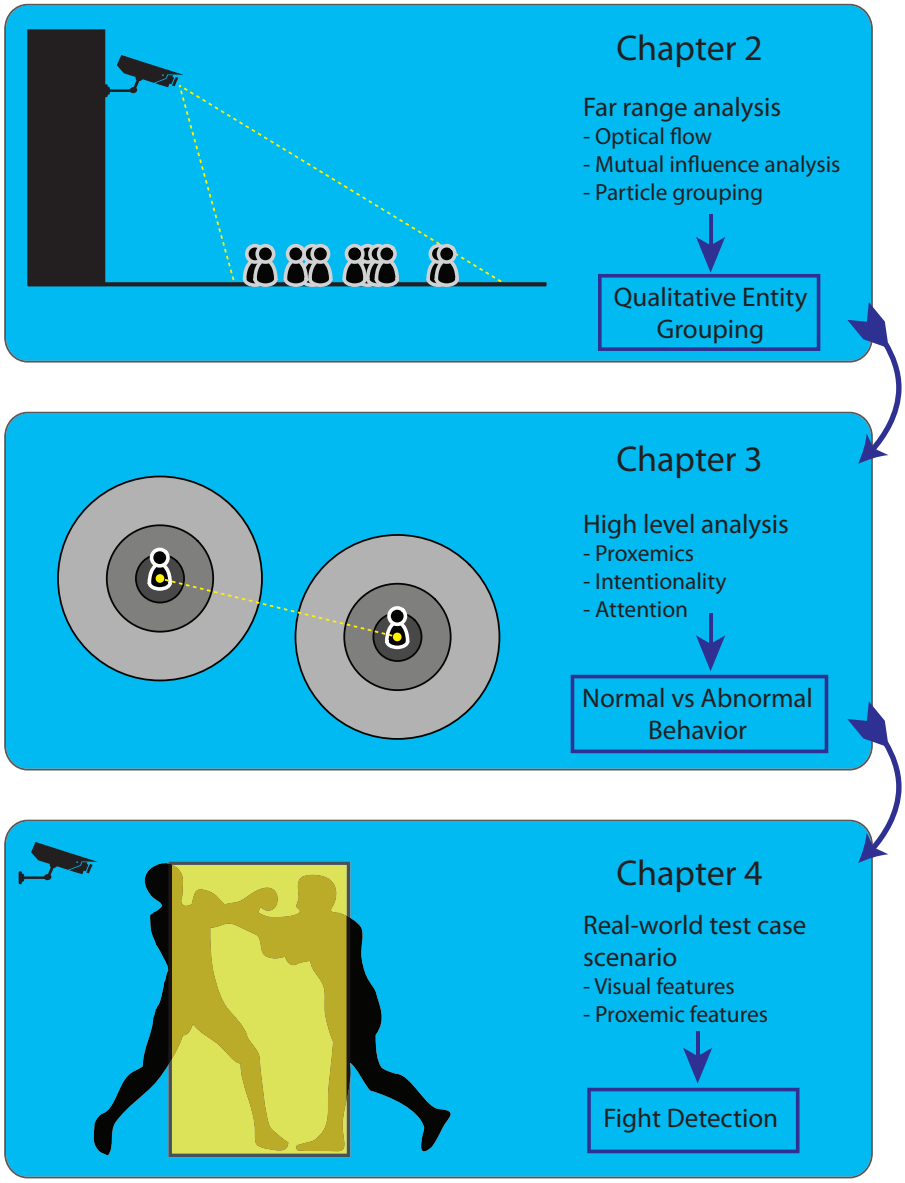


Figure 1: The illustration of the multi-scope approach proposed in this thesis.

In this thesis we are approaching visual behavioral analysis focusing on single person and group interactions. A social interaction differs from an action because of the cooperative purpose among two or more subjects. In computer vision we refer to interactions when more individuals are performing a simple or structured action that requires mutual cooperation between them. Stemming from this concept we are

going to perform an exhaustive multi-scope analysis with the objective of detecting different types of interactions (see Fig. 1).

In video surveillance we are often facing far range monitoring tasks. Placing cameras in such way of covering a wide area is a common strategy in order to breakdown the deployment costs, to augment the coverage of the monitored area and also for structural reasons. However, far-range often means also smaller resolution of the single object with consequential issues in detecting its dynamics. In these cases the most effective analysis is related to the study of motion. In chapter 2 we propose a novel method to group moving objects according to their motion properties. The intuition stems from the dynamics of each moving particle, if similar it is likely to belong to the same group, on the other hand it might appertain to a different moving object. Far-range analysis can give a good qualitative overview of the crowd's situation in the monitored area; however, in order to have a finer examination of people's behavior we need to move closer and detect semantically meaningful objects as, of course, people. In this thesis we are also investigating an original type of crowd that at the best of our knowledge has never been inspected in computer vision before, named *spectator crowd*. We will propose a brand new, over-annotated dataset along with a deep testing of state of art algorithms on people detection and counting, head pose estimation and a novel task, that is of relevant interest for this type of crowd, called *spectator categorization* that aims at finding group of supporters for the different playing teams.

When persons are detectable, we can do a deeper reasoning: when a pair of subjects know each other they have the tendency to adopt a certain type of behavior while interacting. This attitude is remarkably different in the case in which the two interacting persons are in the condition of not knowing each other. These perceptive cues depend not only on their intimacy level, but also on cultural background, religion, character etc. This features might be well rendered by the distance they keep from each other. The branch of semiotics that studies this type of relations is called proxemics and will be discussed in chapter 3. We will analyze the variation of these proxemic parameters in order to discriminate three different types of interacting behaviors, namely: casual/absent interaction, normal interaction and abnormal interaction. In the second part of the chapter, we propose a collaborative application based on proxemic analysis, where the dyadic interaction of a pair of individuals drives an automatic composition algorithm that translates the movements in a customized melody.

The analysis of social interactions is often considered as a natural expansion of action recognition. When we analyze an interaction, however, we are approaching a type of action that has further constraints that are resumed by the compulsoriness of the cooperation between the two interacting parts. For this reason much of the information is enclosed in the mutual causality of their movements. Besides this,

many types of interactive behaviors are strongly different when performed naturally, and the most challenging datasets available in literature are staged, retrieved from movies or performed by amateur actors. In chapter 4 we are contributing to fill this shortage of natural interactions videos focusing mainly on authentic fights. The fight is a particular type of interaction that is not made up of any structure. This makes the detection of it much harder with respect to a simpler one like a hug, a kiss or an handshake. We tackle this problem exploiting the information coming from both interacting individuals, focusing on proxemic cues and visual features.

This final results show with a good amount of confidence, the state of the art in detection and localization of unstructured social interactions in real-world scenario through the mutual contribution of low and high level features.

The main contributions of this thesis are the following:

- A novel method to find particle motion affinity based on cross/self influence matrix, with the objective of grouping entities that have a coherent motion and destination (chapter 2).
- A new dataset for spectator crowd along with game and attendance annotations captured by different cameras (chapter 2).
- A deep state of art testing of crowd analysis algorithms on *spectator crowd*. Discussion of the problems risen by the task and the proposal of some possible solutions (chapter 2).
- The proposal of a novel task related with *spectator crowd* that aims to detect different groups of supporters in the audience relating them to their supporting team (chapter 2).
- A novel method to extract proxemic information in social behavior analysis, combining three different measures related to distance between subjects, intentionality and attention (chapter 3).
- An applicative framework[1] that aims to translate proxemic cues in music using computer vision algorithms exploiting the collaborative creativity of the participants to the experiment (chapter 3).
- A new fully annotated dataset on fight detection in real world scenario (chapter 4).
- A novel method to detect and localize pairwise unstructured physical social interaction based on the definition of an interpersonal space (chapter 4).

1 Exposed during *La notte dei ricercatori 2012*, *ICT days 2013* and at MART museum of Rovereto for 3 weeks in august 2014.

2

FAR-RANGE ANALYSIS: AN OVERVIEW ON CROWD MONITORING

In this chapter we focus on the far-ange social interaction analysis. We consider two different types of crowds. In the first one we look at the crowd as moving deformable objects in a monitored area, we analyze the motion and infer groups of people sharing the same motion information. The second type is a static crowd called Spectator Crowd. In this case we highlight its characteristics and properties, easing the evaluation through the proposal of a brand new massively annotated dataset recorded during an important hockey tournament held in Trentino during December 2013.

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 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 [145], spatio-temporal cuboids [70], appearance or motion descriptors [6], spatio-temporal volumes [76], dynamic textures [80], calculating on top of the flow information. Such information is then employed to learn different dynamics like Lagrangian particle

Parts of this Chapter appear in:

- Rota P., Ullah H., Conci N., Sebe N., De Natale F. G. **Particles cross-influence for entity grouping**, IEEE EUSIPCO 2013.
- *(Submitted)* Conigliaro D., Rota P., Setti F., Conci N., Cristani M., Ferrario R., Bassetti C., Sebe N. **The S-HOCK Dataset: Analyzing Crowds at the Stadium**, IEEE CVPR 2015.

dynamics [109], and in general fluid-dynamic models. The most important applications of crowd analysis are abnormal behavior detection [80], detecting/tracking individuals in crowds [71], counting people in crowds [25], identifying different regions of motion and segmentation [119].

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 [49, 50, 83] – can be divided into four broad categories:

1. *prosaic* [83] or *casual* [13, 51] crowds, where members have little in common except their spatio-temporal location (e.g., a queue at the airport check-in counter);
2. *spectator* [83] or *conventional* [13, 51] crowds, a collection of people who gather for specific social events (e.g. cinema/theatre/sport spectators), and within which one may find
3. *expressive* crowds [13,51], 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).
4. *demonstration/protest* [83] or *acting* [13, 51] crowd, a collection of people who gather for specific protest events (e.g. mob/riot/sit-in/march participants);

In the first part of this chapter (Section 2.1) we will focus on *prosaic/casual* crowd inferring groups of entities in a monitored area through interesting points motion analysis. In the second part (Section 2.2) we will inspect the *Spectator Crowd*, providing a huge and brand new annotated dataset along with some benchmarks for three different computer vision problems.

2.1 CROWD GROUPING THROUGH PARTICLE SOCIAL MOTION ANALYSIS

Understanding human activities through vision-based technologies has been a challenging and attracting research area since the beginning of the past decade [59, 94]. Thanks to the increase in performance of detection and tracking algorithms, the research focus has shifted towards the provisioning of a higher level of interpretation of the visual scene, which includes the identification of semantically meaningful events, such as, for example, role detection [44], people grouping [103], as well as crowd assemblage and crowd flow analysis [126] to name a few.

In [30] the authors exploit the contextual information as a relative contribution of distances and orientations in relation to a reference subject considered as an *anchor*. This information is then processed using a Spatio-Temporal Volume representation

combined with Random Forests, to infer simple individual actions. In the framework of activity recognition, the authors in [5] propose a dual approach (top-down and bottom-up) to classify human activities jointly exploiting multiple detectors information and higher level behavioral features, also based on context. In the work by Lan et al. [74], the authors analyze the contextual information in a sport events so as to infer individual and group behaviors relying on the overall game situation.

However, in case the number of moving subjects in the scene becomes dense, and the chance of incurring in severe occlusions increases, detectors and trackers are likely to fail, and more generic approaches that analyze the motion flow should be exploited, as it commonly occurs in crowd motion analysis [134]. Although in this way the notion of *person* is neglected, due to the absence of an explicit detector, it is still possible to estimate, for example, the density of people, and the aggregation points in the monitored environment. This turns out to be an efficient pre-processing step for any further and more detailed analysis.

The rest of the section is structured as follows. In Section 2.1.1 we introduce the particle based method and the cross-influence computation; in Section 2.1.2 we describe the learning phase and the grouping structure, and in Section 2.1.3 we present the experimental results obtained on the recently released high definition UCLA and BIWI datasets. Conclusions remarks are discussed in Section **??**.

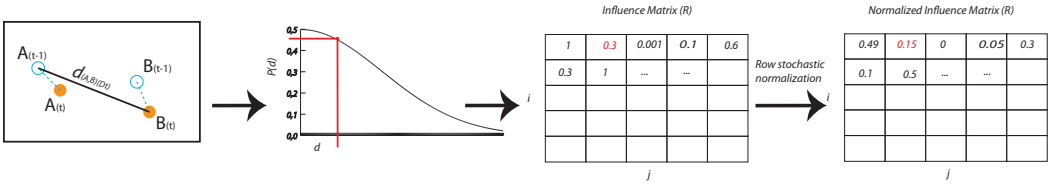


Figure 2: Computation of the particles influence matrix.

2.1.1 *Particles Mutual Influence*

In order to overcome all the problems issued by the appliance of detectors and single person tracking algorithms, we decided to focus on a particle based approach. In our model we assume that each particle corresponds to an entity and has attractive and repulsive forces upon other particles surrounding it. Under this hypothesis, each particle can be classified not only on the basis of its own motion characteristics, but also in relation to the context, in this case provided by its neighbors.

As proposed in [99], the state of an entity $c \in C$ at time t, and defined as $h_t^{(c)}$ can be derived using a Markovian assumption, making it a direct temporal consequence of the state in the preceding temporal observation $(t-1)$. The influence among particles can be expressed by the so-called influence matrix. The influence matrix is

a row stochastic matrix of dimension $C \times C$ where C is the number of the particles (entities) present in the given time window. In order to compute a single value of the matrix (see Fig.2) we assume that each particle relates with the others according to a Gaussian distribution, therefore the influence of c_i on c_j is computed as in Eq. (2.1), where d is the Euclidean distance between c_i and c_j.

$$R^{(c_i,c_j)} = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{d(c_i(t),c_j(t-1))}{2\sigma^2}} \tag{2.1}$$

As can be seen in Eq. (2.1), the motion information is already comprised by the formula, distance is computed indeed between the paticle c_i at time t and the particle c_j at time $t-1$.

Figure 3: An example of particle initialization (top) and after pruning (bottom).

Particles are then classified in two states (h), *grouped* (G) or *alone* (A) according to the model in Eq. (2.2) and Eq. (2.3).

$$S\left(h_t^{(c_i)} \Rightarrow G\right) = \sum_{c_j \in \{1...C\} \wedge c_j \neq c_i} R^{(c_i,c_j)} \times P(G_t / h_{t-1}^{(c_j)}) \tag{2.2}$$

$$S\left(h_t^{(c_i)} \Rightarrow A\right) = R^{(c_i,c_i)} \times P(A_t / h_{t-1}^{(c_i)}) \tag{2.3}$$

In the equations above, R represent the social influence matrix. The highest value of score S will dictate the state of the current particle in the current time window. For the purpose of this work just the *grouped* particles are considered relevant and passed to the feature extraction step described in Section 2.1.2. Particles marked as *alone* are instead pruned and not considered in the further processing steps. A visual example of pruned particles annotated in red is shown in the second row of Fig. 3.

In our work, particles are generated through the Good-Features-To-Track algorithm [124], and tracked by the Lucas-Kanade optical flow [14]. The influence matrix is computed at discrete steps at the end of each tracking period.

2.1.2 *Entity Grouping*

The objective of the features extraction process is to identify low-level information relative to the particles interaction. Features are extracted only for the particles obtained from the mutual influence model. In our approach we have selected the average distance among the particles and their density as two representative elements to infer the interaction among particles. In fact, proximity, which is partially exploited also in the influence model measures the instantaneous relationship among neighboring entities. At the same time, the higher the density of the particles, the higher the chance for them to interact.

For both features, orientation is used as a prior, meaning that particles are considered in the same group, only if their relative offset in terms of direction of motion fall in a predefined range. For the purpose of visualization, in Fig. 4 (a), a set of synthetic entities are shown where a reference entity, annotated in blue, is grouped with the neighboring entities annotated in red. On the contrary, two entites are not included in the same group since their orientations do not conform to the orientation of the reference entity. Moreover, a reference entity annotated in yellow and neighboring entities annotated in red are shown in Fig. 4 (b). These entities constitute a group, as shown in (c), according to the compliance in terms of density and mutual distances with the reference entity.

In order to properly weight the features we have selected for entity grouping, we have trained a feedforward multi-layer perceptron (MLP) neural network. The motivation for exploiting MLP is in its substantial ability, through backpropagation,

to resist to noise, and the dexterity to generalize. To group the particles from the preceding stage of the mutual influence model, the average distance of a reference particle with its neighbors is accumulated and averaged. A particle is only considered for grouping with a reference particle if its relative orientation is compliant with the orientation of a reference particle. The density and average distance of the reference particle are fed as an input to the MLP.

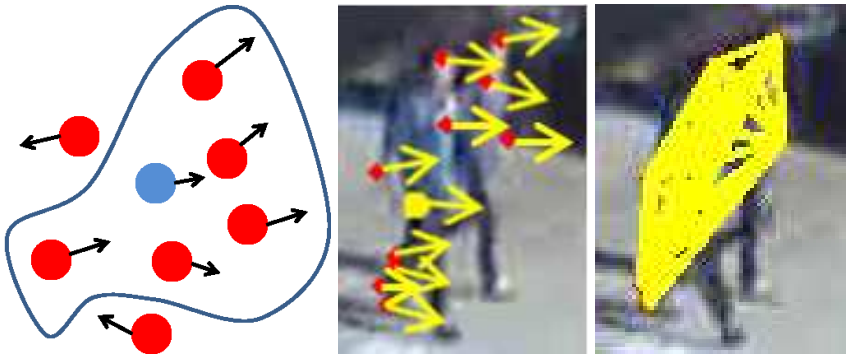
The output y is obtained by the propagation of input x through the hidden layers as in Eq. (2.4), where y^0 is an input vector.

$$y^0 \xrightarrow{W^1, b^1} y^1 \xrightarrow{W^2, b^2} .... \xrightarrow{W^L, b^L} y^L \tag{2.4}$$

In MLP networks, there are $L+1$ layers of neurons, and L layers of weights. During the training stage, the weights W and biases b are updated so that the actual output y^L becomes closer to the desired output d. For this purpose, a cost function is defined as in Eq. (2.5).

$$E(W, b) = \frac{1}{2} \sum_{i=1}^{n_l} (d_i - y_i^L)^2 \tag{2.5}$$

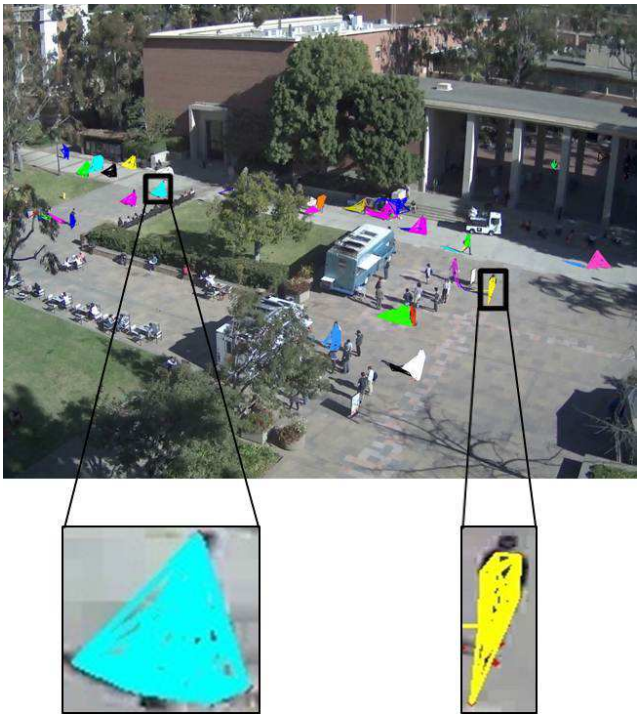
The cost function measures the squared error between the desired and actual output vectors. The backpropagation is gradient descent on the cost function in Eq. (2.5).



(a) (b) (c)

Figure 4: Entities grouping. Synthetic example of moving entities (a), moving entities obtained from the particles mutual influence model (b) and grouping implemented according to the motion and density features (c).

(a)



(b)

Figure 5: Input frame (a), entities grouping with the zoom on two sample groups (b).

Therefore, during the training stage, weights and biases are updated according to Eq. (2.6) and Eq. (2.7).

$$\Delta W_{ij}^{l} = -\eta \frac{\partial E}{\partial W_{ij}^{l}} \tag{2.6}$$

$$\Delta b_{ij}^{l} = -\eta \frac{\partial E}{\partial b_{ij}^{l}} \tag{2.7}$$

The learned weights and biases are used to predict groups from the inputs during testing stage.

In Fig. 5, the tracked groups of entities are shown. Two groups, annotated in cyan (left) and yellow (right) respectively, are zoomed and shown in the third row of Fig. 5. Primarily, entities are snipped with the particles mutual influence model and propagated over a predefined temporal window to associate them in groups in consonance with the features. At the same time, these groups are then mapped to a new set of snipped entities, with mutual influence model, which are then tracked over the same temporal window and the re-association process is repeated over time.

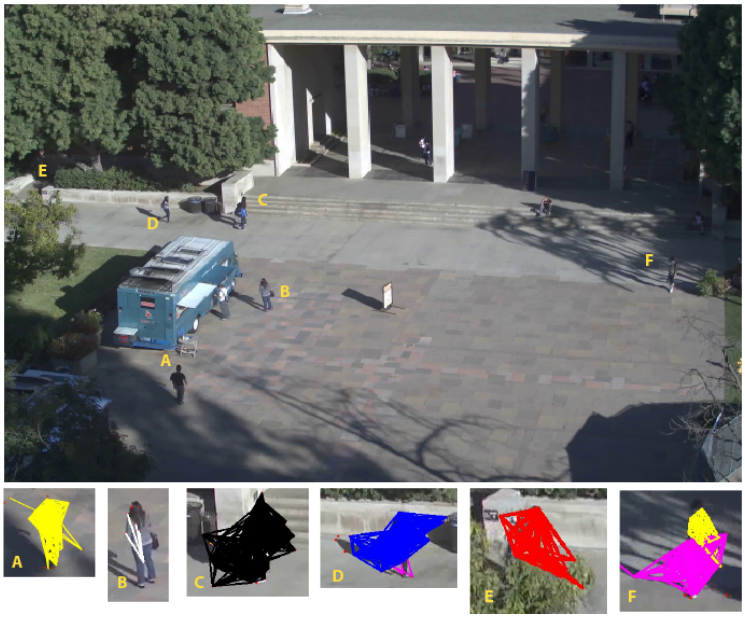
2.1.3 *Results*

For the experiments, we have considered two datasets: the UCLA Courtyard dataset [5] and the BIWI Walking Pedestrians dataset [102]. The UCLA dataset consists of two distinct scenes from a wide top/side viewpoint of a courtyard at the UCLA campus. The dataset comprises of a 106-minute video, 30 fps, and 2560 x 1920 resolution. The dataset presents human activities including walking, talking, riding-skateboard, riding-bike and driving car. The BIWI dataset includes 2 videos at the resolution of 640 x 480, 25 fps. Here we have considered only the ETH sequence because of the exclusive presence of pedestrians.

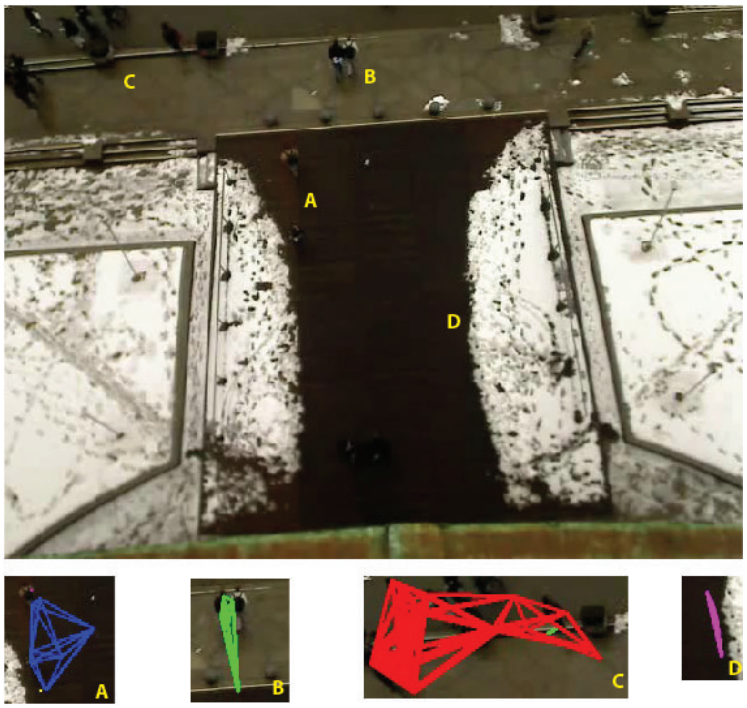
For the influence model, we have configured the following parameters: the length of the time window (time lapse between two observations in the influence model) has been set to 45 frames and the standard deviation of the Gaussian in Eq. (2.1) has been set to 0.8. These parameters are kept constant for both sequences to demonstrate the capability of generalization.

The neural network has been configured considering one input layer, two hidden layers and one output layer. The input layer consists of two neurons, each hidden layer consists of three neurons, and a single neuron is allocated to the output layer. The configuration of neural network in terms of number of layers and number of neuron does not affect the performance significantly. To extract the input features, the relative orientation with a reference particle is set to ± 30 degrees. Furthermore, the distance threshold from the reference particle is set to 80 pixels. For the purpose of training, we exploited 1000 training samples, where each sample is a vector of two observations namely; average distance and density of particles.

The obtained results are displayed in Fig. 2.1.3. The first image shows an example of the method applied on the UCLA dataset, where we can notice a very clear group composition, especially for zones A, D and E. In zone B the number of particles is



(a)



(b)

Figure 6: Particle influence and entity grouping. Results obtained on the UCLA dataset (a) and on the BIWI dataset (b). For visibility, labels are super-imposed on the original frame and the corresponding grouped entities are zoomed in lower row.

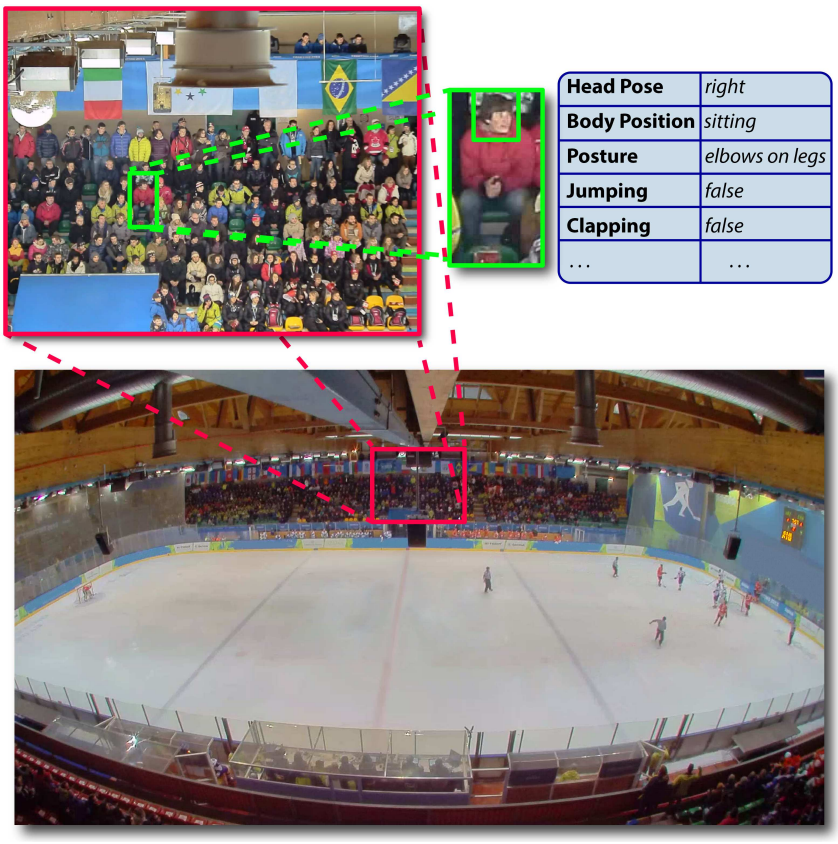


Figure 7: Example of images collected for both the spectators and the rink, plus the annotations.

not dense as in the previous cases, but still grouping is possible since the distance and density features of the entities are sufficient for the neural network. In zone F, instead, two groups have been detected instead of a single one; this can be ascribed to the severe shadowing, in which the pedestrian is located where, in fact, features of the entities are mainly segmented into two groups.

The quality of the results is also confirmed by the ETH sequence, where the groups are well defined (zones A, B, and C). However, in this case a few mistakes are also present (zone D). This is most probably connected to the limited resolution and the compression artifacts.

The UCLA dataset have much better results in terms of grouping not only because of the resolution but also because the bird eye view is less accentuated and the illumination conditions are considerably better.

2.2 SPECTATOR CROWD ANALYSIS

Considering the taxonomy of the crowd proposed in the introductive part of this chapter, we can certainly say that all the approaches in the computer vision literature focus primarily on casual [25, 71, 109], and protest crowds [72], 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 Office[1], disorders at stadiums caused 561 arrests and 2,273 banning orders 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 world[2].

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. The project *Oz* [3], founded by Provincia Autonoma di Trento 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 Canazei (Italy) on December 2013, focusing on the final 4 matches of the tournament. SHOCK 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

1 Football-related arrests and football banning order statistics, Season 2013–14, available on-line at `http://goo.gl/IkIzMV`.

2 See at `http://goo.gl/xWWFyf`.

3 The website with further info about the project are here: `http://disi.unitn.it/rota`

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

The rest of the chapter is organized as follows: The details of the data collection and labeling are reported in Sec. 2.2.1; the tasks of people detection, head pose estimation, and fan identification are introduced in Sec. 2.2.2, focusing on contextualizing the problem, discussing the related state of the art, presenting the considered baselines and our approach, and discussing the results obtained.

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	far-left / left / frontal / right / far-right / away
Body position	sitting / standing / (locomotion)
Posture	crossed arms / arms alongide body / elbows on legs / hands on hips / hands in pocket / crossed legs / etc.
Locomotion	walking / jumping (each jump) / rising pelvis slightly up
Action	waving arms / pointing / clapping (each clap) / hugging somebody / kissing somebody / opening arms / etc.
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 project's website.

2.2.1 *Data Collection & Annotation*

The UNIVERSIADE TRENTINO 2013 was held in Italy in the 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 of the ice hockey turnament (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 brougth 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 chapter):

- 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?

From these data, we used 3 matches as source of training and validation data, while the fourth match (the very final) for providing testing data: this has been for

stressing the generalization capabilities of all the algorithms, since in each different match we had 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 15 sequences.

Each sequence has been annotated frame by frame, spectator by spectator, by a first annotator, using the ViPER format [38] and the toolkit proposed in [1]. Such 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, partially listed in Table 1[4].

Among the whole set of possible features that can characterize the human dynamics, we selected the annotated *elementary forms of action* [83] 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, gesture, etc.), especially in public places; and (b) the so-called *crowd behavior*, i.e. social interaction in large gatherings [49, 50, 83], in particular sport spectator gatherings.

2.2.2 *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.

2.2.2.1 *People Detection*

People detection is a standard and still open research topic in computer vision, with the HOG features [35] and the Deformable Part Models [45] as workhorses, and plenty of alternative algorithms [40]. 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 –

4 More information about the values of the labels will be given in the supplementary material

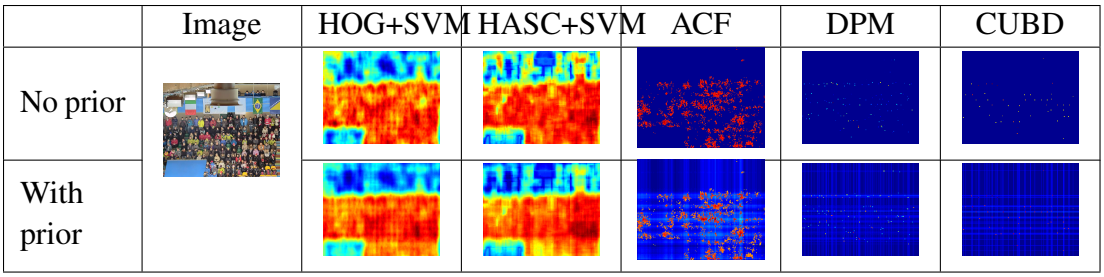


Figure 8: Qualitative results for people detection algorithms. Detection confidence map for each method with and without imposing the grid-arrangement prior. (best viewed in color)

	Simple Search			Grid Prior		
Method	Prec.	Rec.	F_1	Prec.	Rec.	F_1
HOG + SVM	0.743	0.561	0.639	0.662	0.709	0.684
HASC+SVM [118]	0.365	0.642	0.465	0.357	0.685	0.469
ACF [39]	0.491	0.622	0.548	0.524	0.649	0.580
DPM [45]	0.502	0.429	0.463	0.423	0.618	0.502
CUBD [41]	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.

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. [8] used the Hough transform to overcome the non-maxima suppression stage for detecting multiple istances of the same object, while San Biagio et al. [118] proposed a new descriptor able to treat complex structural information in a compact way. To overcome occlusion issues, Wu and Nevatia [143] used a number of weak part detectors based on edgelet features and Eichner et al. [41] fused DMP [45] and Viola-Jones [97] detectors to identify upper bodies. Finally, Rodriguez et al. [112] 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 [35] 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 [118]. 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 [39] 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) [45] combines part's templates arranged in a deformable configuration fed into a latent SVM classifier; and (3) the Calvin Upper Body Detector (CUBD) [41], 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 transform 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 72×112px 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 overlap between the predicted and annotated bounding boxes is more than 50% of the union of the two rectangles. 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. 8, 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.

2.2.2.2 *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 [28, 68, 146]. 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. [96] 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 [133], 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).

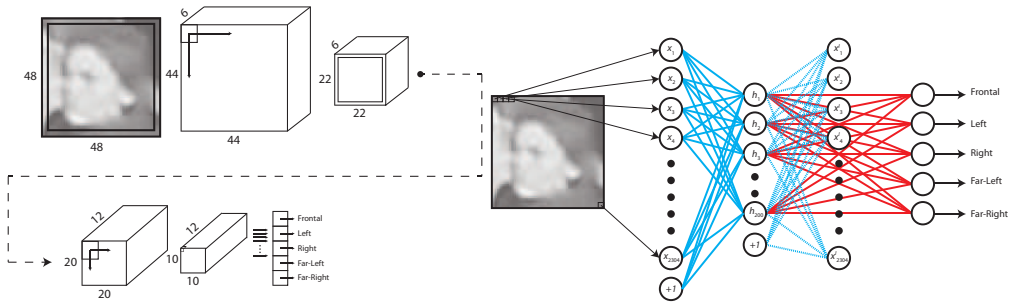
In order to overcome this issue, we propose two novel approaches based on Deep Learning, with comparable results but much faster. The choice of Deep Learning is motivated by the large number of effective approaches in the object recognition literature, witnessing its versatility in many scenarios [63, 73, 77, 128, 129].

In particular, we evaluate the performance of the Convolutional Neural Network (CNN) and the Stacked Auto-encoder Neural Network (SAE) architecture. In both methods 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. 10

(a) (b) (c) (d) (e)

Figure 9: Examples of the five head poses that are present in the dataset, in order (a) to (e): *far left*, *left*, *frontal*, *right*, *far right*.

(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. 10 (b), the input images are fed into an auto-encoder with hidden layers of size $\mathfrak{h} = 200$, 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.



(a) (b)

Figure 10: (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.

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 perfomances. 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

Method	AVG Accuracy	Training time [sec]	Testing time [sec]
Orozco et al. [96]	0.368	105303	6263
WArCo [133]	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.

as *far left* and *far right*. This has been detailed to the annotators during the data labeling (see Fig. 9).

The 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. [96], WArCo [133], 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. (2.8)

$$P_A^{(c)} = \sum_{i=L^{(c)}}^{U^{(c)}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - m^{(c)})}{2\sigma}} \tag{2.8}$$

Method	Accuracy
AS2007 [3]	0.592
MMS2010. [84]	0.559
Our	0.621

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

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.

$$c = \arg\max_{c} \left(\alpha P_A^{(c)} + (1-\alpha) P_N^{(c)}\right) \tag{2.9}$$

The final decision is taken according with the Eq. (2.9), 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.

2.2.2.3 *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.

Spectators categorization can be considered a subtask of the crowd behavioral analysis, which is generally associated with human activity analysis [2, 52, 105]. As stated by Jaques et al. [61], 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 usully start from optical flow to extract global information from the crowd. In [3], 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.

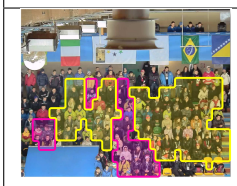
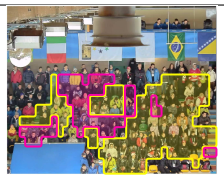
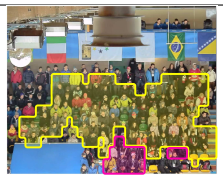
GT	AS2007 [3]	MMS2010. [84]	**Our**
			

Figure 11: Qualitative results for spectators categorization. The colored areas represent the two groups of spectators supporting different team.

More recently, Mehran et al. [84] 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 [3, 113], 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 work we also present a new method for spectators categorization, whose framework can be extended to previous methods. As most of the holistic approaches, our method also starts from optical flow computation. Then we decompose the flow map into a set of overlapping patches and we describe each patch with a feature vector of five elements: 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 [46], obtaining an instantaneous grouping 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^f_τ where each entry $H^f_\tau(i,j)$ measures the similarity between patches p_i and p_j considering the history of the patches until frame f of the video. $H^f_\tau(i,j)$ can be computed as:

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

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. The final PSH represents the history of the similarity between the patches. Thus we can perform a complete linkage hierarchical clustering on the reciprocal of PSH to obtain the spectators categorization.

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 Np=585 patches with size 64×128px 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 crowd extraction and spectators categorization tasks. Instead Figure 11 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).

2.3 DISCUSSION

In this chapter we have analyzed the social interaction activities from a broad point of view, mostly focusing on crowd.

In section 2.1, we have proposed a method to detect and track moving entities using a particle-based approach. According to the mutual influence model, particles are analyzed on the basis of their dynamic properties. For this purpose, we have extracted the average distance and density features for the particles sharing similar orientation. The obtained features are then exploited to train a multi-layer perceptron neural network, using the back propagation algorithm. Experimental results on two

benchmark datasets, UCLA and BIWI, have demonstrated that our method can be efficiently used to detect and track moving entities present in the scene at a low computational burden, thus saving precious resources for any further processing step, compared to more traditional approaches based on detectors.

The second part of the chapter (section 2.2) has introduced S-HOCK, a novel dataset focusing on a brand-new issue, the modeling of the spectator crowd. The 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. Therefore, we are confident that S-HOCK may trigger the design of novel and effective approaches for the analysis of the human behavior.

In the next chapter we are getting closer to the subjects, analyzing their behavior with the help of some basic sociological observations, introducing proxemics as an important feature to discriminate different types of interactions.

3

MID-RANGE APPROACH: PROXEMIC ANALYSIS

In this chapter we analyse the importance of the evolution of the distance among subjects in relation to their social interaction. We introduce the concept of abnormal social interaction that differs from normal interaction on his violent attitude. We propose a framework able to detect such type of dyadic behavior focusing our attention mainly on the global movement of the interacting subjects. In the second part of the chapter we propose a collaborative creativity applicative framework based on proxemic analysis where dyadic interactions are mapped in a composition algorithm that is able to translate in music the movements and the attitudes of the participating subjects.

The research in video surveillance and environmental monitoring has revealed a recent trend in bringing the analysis of the scene to a higher level, shifting the attention from traditional topics, such as tracking and trajectory analysis towards the semantic interpretation of the events occurring in the scene. In particular, behavior analysis in terms of action and activity recognition has emerged as a relevant subject of research, especially for classification and anomaly detection purposes. Important contributions to the field have been proposed by Scovanner et al. [123], in which authors learn pedestrian parameters from video data to improve detection and

Parts of this Chapter appear in:

- Rota P., Conci N., Sebe N., **Real time detection of social interactions in surveillance video.**ECCV 2012 Workshops and demonstrations.
- Morreale F., Masu R., De Angeli A., Rota, P., **The music room**. ACM-CHI'13 Extended Abstracts on Human Factors in Computing Systems.
- Morreale F., De Angeli A., Masu R., Rota P., Conci N. **Collaborative creativity: The Music Room**. Personal and Ubiquitous Computing, 18(5), 1187-1199, (2014).
- Rota P., Dang-Nguyen D. T., Conci N., Sebe, N. **Exploiting visual search theory to infer social interactions**. In SPIE Electronic Imaging.(2013)

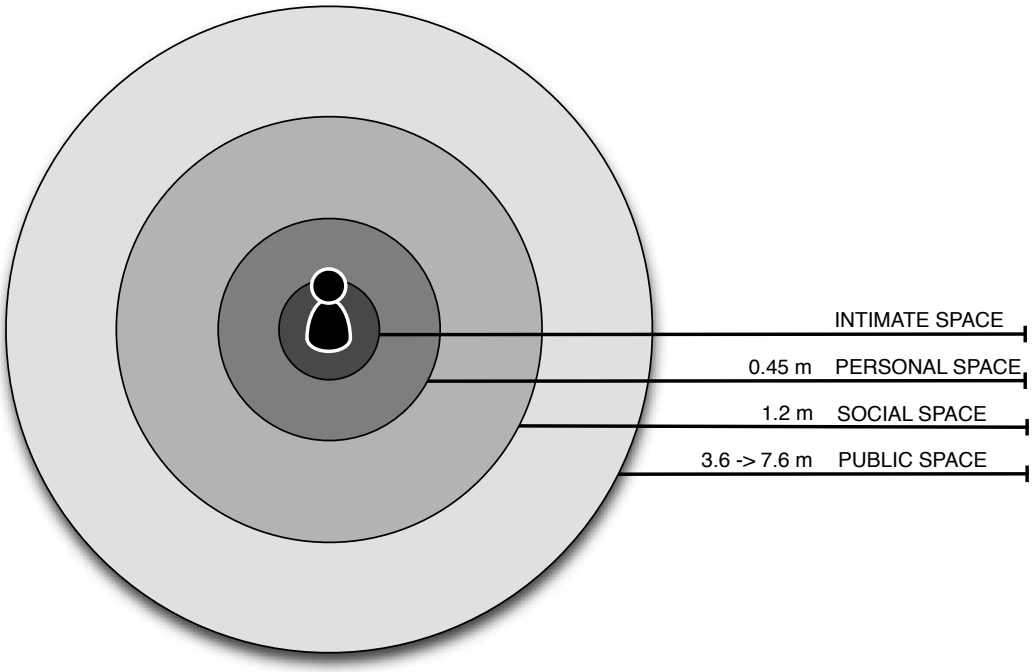


Figure 12: Definition of the proxemics space as defined by Hall in [55].

tracking; Robertson et al. [111] model human behaviors as a stochastic sequence of actions described by trajectory information and local motion descriptors. Bringing the analysis to a higher level of interpretation involves understanding the social relationships undergoing between subjects, thus requiring an extension of the analysis domain also including psychology and sociology. To this aim, the proxemics theory can be effectively exploited to observe the human relationships when captured by a surveillance camera. According to Hall's theory [54, 55], each person speaks a silent language, related to the behavior of the subject, and expressed in terms of motion and body pose. The main proposition of his studies consists of the correlation between the distance among people and the corresponding relationship ongoing between them. An example of how proxemic space is modeled is in Fig. 12.

In section 3.1 we propose a model exclusively based on proxemic features aiming to the detection of the nature of the social interaction between a dyad of subjects. In the following section we are proposing a creative application with the objective of extracting emotional features from a pair of subjects in order to generate a brand new melody composed according with their proxemic parameters.

3.1 PROXEMICS AS IMPORTANT HIGH-LEVEL FEATURE

According with what introduced at the beginning of this chapter, proxemic has a paramount importance in outlining the social interactive situation in a certain monitored area. Following similar principles, Cristani et al. [32] aim at understanding the social relations among subjects when sharing a common space. The authors detect the so-called F-Formations, in order to infer the presence of an ongoing interaction between two or more persons. An approach based on proxemics is proposed by Zen et al. [144] The authors identify proxemics cues in order to discriminate personality traits as neuroticism and extraversion, and use the collected data to construct the corresponding behavioral model. The acquired data is then used to improve the accuracy of the tracking algorithm. A similar approach has been proposed by Pellegrini et al., [102] using the social force model. [85] The solution proposed in [102] considers each subject as an agent, for which themodel of motion has to be optimized, so as to prevent collisions with the other entities moving in the scene. The authors consider every agent as driven by its destination, taking into account, besides position, also additional parameters like velocity and direction of motion. The collected data is then used to model the proximity level between subjects, in order to construct an avoidance function. Cui et al. [34] extract an interaction energy potential to model the relationships ongoing among groups of people. The relationship between the current state of the subject and the corresponding reaction is then used to model normal and abnormal behaviors. The authors also claim that their approach is independent from the adopted tool for human motion segmentation. A hierarchical approach is proposed by Lan et al. [74] where human behavior is described at multiple levels of detail ranging from macro events to low-level actions. Authors exploit the fact that social roles and actions are interdependent one to each other and related to the macro event that is taking place.

In this work we define the interaction as a combination of energy functions that capture the state of a subject in the social context he moves. Since tracking is out of the scope of this work, our goal is to build a classifier to identify and recognize different types of behaviors. A novel aspect we introduce with respect to [102], consists in the insertion of an *intentionality* parameter in the processing chain, targeted at distinguishing between intentional and casual interactions. This term, provided by the proxemics information, is used to weight the interaction patterns acquired in real-time on a sliding window basis. The output of the function is then brought into the Fourier domain, thus removing the temporal correlation of the samples, and eventually fed into an SVM classifier. We have devised three different scenarios: (i) casual interaction, (ii) normal, and (iii) abnormal interaction. The interactions of type (i) refer to non-intentional events, while the type (ii) and (iii) reveal intentional interactions, divided into regular and potentially dangerous events.

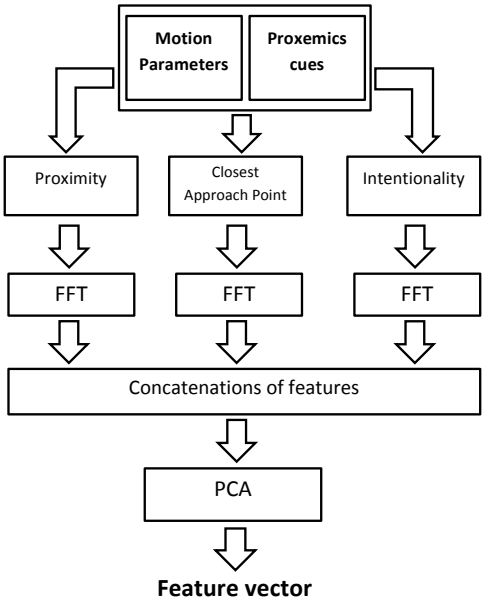


Figure 13: Flowchart of the proposed architecture.

The method has been tested on three datasets specifically chosen for human interaction analysis.

3.1.1 *Methodology*

According to the proxemics principles, distances can say a lot about the relationships going on between people, about their intimacy level, making it possible to distinguish between intentional and non-intentional behavioral cues. This information is generally variable in space in time and depends on the location in which a person stands, on the density of people in the area, but also on cultural and religious differences.

Fig. 13 shows the proposed architecture for social interaction analysis, for which we will provide additional details in the next subsections.

3.1.2 *Proxemics parameters*

In the model we propose, we follow the path covered by Pellegrini et al. [102] in order to capture the salient motion features that can be associated to an interaction. Each subject i is modeled at each time t by a state vector of parameters that takes into account the current position and velocity:

$$S_i(t) = [\mathbf{p}_i(t), \mathbf{v}_i(t)] \tag{3.1}$$

At each time instant t it is then possible to model the distance between each pair of subjects (i, j) as:

$$d_{ij}^2 = \left\| \mathbf{p}_i + t\mathbf{v}_i - \mathbf{p}_j - t\mathbf{v}_j \right\|^2 \tag{3.2}$$

By defining $\mathbf{k}_{ij}^t = \mathbf{p}_i^t - \mathbf{p}_j^t$ and $\mathbf{q}_{ij}^t = \mathbf{v}_i^t - \mathbf{v}_j^t$ and applying the derivative with respect to t in Eq. (3.2), it is possible to find the time instant t^* at which the distance d_{ij}^* between the subjects is minimized.

$$t^* = -\frac{\mathbf{k} \cdot \mathbf{q}}{\|\mathbf{q}\|^2}, \quad d_{ij}^{*2} = \left\| \mathbf{k} - \frac{\mathbf{k} \cdot \mathbf{q}}{\|\mathbf{q}\|^2} \mathbf{q}^\top \right\|^2 \tag{3.3}$$

Eq. (3.3) is the estimate for the closest point (and the corresponding time instant), at which the subjects will most probably meet. However, this piece of information, although relevant to check whether there is chance for i and j to interact in the next future, does not necessarily include details about their interaction level. An estimate can be obtained by building an energy functional between subjects i and j by measuring the evolution of the proximity between them over time:

$$E_{ij}^c = e^{-\frac{d_{ij}^{*2}}{2\sigma_d^2}} \tag{3.4}$$

In Eq. (3.4) σ_d controls the variance of the function in order to make it more or less responsive. The output of Eq. (3.4) can be seen as a *collision warning*, and represents the closest distance at which the two subjects will be, given the current motion parameters (position, velocity and direction of motion). This element is important because it can be used as a hint to predict the future developments of the interaction.

In line with the previous statement, we define an energy function to model the actual distance between subjects. This parameter is useful to obtain a proper modeling of the social behavior, since an interaction is more likely to happen when two persons are closer rather than when they are far apart from each other.

$$E_{ij}^d = e^{-\frac{\left\| \mathbf{k}_{ij}^w \right\|^2}{2\sigma_w^2}} \tag{3.5}$$

In [102], and for tracking purposes, the authors use the term E_{ij}^d as a weight to model the outcome of Eq. (3.4) together with another term depending on the angle between the direction of motion of i and the position of j. Our goal is however different, since we want to understand the dynamics of the interaction. Furthermore, the direction information, is in general noisy, particularly in the case of unrestricted video scenes, and for these reasons it has been discarded from out model.

In order to model the intentionality of an interaction, we adopt the so-called *O-space* [33]. The *O-space* consists of a circular area between the subjects, located in the direction of their gaze. It can be seen as the interaction space, namely the area comprised between two people interacting and facing one to each other.

By means of this definition, the *O-Space* can be used as a selectivity criterion, i.e. to inform about the presence of an interaction. The *O-Space* is in general defined as a static and non-deformable area right in front of the person and is not suitable for dynamic motion models, in which interactions can occur also in case the subjects move (e.g. walking together). Therefore, in our proposal we borrow the idea of the *O-space* as an area of attention of the subject, which can be adopted to infer the intentionality (or causality) of an interaction. In our model the *O-space* is positioned along the direction of motion of the subject and its center varies depending on his velocity. This gives us the opportunity of handling also dynamic interactions, and not only static events.

The position of the *O-space* is defined as:

$$\begin{aligned} O_x &= p_x + a_x \Lambda \sin(\theta) \\ O_y &= p_y - a_y \Lambda \cos(\theta) \end{aligned} \tag{3.6}$$

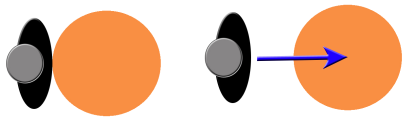
where p_x and p_y are the coordinates of the subject, Λ is the displacement of the subject from the previous frame, a_x and a_y are tuning parameters depending on the field of view of the camera, and θ is the absolute direction of motion. The *O-space* area is used to calculate the intentionality component of the interaction, similarly to what we did for the proxemics information:

$$E^o_{ij} = e^{-\frac{\|\mathbf{k}^o_{ij}\|^2}{2\sigma_o^2}} \tag{3.7}$$

where k^O_{ij} is the distance between the *O-space* centers of subject i and j , respectively. This parameter allows to filter out the noisy information collected by the other terms (for example two people very close but facing in opposite directions), thus reducing the chances of false positives returned in the presence of casual interactions of subjects standing nearby. The *O-space* model we have adopted is shown in Fig. 14.

3.1.3 *Feature Ectraction*

Following the flow chart in Fig. 13 we collect the proxemics values $E^d_{ij}(t), E^c_{ij}(t), E^o_{ij}(t)$ in a given temporal window (128 samples in our examples), and at each time instant we apply the FFT (Fast Fourier Transform) (3.8) on the window samples. At this stage, the importance of the FFT is to eliminate the temporal correlation of the



(a) (b)

Figure 14: O-space modeling. The figure represents the two cases in which the subject is (a) standing still, and (b) when he is moving from left to right. In the latter case the O-space shifts in the direction of motion proportionally with its velocity.

samples by only considering the contribution they bring into the interaction in terms of dynamics of that specific event.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\imath 2\pi k \frac{n}{N}} \qquad k = 0, ..., N-1 \tag{3.8}$$

The next step consists of concatenating the three sets of features to construct the feature vector that will be analyzed by the classifier. This process is carried out at every time instant, resulting in a large number of parameters (128x3). Therefore, we apply a dimensionality reduction through Principal Component Analysis (PCA). Accordingly, the training set is arranged in a $n \times m$ matrix where n is the number of samples and m the number of features. From the matrix X the eigenvalues of the related zero mean covariance matrix are extracted and the obtained vector is sorted by magnitude in descending order. The first value is the so-called principal component. From eigenvalues vector we can compute eigenvectors $m \times m$ matrix.

$$Y = W_s^T X \tag{3.9}$$

As shown in (3.9) the feature space has now been reduced, restraining the training set to a new matrix of size $n \times s$ where $s < m$ is the number of eigenvectors that we consider as relevant for our analysis.

Now that we have constructed our training set, we adopt a similar procedure for prediction. Each new incoming sample consists of a $1 \times m$ vector that is processed as X in (3.9) obtaining as output a $1 \times s$ vector. This new vector is the input for the SVM, from which we will classify the type of the ongoing interaction.

3.1.4 *Classification procedure*

After obtaining the reduced feature space, classification is computed using a kernel based SVM. Since the classification output strongly depends on the data used for

training, let us briefly see what are the main steps we follow to obtain a reliable training set:

- Select the training videos representing the three classes that we want to classify with the frame-by-frame interaction labeling (manually done in a previous stage);
- Compute the interaction values as presented in Section 3.1.2 for the whole duration of the video;
- Segment the interaction values in accordance with the labels;
- Run the sliding window over the segmented interaction values, and consider each step as a feature vector;
- Transform each feature vector in the FFT domain and reduce the dimensionality using Principal Component Analysis;
- The resulting arrays consist of the features space for the classifier, which will be tuned by cross validation optimization to estimate the best configuration for the class separation.

It is worth noting that samples for training are picked randomly and in equal number for each class from the dataset, in order to avoid any possible bias in the training and to prevent overfitting of a particular class with respect to the others.

In the test phase the procedure simply consists of collecting the sliding window data at each time step, compute the FFT transform and the PCA decomposition using the training eigenvectors, thus building the new space. Data are then sent to the classifier for the final class prediction.

3.1.5 *Results*

3.1.5.1 *Datasets*

To validate our method we have used three different datasets: our own dataset[1] (called from now on: Social Interactions - SI Dataset), a selection of video sequences collection of YouTube CCTV videos (different contexts) and some sequences taken from the Behave dataset.

The SI Dataset has been acquired to specifically address the topic of interactions analysis. Therefore, we provide a brief explanation of its content. The set consists of 12 fully annotated video sequences of different length recorded at 25 FPS.

1 This dataset has been collected by the authors for surveillance purposes and it is available at https://dl.dropbox.com/u/4098070/ECCVW.zip

Sequences mainly represent regular daily life behaviors such as people chatting, walking together or simply crossing each other. The dataset also includes more critical types of interactions, simulating fights. The video sequences are recorded outdoor, under three different views, for which we will use here only the bird's eye view for similarity with the other datasets. For our experiments, and considering that tracking is out of the scope of this paper, we use the collected ground truth, from which it is possible to extrapolate all the necessary parameters required by our method.

The YouTube dataset is composed by 4 video sequences recorded in as many different locations. This dataset is not homogeneous because the videos come from different sources, with different view angles and different fields of view. For there reasons the videos are very challenging, since they represent real-life situations, and are not acquired with any specific purpose.

From the Behave dataset [9] we have included in the experiments two different segments regarding different behavioral situations. Also here videos are acquired from far range, and are only partially annotated. We have then collected the corresponding ground truth.

3.1.5.2 *Experiments*

As mentioned in Section 3.1.1, classification is achieved via a multiclass SVM with Gaussian kernel. The number of training samples for each dataset is 1200, balanced over the three classes (400x3). In the training phase the best SVM parameters have been estimated by cross-validation.

The testing phase takes as input the SVM parameters and the interaction parameters used to compute the interaction measure. These parameters are estimated through an exhaustive search and they differ in relation with the properties of the monitored area (range, field of view, angle). The proposed architecture allows computing the interaction measure on-line, without waiting for the end of the interaction. In Fig. 3.1.5.2 the energy functions obtained from three different sample sequences are shown.

In terms of numerical results we present two different tables, where it is possible to observe the effectiveness of our approach, especially in unconstrained scenarios, in which the interpretation of the interactions could be problematic. As it can be noticed from Table 5 and Table 9, the algorithm performs in general well especially in detecting the presence of an interaction, in all three datasets used for testing. As far as the class 3 is concerned (anomalous events) and considering the complexity of the task, the improvement given by the O-Space term is considerable (more than 20% in precision) due to the capability of better isolating the interacting subjects. A graphical presentation of the classification process is shown in Fig. 16. Here, each line reports three snapshots taken from the different datasets, each of them

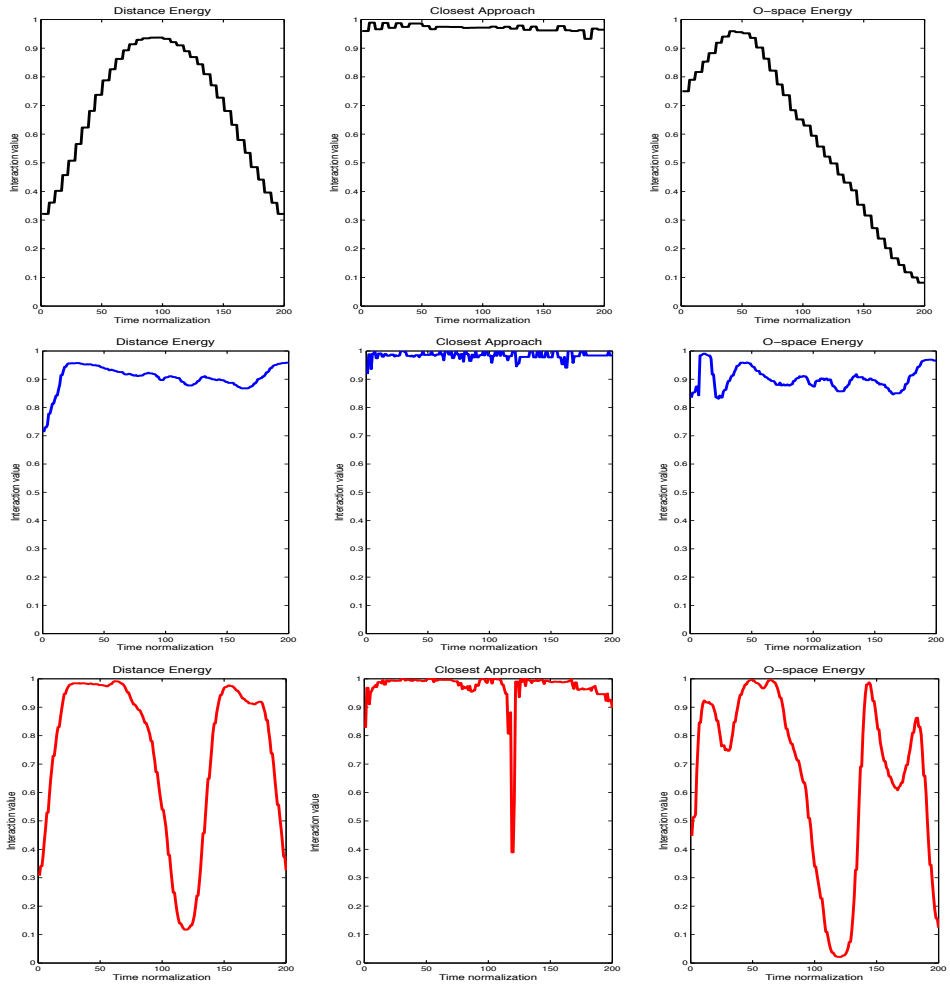


Figure 15: Energy functions for distance (left), closest point of approach (center), and O-Space (right), in presence of two people crossing (first row), chatting (second row), and fighting (third row).

representing one of the classes. White lines (left column) indicate that no interaction is currently ongoing, yellow lines (center column) refer to normal interactions, while red lines (right column) indicate the presence of an abnormal event.

3.2 MUSIC ROOM

The fostering of creativity has recently become a key educational, social and economic priority in many parts of the world [121]. There is a growing consensus that creativity, arts and wealth are deeply interrelated and that creativity stems from a collaborative context, rather than from the mind of an isolated individual.

Figure 16: Sample interactions taken from the three datasets. The first column indicates casual interactions, the central column refers to normal interactions, while the last column signals the presence of abnormal interactions.

Contemporary research aims at a psychological and sociological understanding of the contexts which favour creativity and of the analysis of the meditational role of cultural artifacts, such as signs, tools and technologies.

This applicative work describes our experience with The Music Room, an exhibition designed to allow people with little or no musical education to compose original music while moving in an interactive space. In this exhibition, the space becomes the instrument and the stage of an act of creativity that has so far been confined to a small elite of educated musicians. Indeed, due to the complexity of playing and composing, most people can only experience music by dancing or listening to compositions created by somebody else. Only professional musicians can master the intrinsic issues of creating quality music. New technologies, such as ubiquitous computers, touchscreen devices and tracking systems have been designed to deal with the issues related to the creation of original music [11,21,64] by replacing traditional instruments with more intuitive devices. These technological advancements provide novel possibilities by leading a new set of design issues: finding new metaphors to ease the art of creating music and thus opening this creative area to a wider and

Table 5: Performance comparison of the proposed algorithm with and without the O-Space energy on the three datasets.

		O-space Method			Without O-space Method		
		Precision	Recall	HitRate	Precision	Recall	HitRate
SI	Casual	93,3%	94,1%	88,5%	93,5%	91,1%	86,1%
	Normal	75,1%	76,1%		63,3%	77,5%	
	Abnormal	55,3%	48,7%		55,4%	45,3%	
Behave	Casual	75,8%	93,8%	90,1%	75,3%	93,8%	90,1%
	Normal	98,0%	93,9%		97,7%	94,3%	
	Abnormal	42,5%	27,5%		43,6%	22,6%	
YouTube	Casual	88,2%	90,2%	82,7%	66,3%	76,2%	60,1%
	Normal	84,4%	80,2%		70,9%	43,7%	
	Abnormal	38,2%	40,3%		11,0%	17,7%	

Table 6: Confusion matrices for the three datasets obtained using the O-Space energy.

		Casual	Normal	Abnormal
SI	Casual	94,07%	3,68%	2,25%
	Normal	18,49%	76,17%	5,34%
	Abnormal	37,56%	15,34%	47,10%
Behave	Casual	93,82%	3,53%	2,65%
	Normal	3,98%	93,92%	2,10%
	Abnormal	61,83%	10,92%	27,25%
YouTube	Casual	90,16%	5,00%	4,85%
	Normal	12,74%	80,23%	7,03%
	Abnormal	29,71%	29,92%	40,37%

untrained audience. Current research is focusing on the development of methods to assist music composition by hiding part of the complexity of creation to the user [31]. Such methods partly fail to reach their goals because they rely on traditional sonic and musical metaphors (e.g. scales, sequencers and filters) that are unlikely to convey any meaning to non-musicians. To overcome this limitation, a few works in artificial music composition have been recently trying to search for interaction paradigms into different domains [20, 110]. This approach is consistent with the Blended Interaction conceptual framework suggesting that computer-supported collaboration should map elements through different domains by combining real world concepts with digital analogies [62].

With this purpose in mind, we aim at exploring new interaction metaphors, which are supposed to match a series of requirements: they have to be available to everybody, social, intuitive and naturally connected with music. Emotion seems to be the element that best meets these requirements. In each culture, music is one of the arts that can most effectively elicit emotions [7, 47]. Significantly, it has always been connoted as emotional [65]. Bodily movements, which, in the different declinations of dancing and conducting, are traditionally associated to music [19,65], are the most appropriate medium through which emotions can be conveyed. The Music Room combines these metaphors by providing a space where people can compose music expressing their emotions through movements. It is designed to be experienced by couples according to the metaphor of love as closeness between people. Two users direct an original piano music composition by providing information about the emotionality and the intensity of the music they wish to create. In order to communicate these factors, users interact with each other by moving in the space: the proximity between them influences the pleasantness of the music, while their speed influences the intensity. These intuitive compositional rules open endless possibilities of creative musical expression to everybody. The music is generated by Robin, an algorithmic composition system that generates original tonal music in piano style by following a rule-based approach. Proxemics information is acquired by a vision tracking system. It is then converted into emotional cues and finally passed to Robin.

3.2.1 *Related Works*

The design of The Music Room spans several research areas. The rules of the compositional system are inspired by research on the Psychology of Music, while Robin is founded on existing approaches for algorithmic composition. The idea of exploiting the metaphor of gestures and emotions is partially influenced by previous collaborative interactive systems for music expression. Computer vision solutions have been adopted and preferred to other technologies such as wearable sensors, in order to capture the motion of the participants and analyze their level of interaction, so that they could move without any constraint in their behaviors.

3.2.1.1 *Algorithmic music composition*

Two approaches have so far been adopted to address the challenge of algorithmic music composition: one is based on the editing of existing musical material, the other on algorithmic generation of new melodies. In the editing approach, some compositional parameters (e.g. tempo, note duration, mode, legato phrasing) are modified in a series of excerpts of music [93, 136]. For example, the work by

Oliveira et al. composed music by playing-back existing melodies where a number of parameters were dynamically changed according to emotion variations, described in terms of valence and arousal [93]. This is a simple and elegant solution, yet the limited number of musical parameters that can be used to edit existing tracks represents a major deficiency. The variation of pre-composed melodies is restricted, and outcomes are often repetitive. As a result, users are confronted with a limited range of scarcely original songs, which is likely to make their experience less appealing.

The generative approach has been widely explored in the last decade, often supported by a dimensional description of emotion in terms of valence and arousal [78, 137, 138]. Three basic methods have been proposed: rule-based, learning-based and evolutionary composition [132]. The first two methods differ in the procedure used to define the compositional rules [108, 127, 138]. In rule-based methods, knowledge of music theory is manually coded into the system. The diversity of possible outcomes depends on the amount of taught rules.

In learning-based methods, the system is trained with existing musical excerpts and rules are automatically added [17, 57, 125]. Most of these works make use of Markov chain models to represent efficiently musical patterns: the Continuator is a system that automatically generates music according to previously learnt, arbitrary styles [98]. Whilst the learning method has the clear advantage of decreasing the reliance on designer knowledge, the quality of music is heavily dependent on the training set.

Evolutionary algorithms allow the creation of original and unique music by means of mechanisms inspired by biological evolution. This approach ensures an original and unpredictable output [48, 87, 142] but the music might sound unnatural and structureless if compared to rule-based systems that are generally superior by virtue of the context-sensitive nature of tonal music [91].

3.2.1.2 *Collaborative musical systems*

Research on the design of collaborative systems for playing music has been growing in the last years [10]. Some of these systems target users that have at least a minimum musical training. In this category, we can identify tangible musical interfaces such as the Reactable [64] , Jam-O-Drum [11] , AudioPad [115] and GarageBand for the iPad. Despite the appealing interface and collaborative interaction of these systems, the profuse amount of sonic and musical inputs (e.g. envelopes, sequencers, source generators) tends to make their use difficult for non-musicians. Another category focuses on the concept of active listening, a novel generation of musical systems which are addressed to inexperienced users [115]. In these systems, users can interactively control the music content by modifying it in real-time, while they are actually listening to it [22].

Several works have been trying to enable groups of people to shape musical contents through collaborative interaction [20, 21]. In 2011, Taylor and her colleagues developed humanaquarium, an installation aimed at investigating audience engagement with creative experiences [131]. While two performers are playing music inside a glass cube, the audience can alter their performance touching the surface of the cube itself. This project pursues the goal of a collaborative control of the audio-visual content, enabling participants to 'jam with the performers'. In the Urban Musical Game, installation visitors interact with augmented foam balls to manipulate the pre-composed music material [110]. TouchMeDare aims to encourage strangers to collaborate. In fact, two or more people can compose music by interacting through a canvas: the pre-composed music samples, though, are only triggered when the canvas is simultaneously touched by both users [135].

Sync'n'Move allows users to experience music in a social interaction context [136]. Two users freely move their mobiles and the complexity of music orchestration is directly proportional to the synchronization of their own movements. Accordingly, if synchronization fails, there will be no orchestration at all; if synchronization is only partially achieved, orchestration will be quite elementary; in the case of perfect synchronization, the level of orchestration will reach its peak. In Mappe per Affetti Erranti [20], a group of people can experience active listening by exploring pre-composed music and navigating a physical and emotional space. The installation encourages collaboration, as music pieces can only be listened to in their full complexity if the participants cooperate with each other by moving through the space.

Other studies endeavor to exploit expressiveness and emotions to influence the status of the system [20, 21], through body gestures [81] and dance movements [19]. Video analysis techniques have often being exploited in performing art. They range from simple position tracking to complex full-body and hand movements tracking by means of sophisticated architectures. EyesWeb, for instance, is a platform for the design and development of real-time multimodal systems and interfaces, and it was specifically designed to track the gestures of performers [23, 81].

3.2.1.3 *Social interactions via proxemic cues*

The proxemic cues have been exploited before in order to develop ubiquitous systems. One of these applications has been proposed by Eriksson et al. [42]. In their work the authors have proposed a conceptual framework, which emerged from the observation of movement-based interactions; they explain the concepts of space, relations and feedback using four different applications. Motion also represents important information for the acquisition of new visual features, allowing the distinction between ongoing and completed activities [79]. The work of Greenberg et al. provides an interesting example of exploitation of proxemic rules in an interactive

scenario [53]. An interactive screen displays different information to passers-by, mirroring different interaction styles which are influenced by the distance between the user and the installation and the attention of the user.

3.2.2 *Music Room*

New technologies and the increase in computational power are changing live performances with sensors-augmented instruments and widening musical expression with interactive devices such as tabletops, tablet applications and tangible and haptic devices [64, 95]. Technological advancement, however, needs to develop new design concepts in order to simplify the access to playing and composition, if it is to make the experience of musical creativity available to everybody. The design of interactive systems for non-musicians introduces a new conceptual issue. The language in which one communicates their artistic intentions cannot rely on musical paradigms that are unknown to most of the people. The avoidance of musical input in the process of music creation has two main effects on the design process.

First, in traditional musical instruments (e.g. piano, brasses, synths, percussions) a mechanical input is usually directly associated with a musical output. This does not hold for interactive musical systems, where the input can be separated from the music generation system, as there is no need to force the user to direct the music note by note. An intermediate layer gathers input information, which is then fed into a lower-level module represented by the actual music engine. Second, the purposes of music creation in contexts of collaborative interactive spaces should be reconsidered, since users do not expect to create masterpiece music. The design needs to focus on the experience of users and on the social aspect of interaction and not only on the musical outcome itself.

The search for new input metaphors should be consistent with these changes of purpose. The language of emotions and body gestures can be a proper bridge between users' intention and the musical engine. This pervasive language allows humans to convey rich information which is clearly linked with the musical domain. The Music Room is based on the fusion of movements and emotions. It is designed to be experienced by a couple, according to the paradigm of love: two users direct the musical output by supplying the system with information about the emotionality and the intensity of the music they wish to create. To communicate these factors they interact with each other by moving through the space. Their behavior expressed by proxemic cues in terms of relative distance and speed is analyzed to map the pleasantness of music and its intensity respectively. This tool enables users to create music with the desired emotional characterization, merely relying on their own movements and emotions.

3.2.2.1 *Design process*

As discussed by Wiethoff et al. [141] , the design of interactive spaces for collective creativity presents some unique challenges. The Music Room evolved through three months of user-centered iterative design, involving a conceptual stage enriched by early evaluations of scenarios and storyboards and continuous testing of an evolving experience prototype. The ultimate goal of the process was to enable everyone to experience musical creativity through social interaction. To compose their music, users needed to collaborate with each other in order to get the desired level of emotionality and intensity. During the conceptual design phase, we envisaged two basic scenarios, allowed by the manipulation of the proxemic cues of proximity and speed.

- *Acting scenario.* Users perform a drama and create its soundtrack at the same time. They can act out and perform a storyline, for instance a tragic event with a happy ending. Thus, at the beginning they could stand far from each other, eventually moving to opposite corners of the room. After this tense and tragic stage, they would gradually get closer. As they did so, the music would assume an increasingly cheerful character.

- *Dancing scenario.* Users enjoy the installation by dancing to the music played in the room. Two interaction styles were envisaged: either users passively feel the music, by synchronizing their movements with the beat, or they actively influence it, as to map the dancing style they desire. If they stand far from each other they will achieve a jazzy, improvisational style; on the contrary, if they stand close together they will get a more romantic music.

These scenarios were enriched with graphics and storyboards and used as probes in a design workshop involving 12 user experience researchers, and in 5 contextual interviews with students. Then the basic technological architecture of The Music Room was assembled by interfacing Robin, the music composition system and a visual tracking system. This architecture was tested and fine-tuned off-line through video analysis of scenes specifically recorded for the project and showcasing the two driving scenarios.

As the architecture developed into a stable system, the evaluation moved into the laboratory and, only the final week before the Researchers' Night (due to some organizational constraints), into the actual display area. That week was an intense exercise of choreography, implementation and evaluation with a diverse sample of visitors. Design decisions were verified and eventually modified. For instance, the idea of introducing speed as the main interaction channel for controlling music intensity emerged from behavioural observations and user comments. In particular, an eight year-old child testing the room with his mother used it as an augmented

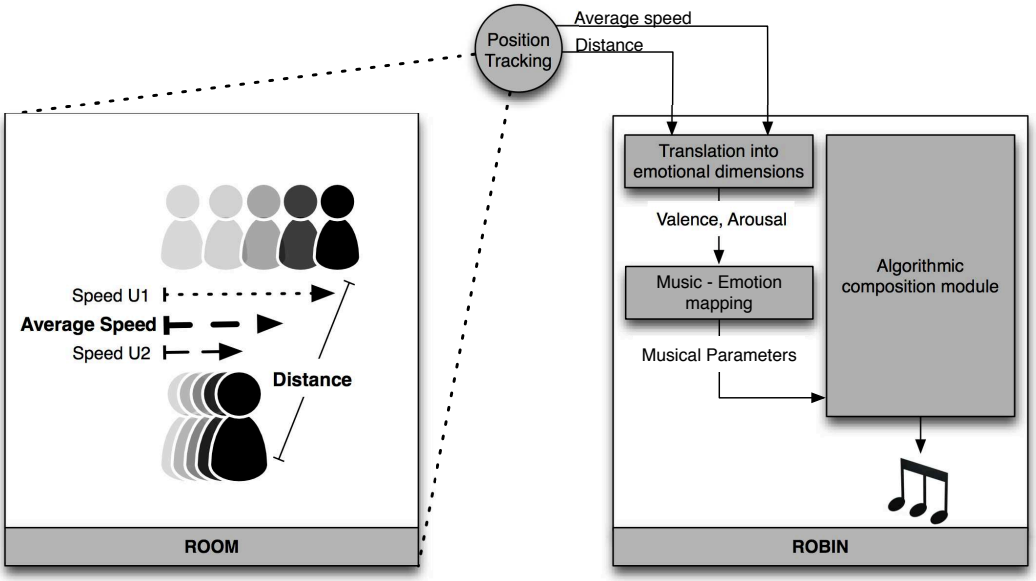


Figure 17: Architecture of The Music Room.

playground engaging in a musical chase and tag competition. The same requirement was also clearly requested by some student visitors.

3.2.2.2 *Architechture*

The Music Room is composed of two main building blocks: a visual tracking algorithm and Robin, the music composition system (Figure 17). The first module consists of video analysis tool aimed at detecting the moving objects captured by a camera installed in the room, in order to extract proxemic cues from the participants' behavior. These values are sent via OSC protocol (Open Sound Control) to Robin, that, in turn, transforms the proxemic information into the emotional cues of valence and arousal, and then into musical variables. Finally, these musical variables are fed to the Algorithmic composition module, thus determining a change in the generated music.

3.2.2.3 *Extraction of the proxemics cues*

Motion of the participants in The Music Room was captured through a bird-eye video camera installed on the ceiling of the room which looked downwards. This configuration of the camera allowed us to minimize the risk of occlusions, which is intrinsic to any motion tracking application, thus limiting the occurrences of false and missed detections. The detection of the moving subjects has been implemented by applying a standard background subtraction algorithm [67]. The obtained foreground

information was then processed by the CamShift tracking algorithm [16]. CamShift has been chosen mainly because of the reduced computational burden, which enabled us to keep up with real-time constraints. The proxemic cues which have been considered in our scenario are distance and speed, as they provide a reliable picture of the ongoing interpersonal relationships.

Figure 18 displays a view of the room as seen by the camera. Two different instances of the interaction are shown on the left, while the output of the detection and the tracking module are portrayed on the right. In particular, the algorithm capability of dealing with partial occlusions, occurring when the two subjects approach, can be observed in the bottom-right image.

However, even in this simple environment, the tracking algorithm may fail to return the right positions of the subjects or wrongly track some artifacts created by shadows. To overcome this issue we assumed that the subjects are the two larger blobs in the area. Even after this precaution sometimes the trak drifts or simply one of the two subjects splits his blob in two parts forcing CamShift algorithm to follow the wrong track. For this reason we proposed a *human-in-the-loop* framework that corrects this circumstance.

$$\check{D}_{ij} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_1-x_2)^2+(y_1-y_2)^2}{2\sigma^2}} \tag{3.10}$$

The position of the participants is then tracked progressively over time and the distance between them has been calculated and normalized using Eq. 3.10. The subjects' speed has been computed averaging the distance they cover in each frame, averaged over a time window. The global speed is then computed averaging the speeds of the two participants.

3.2.2.4 *Robin*

Robin is the result of several years of research carried out at the Experiential Music Lab of the University of Trento. The compositional approach is rule-based: the algorithm is taught a series of basic compositional rules of tonal music that are used to create an original composition in Western classical music. A set of compositional constraints guarantees musical coherence and a pleasant output. The major novelty of the algorithm, compared to related works [125, 127, 138], lies in its potential to adapt in real-time to users' intention. Depending on the user input and on the internal state of the system, which is determined by previously generated music, the best possible choices of harmonic progression, chords, rhythm and melody are computed at each bar following a statistical approach. Music quality is also improved by the iteration of short themes simulating features such as chorus and verses, that are typically found in every musical genre. At each new step, the system statistically

Figure 18: Two different instances (a,c) of the same experiment with the respective foreground maps portrayed (b,d). The second situation (c,d) shows how the algorithm is effective even when the two participants are close to each other.

decides whether to repeat the main theme according to the current state of the system or to progress to the generation of a new one. Details of the algorithm can be found in [88].

The information supplied by the proxemic cues computed by the tracking system conveys the intended emotionality. Distance and speed explicitly communicate the users' moods, as well as the desired level of intimacy according to the analogy of love. At this end, high proximity is mapped with positive emotions and low proximity with negative emotions; high speed with intense emotion and low speed with mild emotions. By matching the values of speed and proximity to emotion, Robin adapts the musical flow.

The mapping between emotion and musical parameters is based on related literature which puts the emphasis on mode and tempo as key factors for emotion elicitation [66, 89]. Robin creates emotional music primarily by modifying mode and tempo. In addition, it alters melody direction, theme re-currence, volume and consonance. This set of elements counts three discrete (mode, melody direction,

Table 7: How movements are mapped into music.

	INTIMATE SPACE	PERSONAL SPACE	SOCIAL SPACE
SLOW MOVEMENTS	Major mode High tempo Low volume Theme recurrence	Minor mode Mid tempo Low volume	Dissonant Very low tempo Low volume
Emotional Outcome	Romantic, serene	Sad, boring	Tragic, unpleasant
FAST MOVEMENTS	Major mode High tempo High volume Theme recurrence	Minor mode Mid tempo High volume	Dissonant Very low tempo High volume
Emotional Outcome	Happy, joyful	Frightening, depressive	Intense, unpleasant

theme recurrence) and three continuous variables (tempo, volume, consonance) [88]. The proximity value influences the emotionality of the song by means of mode, harmonic direction, consonance and tempo (BPM). High proximity is mapped into major mode, rising melody and fast tempo; medium proximity is mapped into minor mode, descending melody and medium tempo; low proximity is mapped into dissonant melody (a number of out-of-scale notes) and unnaturally changing tempo (very fast notes followed by long ones). In addition, in order to increase the difference between positive and negative situations, the recurrence of the theme was treated as an expressive quality of music. In his pioneering work, Emotion and Meaning in Music, Leonard Meyer explained how fulfillment of expectations can elicit positive emotions while listening to a piece of music [86]. We operationalized the concept of fulfillment of expectations by presenting a theme numerous times. In contrast, we avoided any repetition to induce negative emotions. Theme recurrence was triggered by users through proximity: the closer they were, the higher was the probability of theme repetition. Lastly, the speed of users determined the dynamics of the song by means of the volume: fast movements were linked to high volume, while slow movements were linked to low volume. Table 7 summarizes the main differences in music parameters and emotional output in 6 conditions, according to the two proxemic cues manipulated: proximity and speed.

The algorithm was implemented in SuperCollider and it communicated via OSC (Open Sound Control) with the tracking system. The output of the SuperCollider patch was a score in MIDI format. Logic Pro, a Digital Audio Workstation, transforms the MIDI flow into piano music.

3.2.3 *Evaluation*

The first version of The Music Room was presented at the EU Researchers' Night on September 28th 2012 in the city center of Trento (Italy). The EU Researchers' Night is a European event which involves 300 venues and is aimed at showcasing research results to a broader population. In Trento, the Researchers' Night took place from 5pm to 2am and it hosted almost 90 demos and installations. The audience attending the event was quite varied: among others, it included students, researchers, families with children, and bystanders.

3.2.3.1 *Settings*

The installation took place in an empty 25 sq m room. The room setting had to be as static and minimal as possible, as to draw the attention of people to the musical cue only. The only visual decorations in the room were a sticky tape square delimiting the walking area on the floor and eight spotlights placed in the corners, four on the floor and four on the ceiling. All spotlights were hidden by yellow and orange decorative papers. The display area was separated from the control area (i.e., a table equipped with some computers and a mixer) by a thick curtain. This increased the level of intimacy of the experience. Each couple was explained the rules of The Music Room before starting the experience: music would be influenced by their own movements.

3.2.3.2 *Procedure*

Before entering the room, people signed an informed consent providing a brief introduction to The Music Room, explaining that their behaviour was videotaped, and asking them to express a preference on how they wanted the video to be treated (immediately deleted, used only for research purposes, published to showcase the installation). Then, they received a short verbal introduction to the installation and its features. In particular, they were informed that their proximity level and the speed of their own movements would enable them to influence the music. Visitors were invited to experience the installation for as long as they wished. However, several couples were invited to leave because of the long queue which gathered around the installation. During the first hour the waiting time was relatively short (less than 15 minutes) but it gradually increased to almost two hours at around 9 PM. Hence, we

resorted to a queuing system: the names of those who wished to try the installation were written on a list, so that they could get on with their visit and only come back to the room when their turn would actually arrive.

After they had left the room, couples were asked some questions about their experience. Three researchers in parallel interviewed the couples. A qualitative approach was chosen, as this was the first demo and there were no clear expectations on how the audience would react to this novel experience.

1. What do you think about the installation?

2. Did you feel as if you were controlling or following the music?

3. What would you change? Is there anything that you did not like in particular?

In addition to the above-mentioned questions, couples were also encouraged to express any further comment if they wished to do so. Finally, people were given a card with a URL address and a personal code, which could be used to download their own music composition as an mp3 file.

3.2.3.3 *Data analysis*

A hidden camera was mounted in the room to audiotape the visitor's performance from a perspective which allowed to see their main emotional reactions[2]. For the video analysis we managed to use 85 videos. Out of a total of 87 couples, only two did not consent to the processing of data for research purposes. Videos were analyzed by 3 researchers and coded thematically. The variables of interest were defined following a combination of top-down and bottom-up code development [15]. Thematic analysis is a general method that involves the creation and application of codes to qualitative data. 'Coding' refers to the creation of categories in relation to data. Some categories, related to the behaviour of participants in The Music Room, were coded top-down following the two design scenarios proposed during the conceptual design (i.e., acting and dancing). Other categories (e.g. room exploration, running) were derived from the observation of the videos. The videos were initially watched by the researchers independently, and the new coding scheme was iteratively defined by discussion. Inter-rater reliability was finally computed yielding satisfactory values (88% agreement). SPSS was used as the tool of analysis whereby relevant couple's behaviour during the sessions were identified and coded consistently.

Answers to interviews were loosely transcribed with pen and paper during the event, as no recording devices could be used due to the chaotic contexts where

2 Videos from The Music Room can be viewed at http://youtu.be/92UDoy8QCDs and http://youtu.be/qbmETsxcVc0.

the interviews were performed. The transcripts were then transcribed on a digital support. The coding process was guided by the three main questions.

3.2.3.4 *Results*

From the opening, The Music Room was constantly busy. Both attendee reviews and the long line suggested that The Music Room was extremely successful. In total, 87 couples (174 people) experienced the installation. The age of visitors was the most variegated: 57 couples of young people (roughly between 16 and 30 years of age), 19 couple of adults (30-60), 5 couples of children (younger than 16) and 6 couples composed of a child and an adult. Session lasted on average 5 minutes (from a minimum of 1m30s up to a maximum of 10 minutes), although these data are biased by the long queue that forced many participants to limit their session.

A common behavioural pattern emerged over time in the performance of the majority of the couples (75%). When entering the room, they initially explored both interactive dimensions of the installation. In this phase, people moved in the space to perceive how proximity affected music, and they changed their speed often ending up running. During this time, they appeared to be more interested in testing the interactive space, rather than enjoying the musical experience per se. However, they eventually found themselves dancing to the music they were creating. The remaining 25% of the couples immersed directly into the experience, dancing or moving closely, and skipping the first exploratory phase.

The most common behaviors in the room were dancing, walking and running. As predicted by the dancing scenario developed in the conceptual design phase, most couples (72%) spent at least 30 seconds dancing. The majority of them (75%) danced closely, often following the steps of traditional couples dancing in western culture. The level of intimacy varied within and between couples, and different dancing patterns emerged. At times, people engaged in slow dances: the lead partner held their hands on the following partner's waist, who draped their arms on the lead's shoulder. They also engaged in waltz style dancing: touching the partner's right hand and hugging their waist with the left. Sometimes, instead, they danced independently, often paying little attention to the partner. In these cases, they often closed their eyes. A significant number of couples (45%) kept moving the whole body (arms, hands, legs), jumping, running after each other and even twirling around. This behaviour was common among male children, who completely disregarded the dancing possibilities but tended to transform the room in a tag playground. Female children engaged in both behaviour, naturally passing from dancing to playing. Figure 19 displays some of the most common scenarios.

The acting scenario we envisaged during the conceptual design phase was less common. Only 25% of the couples spent some time acting as if they were performing a drama or playing an instrument. In around 20% of these cases we could identify

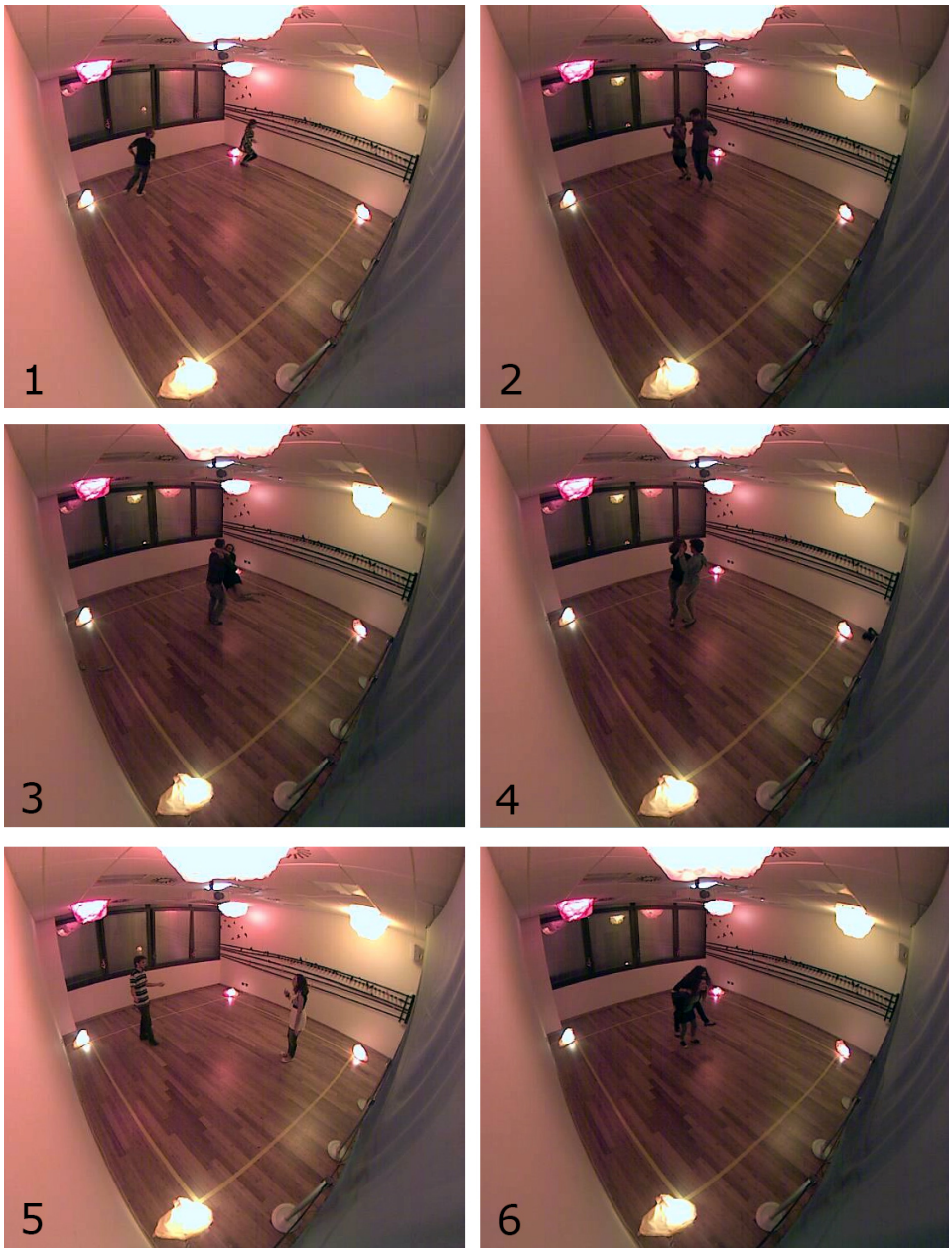


Figure 19: The hidden camera used for the video analyses revealed a series of recurrent behaviors: 1) Running after each other; 2) Jumping; 3-4) Intimate dance; 5) Discussions; 6) unexpected behaviors.

a clear leader in the couple, as if one person was taking charge of directing the scene. In general, the room allowed a considerable amount of communication and collaboration. Almost 59% of the visitors were constantly talking to each other in order to define a joint composition strategy (e.g., going to the opposite sides of the room, getting close, running, doing sudden movements). The remainder of the sample preferred a more intimate and less planned experience, though almost everybody had some form of social interaction mediated by words and/or physical contact.

By observing people's behaviour it was evident that everybody deeply enjoyed the experience. People were often laughing, smiling or openly commenting on it. These impressions were confirmed by the short debriefing interview. All users but two regarded their experience as extremely exciting. The most common words used to describe it were '*cool*', '*interesting*', '*unique*', '*intimate*', '*pleasant*', and '*relaxing*'. Several people also commented on the quality of the music, which was often perceived as barely distinguishable from that produced by a human musician. The two people who did not like the experience complained about a lack of interactivity in The Music Room as if they wanted something physical to be explicitly manipulated. The question concerning the perceived level of control on the musical outcome evidenced a dichotomy. Nearly half of the users reported that they felt as if they were actively controlling the music, while the other half declared that they were mainly following the music, only having the impression of playing an active role in a few situations.

Most people just replied 'nothing' to the question concerning possible improvement and desirable changes. However, some interesting suggestions were collected. The most important complain stressed an annoying latency between the users' movements and the sound they generated. Several interviewees argued that the response should have been prompter. Other people would have appreciated a wider range of musical genres and instruments, or suggested the possibility of associating a different instrument to each person. Some visitors also suggested extending the set of actions through which users can influence the music, including hand and arm gestures. Several users pointed out that the system might be potentially exploited for therapeutic purposes and for the realisation of soundtracks for short movies and video games. Finally, several participants provided some interesting observations concerning the layout of the room. They suggested setting up The Music Room in a larger and darker space, as to foster intimacy. Other people suggested using lights to improve the general choreography of the room: they pointed out that light might be used to recreate a nightclub context, and perhaps change according to the movements of users.

3.2.3.5 *Discussion*

The video analysis and the interviews confirmed the suitability of the two scenarios we had envisaged at the conceptual design stage, with a strong predominance of the dancing scenario. Natural inclination might account for this behaviour, as people usually start dancing when confronted with music. Some interviews, supported by data gathered by the video analysis, put the emphasis on the expectations that people had toward the installation. Even though they had all been informed on how they could interact with the systems, almost half of them repeatedly moved their whole body, expecting their hands and arms movements to produce a change in music. A relevant number of people frequently expressed their enthusiasm towards the experience by twirling, jumping and playing with the partner even though they were aware that those movements were not influencing the music. This behaviour witnesses to the desire of some users to exert a greater control on the musical output and on the creative process itself. Several interviewed explicitly reported this wish: "*I'd like to control music by means of arms movements as well*", "*I'd like BPM to be synchronised with my steps*", "*I'd like the two of us to control distinct instruments*".

Video analysis showed that almost 40% of the couples did not collaborate as expected, preferring an isolated experience, sometimes accompanied by solitary dancing. Future versions of the installation should definitely include ad-hoc questions, aimed at further investigating this behaviour; yet, it can be argued that it is ascribable to shyness, lack of intimacy between users or to the personal attitude of some people towards collaborative contexts. In some cases, that desire to individually change the music which has been mentioned above can also account for this behaviour. Interestingly, several other unexpected themes emerged. Two couples of skilled dancers analogously reported that directing the same music they were dancing to was a completely new experience and expressed strong a interest in adopting a similar approach during their performances. Some users declared that for the first time they were not worried about dancing awkwardly, because it was music that was actually following their own movements. Their movements, the music they generated, their dancing in reaction to the music etc. originated a feedback circuit.

The most critical issue was attributed to the slow response of music to users movements. Even though the musical system is specifically designed to synchronise music generation with user inputs, we purposely decided to avoid sudden changes in music. This choice was mainly dictated by aesthetic reasons: the phraseological structure of music should be preserved even in case of sudden changes in the emotional input. Nevertheless, this issue should be regarded as a non-fulfilment of our expectations, and as such it deserves further analysis. The next versions of the system will delve into the trade-off between expressive freedom and ecological validity of musical outcome. Had the control of the composition been tied to individual beats instead of a bar, how would this have affected the user experience?

Is there a balance between expressive freedom and natural and pleasing music for this particular kind of installations? This questions spark interesting debates about the nature and the ultimate purpose of the system. If users were allowed to directly control lower-level features, a different learning curve should be expected. This change would also imply a wide range of musical possibilities and a high involvement of users.

Another interesting aspect that emerged from the evaluation was the dichotomy between users who felt as if they were controlling the music and users who felt as if they were merely following it. Some of the users belonging to the latter group claimed that: "*At first, we did try to control the music and it partially worked. Then we just followed it*." or "*I felt passive, I was not really controlling the music*". As feeling passive during the experience may be ascribed to minor problems with the tracking, fixing this issue would possibly increase the percentage of people who feel like actively controlling the music. At this regard, we speculated that such a sharp divergence might be due to the fact that those who mostly enjoyed the installation were also those who felt as if they were controlling the generated music. Yet, no correlation between appreciation of the installation and perception of control was found.

Even though the near totality of the participants described their experience as engaging, half of them limited themselves to simply following the music. We had taken this possibility into account, as we specifically formulated one of the three questions at this regard (*Did you feel as if you were controlling or following the music?*). Yet, their creative experience seems to have been unsatisfying, or not as satisfying as we had expected. Whether their experience can still be considered creative and to what extent the collaborative element actually fostered creativity is still debatable. For future implementations of The Music Room we plan to perform ad-hoc creativity evaluation techniques, as to measure the creative outcome in a defined framework such as consensual assessment [4].

3.3 DISCUSSION

In this chapter we have proposed a novel method to infer social interactive behavior among people combining traditional metrics based on distance and velocity, and proxemics cues. Proxemics is handled as an intentionality parameter, giving the opportunity to better focus on the events of interest by considering only the moving subjects whose motion patterns demonstrate a will to interact. The method has been evaluated on three different datasets, confirming the viability of the method in recognizing different types of interactions. One of the datasets, specifically designed for social interactions analysis is provided by the authors as an additional contribution of the paper.

In section 3.2 we proposed an experimental artistic installation where the two participants were induced to collaborate in order to produce a brand new melody as an outcome of their movements.

In this scenario, The Music Room was conceived as an interactive space where everybody has access to collaborative musical creativity. The social aspect clearly emerged, as half of the participants explicitly discussed possible ways of influencing the musical outcome, actively taking part in decisions and establishing decisional roles. The absence of tasks accomplishment forced them to engage in the creative choice of their own acting storyboard, dancing techniques and performance style. The several possibilities offered by this open-ended activity may account for its clear appreciation among users. Individuals and groups could shape their experiences according to their own desires, originating several distinct behaviors.

This success showed open-ended tasks to be a valid approach in the context of creative support interactive systems. The combination of movements and emotions proved to be an appropriate mediator of musical intentions. The introduction of an intermediate layer hiding the complexity of music creation may be of interest to designers and researchers who aim at creating meta-creativity environments where users are encouraged to experience art and to express and elicit emotions. This intermediate layer actually enables users to deal with perceptual and semantic descriptions rather than low-level features that are not accessible to everybody. This suggestion is to be also observed in the context of visual art products, as witnessed by the considerable success of software applications such as Instagram and Photoplay, which allow users to effortlessly lend a certain charm to their own videos and pictures by simply applying filters that can simulate vintage and professional cameras and recreate blurring effects.

In next chapter we are going to project the findings of the proxemic analysis on a real-world scenario, fusing the high-level information with low-level visual features contribution.

4

LOCALIZING UNSTRUCTURED SOCIAL INTERACTIONS IN REAL-LIFE SCENARIO

In this last chapter we are observing human interactive behavior from a closer perspective. We are merging the high level information from proxemic theory with local visual descriptors to detect and localize fights in a real world scenario.

Real life videos are of paramount importance to benchmark video analysis algorithms, especially in surveillance scenarios. In this chapter we are focusing our attention on two shortcomings of computer vision literature. On the one hand we provide a method to detect and localize dyadic human interactions based on the interpersonal space. On the other hand we present a fully annotated dataset consisting of 59 different videos of a specific type of interactions, namely urban fight situations. Videos are collected from YouTube, and are meant to provide researchers with a common ground for comparison. The proposed dataset can be considered as one of the most challenging annotated video collection concerning dyadic interactions, due to the intrinsic intra-class variability that characterizes real fights. The idea stems from the significant difference between an action performed by a single subject and an interaction between two persons. In the first case all the visual information is concentrated on the subject, while in the latter case the action of a person is related to the interacting person's attitude, following an action/reaction principle. This kind of behaviors is significant especially in natural and real-life scenarios, in which people are moving freely without the awareness of being recorded. We provide an extensive experimental analysis on this dataset and demonstrate that the visual information extracted in the area associated to the interpersonal space plays a fundamental role in detecting fights. We also validate our algorithm in distinguishing between urban fights and street dancing, a class of interactions that exhibits similar features but in a more structured fashion.

Parts of this Chapter appear in:

- Rota P., Conci N., Sebe N., Rehg J.M. **Real-life Violent Social Interaction Detection**. IEEE ICIP 2015 *(submitted)*

Figure 20: Snapshots taken from the proposed dataset. We intend to emphasize the intra-class difference in fight interactions even if all samples are captured in the similar urban scenarios.

In the last few years, computer vision and machine learning have brought significant contributions to the creation of models capable of recognizing human actions in videos [12, 75, 82, 122]. The huge interest in this field as well as the performance improvement in detecting and analyzing visual features extracted from videos, have pushed researchers to consider also situations involving more than one person [58, 60, 94, 101].

In the literature many works have been addressed towards algorithms that can be considered as an extension of the models commonly used to classify individual actions [101, 116]. When we talk about interactions, for what concerns computer vision at least, we are considering a collaborative and mutual dependent series of body gestures. This intuition leads us to say that so much information regarding this particular kind of actions is behind the causality and the synchronism of dyadic movements. For this reason, unlike some landmark works on social interactions such as [82], we are not interested in labeling each single video in a category, but our model is conceived to analyze every pair of subjects in the scene in order to infer whether an interaction is taking place or not. In some previous works such as [30, 75] social interaction concept falls into the word *activity* referred to the principal collective action performed in the video. According with the previously mentioned papers, they put in relation all the individual action labels for of each

subject in order to infer the overall activity label. In this work we rely on the direct dyadic connection to inspect the ongoing type of interaction.

Another important element proposed in this work is the concept of unstructured interaction. In the majority of the datasets proposed so far [82, 100, 117], social physical interaction has been considered as a short, well defined set of cues.Let us take hand hand-shaking as an example: two people approach each other, put forward their arms, shake them, and go back in the original position. These sub-activities have unique meaning, no matter what the situation is or who the subjects involved are. The same could be for hugging, kissing, etc. In this work we are interested in those kind of mutual actions that do not have any predefined structure e.g., street fighting.

In order to emphasize this novel aspect regarding social interactions we propose a new dataset of videos retrieved from YouTube with different kinds of physical unstructured social interactions. A particular type of interaction that matches the proposed characteristics is *urban fighting* which is very difficult to analyze. In fact in some public datasets [9, 24, 114, 117] there are several instances of fighting but they are staged and it is always clear that occasional actors are not performing naturally. Another aspect is the length of the events in the datasets mentioned above. Since their purpose mainly regards events classification, the length of every execution is below 5 seconds. In our dataset have many *urban fights* in which the interaction lasts for more than 10 seconds and they often include more fighting instances in a single video. To balance and validate the algorithms we added a further type of physical interactive behavior, namely *street dancing*. In this case, unlike in fighting, the interaction is more deterministic and harmonic, while still preserving its unstaged nature.

As a further contribution, a novel method to detect and localize pairwise unstructured physical social interaction is proposed. The method relies on the definition of an interpersonal area between the interacting subjects in which the motion cues are intensified and therefore are more discriminative. As a further motivation, the interpersonal space includes inherently the proxemic information among the interacting subjects, providing an important contribution that can not be provided by the visual features alone.

The chapter is organized as follows. In Section 4.1 we provide an overview of the related work, in Section 4.2 we present in detail our new dataset, while in Section 4.3 we propose our evaluation methodology. In Section 4.4 we present the detection results, while conclusions are drawn in Section 4.5.

Table 8: A comparison among most popular datasets for social interaction analisys

Dataset	Number of sequences	Resolution	Scenario
UT-Interaction Dataset [117]	20	720x480	Staged
Caviar Test Case Scenario [24]	28	384x288	Staged
BEHAVE Interactions [9]	8	640x480	Staged
Collective Activity Dataset [29]	74	Variable	Natural/staged
UCLA Courtyard Dataset [5]	6	2560X1920	Natural
Hockey Fights Dataset [92]	1000	360x288	Single Scenario
TVHI dataset [100]	300	Variable	Movies
Hollywood2 [82]	2517 (771 are interactions)	Variable	Movies
RE-DID (Our Dataset)	**59**	**1280X720**	**Natural**

4.1 RELATED WORK

4.1.0.6 *Social Interactions*

The study of social interactions in computer vision has generated a big interest especially in the area of video surveillance, where the detection of atomic actions could not be sufficiently representative of the observed scenario, and because in most cases it does not perform well due to the significant clutter in public areas. An example of a high level behavioral model is proposed by Mehran et al. [85] where the authors recognize abnormal events in a social scenario by analyzing the space-time flow of a grid of particles. In [30] the authors propose a method to recognize collective human activities modeling a *Crowd Context* framework where the motion cues of individuals are expressed in terms of proximity to other subjects in the scene. This work is the closest to ours although their model is centered on each subject, so they are not answering to the question *with who is he interacting?*. In [75] the necessity of the contextual information is highlighted to perform a reliable activity recognition. The authors propose a multi-potential approach where different aspects of the activity of each subject in the scene is considered (e.g., image-action, action-activity, action-action, image-activity). In [69] the authors use a mixture of Switching Linear Dynamic Systems to jointly infer common and unusual actions from people trajectories. In [5] a dual method to perform activity recognition is proposed, fusing low-level and high-level approaches to achieve dual inference on actions and activities using AND/OR graphs.

4.1.0.7 *Dyadic Interactions*

With dyadic interaction typically we consider a pair of individuals in the video that are performing a cooperative action. Previous works [18, 114] propose a high level

approach exploiting proxemic cues. In the work by Rota et al. [114] cost functions are generated from the motion trajectories of the subjects, and utilized to classify the interaction type; Calderara et al. [18] propose instead a set of proxemic-based features to detect abnormal behaviors through grammar analysis.

An interesting contribution to dyadic interaction understanding has been proposed by Prabhakar et al. [106], where the authors outline a novel temporal model to extract causality features from unstructured videos. The experiments highlight the possibility to detect redundant movements discarding other visual cues. The work is then extended in [107], with the goal of discovering turn taking interactions exploiting a multiple instance learning model.

4.1.0.8 *Fight Detection*

In terms of detection of violent situations in videos, an early work [90] proposes the analysis of the combination of different visual features (blood, flames, etc.) and audio features (explosions, screams, etc.). Chang et al. [27] propose a multi-camera framework to detect and predict aggressive behaviors between groups of individuals such as gangs in prison yards; in their model each individual is tracked along the monitored area. They propose a hierarchical clustering to define groups of individuals in order to detect predefined behaviors such as loitering, flanking, and aggressive group behaviors. Nievas et al. [92] propose a classification problem to detect hockey fights in short video clips collecting visual features over the whole frame. Hassner et al. [56] propose a real-time detection model of violent crowd behaviors using flow information.

In our approach we focus our attention on dyadic aggressive interactions, so we are not evaluating group aggressions as in [27]; on the other hand unlike [92] we focus our attention on a more restricted interpersonal space collocated between the two opponents in order to prune out the visual features not related to the ongoing event. Our final goal is also to detect the fight situations on-line as soon as they happen in the video.

4.1.0.9 *Features*

Many different ways to extract visual features have been proposed in literature. For our experiments we refer to the work by Tamrakar et al. [130], which reviews different kinds of state-of-art visual features. In particular we will test Dense Trajectories based on Histogram of Gradients [36], Histogram of Flow and Motion Boundary Histogram [37].

4.2 REAL-LIFE EVENTS - DYADIC INTERACTIONS DATASET (RE-DID)

A crucial contribution given by this work consists on the collection of a new dataset for dyadic interactions called Real-life Events Dyadic Interaction Dataset (ReDID). The main motivation that pushed us to propose a new dataset is given by the lack of annotated videos recorded in real-life scenarios, picturing unstructured challenging interactions performed by a pair of subjects.

To support our claim Table 8 shows some characteristics of the most popular datasets published so far. To be more specific, the data proposed by [9, 24, 117] are more focused on dyadic/small group interactions but the number of situations is smaller than ours and, more important, videos are staged so there is no spontaneity in performing actions. Datasets [5, 29] are more dedicated to group interaction and single action recognition; they also lack of unstructured interactions that are substantial in our work. The dataset proposed by Nieves et al. [92] is composed by very short clips (50 frames each) taken from ice hockey games including fight scenes and normal game situations. The dataset has a retrieval/classification purpose, the fights are taken from the same scenario, there is no huge variability in the way the fights take place. Additionally, these fights are far from the violent situations in a surveillance context our dataset directly addresses.. Other datasets [82, 100] are more dedicated to a classification/retrieval activity, since they are composed of a higher number of videos. Moreover, videos in these datasets are short and often recorded in multiple shots from movies.

In Re-DID we propose a collection of 59 single shot videos containing different instances of unstructured urban social interactions that involve physical contact recorded into the wild. The collection of the dataset has been inspired by the availability of many sequences recorded through a Dash-Cam and then uploaded on YouTube. This configuration allowed us to collect 30 different videos of people fighting for real in an urban scenario. Along with fighting samples we also introduce a set of 29 additional different social interactions videos depicting people couples dancing on the street. In both cases the interaction can be considered as *unstructured* since each single instance differs from the rest under many parameters such as temporal length, physical involvement, degree of contact and approaching strategy. Nevertheless, these fights can still be clustered easily in the same category by a human observer. From the computational point of view there are no systematic patterns that can be easily retrieved as in other dyadic interactions such as kissing, hugging, handshaking etc. Under a video surveillance scenario the difference between *urban fighting* and *street dancing* is clearly crucial and detecting fights allows for dealing with them immediately.

In the case of fighting, the literature [9, 24, 117] provides similar situations but they are mostly recorded in a staged scenario simulating fight events among two or more people. This is due to the difficulty in recording naturally occurring fights in real video footages.. Furthermore, the dynamics of these videos is often inconsistent compared to what happens in real situations, and the number of examples is not large enough to obtain reliable statistics.

In Re-DID dataset we propose a collection of events showing different types of fights. We decided to keep the same contextual environment, so all the footages are recorded *on the road*. As a further constraint we wanted videos with similar camera views, single shot and in particular, frontal to the event, so as to capture the scene at the person level.

All the videos in the dataset are retrieved from YouTube; 25 of them are recorded using car mounted Dash-Cams, the remaining have been taken by other devices such as mobile phones. The length of the videos varies from 0:20 to 4:02 (mm:ss) and the resolution has been normalized to 1280x720 for sake of homogeneity, resizing some videos to match the given dimension. The dataset includes of 73 different fight instances and 42 dancing interactions under different lighting (day, night) and weather conditions (sunny, rainy), different original video resolution (native 1280x720, upsampled videos), different camera views (simple wide angle, fisheye, zoomed view), moving and static scenes.

The dataset has annotations of the position of the subjects' bounding boxes for each frame and relative ID, the temporal collocation of the interaction, and the position of the interpersonal spaces precomputed for the ground truth. For what concerns the interaction triggering and ending, we have considered a general rule for the annotation process, starting with the first contact between the involved subjects until a relevant distancing is taking place.

4.3 EVALUATION FRAMEWORK

In this section we describe the method used to evaluate our framework for social interactions analysis. Fig. 21 shows a global overview of the appraoch we propose.

4.3.1 *Dense Trajectories and Visual Features*

In order to capture both shape and motion features, dense trajectories have been adopted. The extraction of the dense trajectories is affine with the proposal by Wang et al. [139]. The interesting points are sampled using Shi and Tomasi algorithm [124] and then tracked using the Farnebäck's implementation [43] of the optical flow. In order to be informative, all trajectories, must respect the following requirements: a consistent displacement (static trajectories and those with a too large displacement

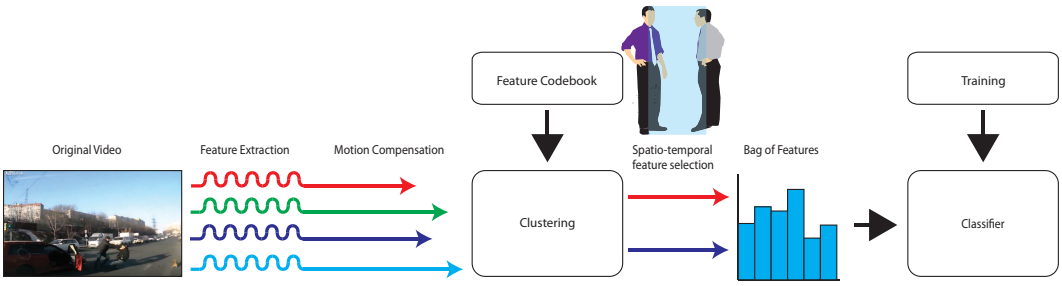


Figure 21: The framework schema for the fight detection evaluation using the interpersonal space.

will be removed), the angle between the trajectory an the ensuing point can not be higher than 90° and those points whose quality factor (according with Good Feature To Track algorithm) is lower than a certain threshold are considered hard to track and therefore discarded. Feature points are sampled on a grid spaced by W and tracked on each scale independently in a pyramidal fashion; short and overlapped trajectories have been pruned, those instead that are longer than a given length are cut in order to prevent them to drift. Most of the videos in Re-DID are affected by high camera motion, that it is fairly fluid in the case of Dash-Cam videos but for mobile recorded videos often degrades into rapid shaking. To improve the cleanness of the trajectories we apply homographic correction according to [140]. The mean coordinate of the points in the trajectory has been considered as the location point for the feature extracted. The descriptors used in this work are essentially shape features i.e. HOG (Histogram Of Gradients), zero-order and first-order motion features as HOF (Histogram of Optical Flow), and MBH (Motion Boundary Histograms) [36, 37]. The HOG descriptor categorizes the gradient information into 8 bins referred to the pixels connected to the current pixel location in a multiscale manner. The HOF descriptor describes the Farnebäck's optical flow information using 9 bins, one more with respect to the HOG that refers to the central location of the 3x3 patch. The MBH descriptor is obtained applying the HOG to the Farnebäck's optical flow image, obtaining two different 8 bins descriptors, one for the vertical and one for the horizontal component. Features are then normalized using the L2 norm.

4.3.1.1 *Interpersonal Space*

We evaluate social interactions according to two different criteria. The first is the *high-level* approach, in which trajectories of subjects are the most informative source. The second way is similar to the one used to classify atomic actions at the pixel-level information, thus analyzing color, gradient, and optical flow information. In fact, the proximity information between subjects and the analysis of the visual features

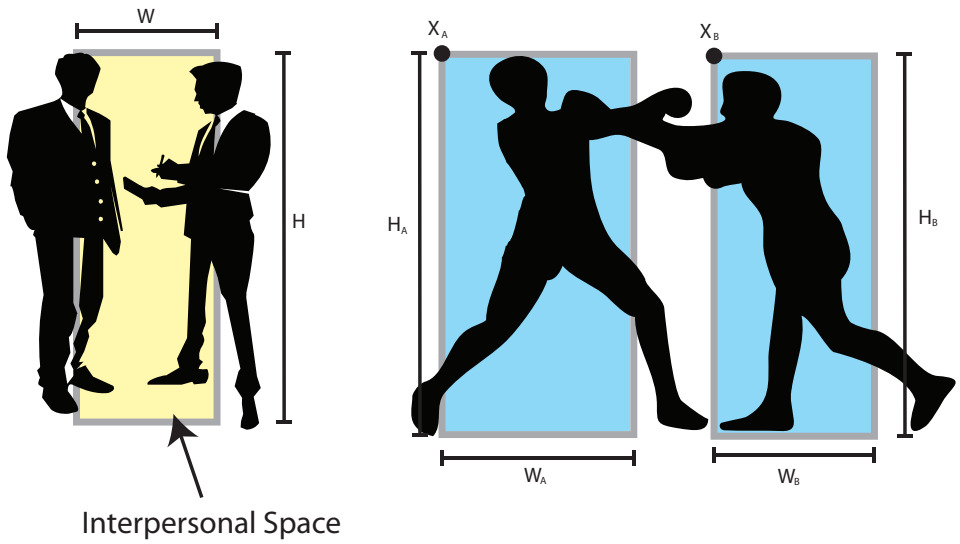


Figure 22: The sketch shows two situations of non-fight and fight, respectively, with the notations used in Section 4.3.1.1 to locate the *interpersonal space* position.

created by the movement of the individuals, can equally contribute to provide an accurate description of the ongoing interaction. To support this claim it is enough to think about the situation when two people are interacting within a close-range distance, in which most of the salient motion information occurs in the space located between the two persons.

Let U_a be the collection of shape and position parameters referred to subject a as depicted in Fig. 22. Let's define what we call *interpersonal space* through the following equations:

$$H = \max(H_a, H_b) \tag{4.1}$$

$$W = |X_a - X_b| \tag{4.2}$$

$$K_w \cdot \min(W_a, W_b) \leqslant W \leqslant W_a + W_b \tag{4.3}$$

$$\left| \frac{H_a}{H_b} - 1 \right| \leqslant K_h \tag{4.4}$$

In Eq. (4.1) and Eq. (4.2) we define the height and the width of the interpersonal bounding box, respectively. In Eq.(4.3) we define the dimension constraints. The first part of the equation prevents the minimum area to be smaller than the narrowest person's bounding box in order to manage the situation in which the involved subjects are occluding each other. The second part is a distance constraint; whenever it is not satisfied we assume that every close-range interaction is not possible. Eq.(4.4) is a

perspective constraint that avoids considering as interacting two subjects that are too far from each other. On top of these considerations the interpersonal space is always centered onto the center of the conjunction line between the two subjects' bounding boxes.

Feature selection has been pursued according with Eq. (4.5) where $\phi_{ab}(t)$ are the selected features related to subject a and b at time t. $\mathbf{x}(t)$ is the whole set of N valid trajectories present at time t.

$$\phi_{ab}(t) = \{\mathbf{x}_i(t)\}_{i=1}^{N} : \mathbf{x}_i \in I_{ab}(t) \tag{4.5}$$

We indicate with $I_{ab}(t)$ the area of the interpersonal space at time t generated by the position and shape parameters U_a and U_b.

4.3.1.2 *Features Processing*

All the trajectories that lie in the interpersonal space are considered, all the others are discarded. The experiments we are proposing are twofold; the first is in the spatial domain, the second consists of collecting the data from a temporal cuboid similar to [120], but unlike them we did not collect data from subvolumes in a uniform grid. In this work the cuboid is defined by the temporal envelope of the interpersonal space bounding box along a predefined Δt. This modification has been necessary since we intend also to localize the interaction. The features are gathered inside the spatio-temporal envelope and processed using a bag of features approach. All the combinations among the subjects are considered but only where the interpersonal space is defined. If the conditions in Eq. (4.3)-(4.4) are satisfied we perform the actual classification based on a linear Support Vector Machine.

4.4 RESULTS

In this section we propose an extensive set of experiments. As previously mentioned, this consists of a combination of simple baseline methods in order to tackle the problem of detection and localization of unstructured social interactions.

In the first place dense trajectories have been extracted cutting all those trajectories longer than 15 frames and pruning those that are not reaching that number of frames. The trajectory density is set through a spacing parameter $W = 5$ and each scale is spaced by a factor $1/\sqrt{2}$. At every frame, re-initialization is performed when no trajectory is found in a 5×5 neighborhood.

As explained in Sec. 4.3.1, features are selected according to the position of subject's bounding boxes and interpersonal spaces (an example is shown in Fig. 23). For each experiment we propose three different approaches: in the first case we collect all the features inside the bounding boxes of the subjects. In the second

experiment we consider only the *interpersonal space* described in Sec. 4.3.1.1. The third is the union of the two. We also propose a further experiment in which we collect all the features that fall outside the people bounding boxes and the *interpersonal spaces*, so as to have an objective validation to prove the viability of our approach. Indeed the results for this last test are much worse than the previous three so numerical results are omitted from the tables (they are reported in ROC graphs instead). For the computation of the interpersonal space we used $K_w = 1$ and $K_h = 0.3$; these parameters are set considering that a human observed from far can be easily approximated by a vertical rectangle and also to avoid to consider people that are lying on a different perspective plan when interacting.

For all the approaches mentioned above, the selected features are collected following the bag-of-features paradigm with L1 normalization, prior to the classification phase on each instance. The descriptors of dense trajectories are clustered using the K-means algorithm, we use $k = 500$ as the number of clusters after a grid search analysis. Two different harvesting methods are proposed: spatial and spatio-temporal. The first consists in the collection of the features in each frame separately, classifying each frame. The second consists in collecting all the features in a spatio-temporal cuboid (as described in 4.3.1.2). In the latter case we used a temporal length $l = 15$ frames.

For classification we adopt the Support Vector Machine (SVM) as binary classifier using LibSVM [26]. The experiments are run using the linear kernel with a training/test based on leave-one-out cross-validation. The soft margin parameter of the SVM has been optimized using grid search tuning.

In the Section 4.4.1 we focus our attention on fight detection in urban scenarios as we consider it as one of the best representations of unstructured social interactions. Our interpersonal space model will be tested on the proposed Re-DID in a frame-by-frame fashion and using a baseline temporal model in Section 4.4.1.2. A validation of the method based on interpersonal space will be showed in Paragraph 4.4.1.1 testing it on the well known UT-Dataset [117]. In Section 4.4.2 we extend the detection to a multi-class problem, adding the class *street dancing*. As proposed in the one-class case a complete set of experiments is proposed.

4.4.1 *Fight Detection and Localization*

The first set of experiments are limited to a single class detection, focusing our attention on *urban fights*.

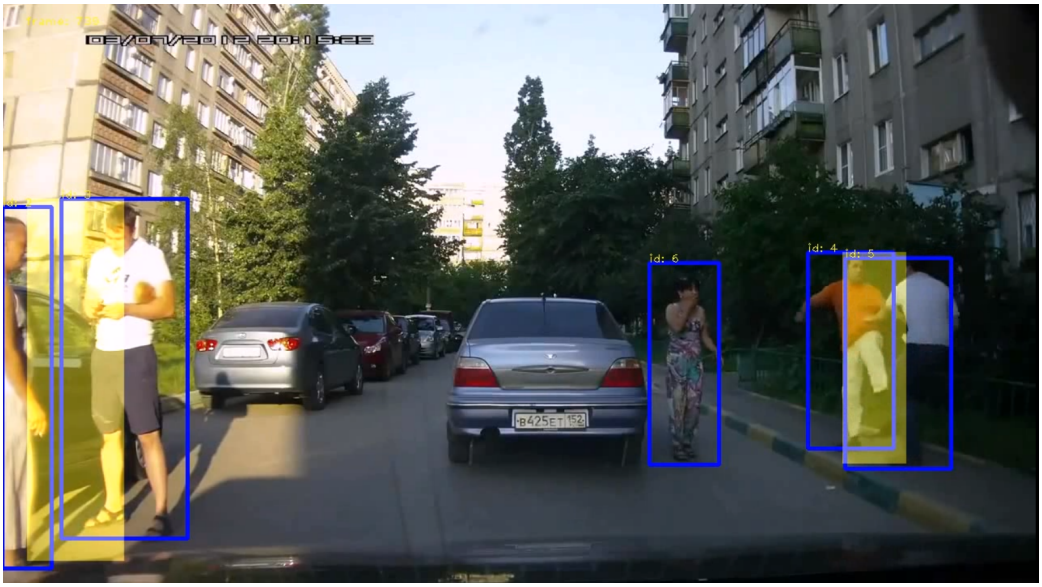


Figure 23: An example of the discriminative capability of the interpersonal space model. In the figure it is highlighted how the two interacting subjects on the left are separated from the subjects on the right, moreover on the far right the two subjects are fighting while the one slightly on their left keeps distance and indeed is not considered as interacting with the previous two.

4.4.1.1 *UT-Dataset*

We use the UT-Dataset [117] to validate our method in addition to the dataset proposed in Section 4.2. In order to fit the videos to our task we put together some routinely social interactions as hugging, pointing and handshaking, considering them as normal type of behavior. On the other hand we gather pushing, kicking and punching as violent interactions. From the results in Tab. 9 we can denote not negligible performance improvement when using the interpersonal space. Analyzing the ROC curves in Fig. 24 (a-b) it turns out, as expected, that the features outside the interpersonal space are much less informative comparing tho those collected inside. In the spatio-temporal cuboid experiments (Fig. 24 b) it is evident a certain blocketing in the ROC curve given by the small number of samples available in the dataset.

As previously stated, this dataset refers to a staged scenario, created on purpose to benchmark very simple structured movements, our method anyway turns out to be very useful even in this situation.

Table 9: Results for Fight detection on UT Dataset and ReDID.

	AUC [%]				
	UT-Dataset		Re-DID		
	Spatial no tracker	Cuboid no tracker	Spatial no tracker	Cuboid no tracker	Spatial tracker
HOG+HOF					
people	75.63	82.36	63.52	67.71	63.55
interpersonal	82.31	87.01	71.21	73.26	75.87
people + interpersonal	81.77	84.51	71.68	71.22	73.59
MBH					
people	76.43	83.77	65.45	64.12	67.43
interpersonal	82.82	88.84	72.14	72.46	76.18
people + interpersonal	82.23	88.39	74.10	70.31	74.73
HOG+HOF+MBH					
people	78.93	84.38	64.94	68.72	65.74
interpersonal	83.41	92.25	72.70	71.33	71.50
people + interpersonal	83.54	88.51	72.09	72.33	73.96

4.4.1.2 *Re-DID dataset*

After getting people's position in each frame of the video we are analyzing one by one each pair of subjects computing their interpersonal space. After checking that the conditions of existence of the space are satisfied (see Section 4.2), we fetch all the valid dense trajectories from the defined space and perform the bag of feature in order to obtain the final vector. The classification is performed using linear SVM.

The results in Tab. 9 are confirming the findings of the previous section. As we expected, the accuracy is lower than in the case of the UT-Dataset because we are now dealing with a real-life scenario where the scene is not always perfectly clear. Anyway our method improves the performance comparing with *people* baseline method. Differently from the experiments on the UT-dataset, the temporal approach does not give much improvement with respect to the spatial approach (frame-by-frame) only. In Fig. 24 (c-d) it is visually evident that the use of the interpersonal space benefits the overall performance.

The last column of Tab. 9 refers to the use of a state of art visual tracker [104] in place of ground truth annotations. The results using a tracker turn out to be slightly better than the ones computed using ground truth, this is because many situations have not been detected by the tracker, especially those with high degree of occlusions, so we can say that these results are proposed in a smaller set.

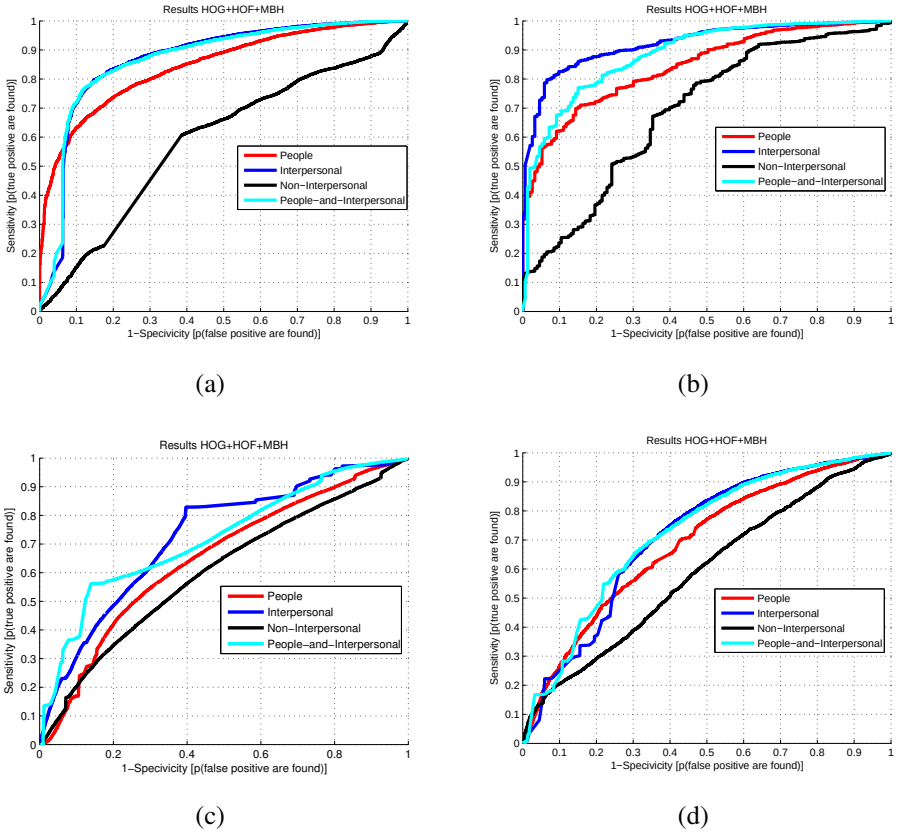


Figure 24: Examples of ROC curves for frame-by-frame fight detection with (b) and without (a) temporal model on UT dataset, and some examples on Re-DID, on spatial (c) and temporal cuboid basis (d).

4.4.2 *Fighting vs Dancing*

In order to extend our experiments and evaluate our findings onto a more challenging scenario, we propose an additional test introducing the new class of *street dancing*. The videos are a collection of different types of dancing clips, all of them recorded in a urban scenario for the sake of consistency with fights. This type of activity shares many properties with *fights*, in both cases we are dealing with unstructured interactions which have no constraints in terms of duration, distance and gestures for the participating subjects. In the dataset there are several instances of different types of dyadic dancing. Movements are generally sudden and causality is a remarkable component. The main difference resides in the general harmony of the dancing interaction itself.

Results are proposed in Tab. 10. Even in this case the proxemic component given by the interpersonal space leads to a clear improvement in the final classification

stage. Moreover, we can also infer that using the interpersonal space alone improves the classification performance as compared to the experiment that includes features in the person's bounding box. As happens in fight detection (Section 4.4.1.2) the temporal information does not bring any remarkable improvement to this bag of feature approach.

Table 10: Results on multiclass experiment on REDID dataset.

Accuracy %		
	Spatial	**Cuboid**
HOG+HOF		
people	40.94	42.46
interpersonal	76.86	71.44
people + interpersonal	75.29	69.97
MBH		
people	41.26	40.42
interpersonal	76.32	71.66
people + interpersonal	74.32	69.85
HOG+HOF+MBH		
people	43.34	43.48
interpersonal	76.74	71.30
people + interpersonal	74.90	69.64

4.5 CONCLUSION

Addressing dyadic violent interactions in a real life scenario turns out to be a harder undertaking comparing with the traditional action/interaction analysis that is mostly performed on staged videos. The new dataset proposed in this work aims to address this shortcoming of the currently available literature. According to the results proposed in Sec. 4.4 we can confirm that proxemic information introduced by the interpersonal space is beneficial in terms of detection of the interaction itself and moreover in the discrimination among different types of attitudes. As future work we intend to gather information from the contextual environment that is, from the human being's point of view, highly important in the discrimination of a violent behavior from normal behavior that is considered harmless.

5

CONCLUSION AND FUTURE WORK

In this thesis we have addressed social interaction analysis from a multi-scope point of view, spanning from crowd to single persons in detail.

In chapter 2 we have highlighted the problems in approaching far range monitored scenes, denoting the complexity in detecting and tracking each single person. We have then proposed a novel approach based on particle motion analysis in order to bypass this problem and to discriminate moving groups of different entities in the watched area. For doing this we have outlined a mutual influence matrix who relates the particles having similar motion properties. The grouping of those particles lead to a remarkable qualitative result in terms of entity discrimination. In the same chapter we have also framed a different type of crowd, named SPECTATOR CROWD. This type of crowd, differently to the one examined in the previous part has a static attitude since it is monitored during big events as concerts, sport games and conferences to name a few. In this work we have proposed a brand new over-annotated dataset related to an international hockey tournament that has been held in Trentino in December 2013. We have tested many state of art algorithms to detect and count people and proposed an innovative solution to estimate spectators' head pose in this low resolution scenario. We have also proposed a brand new method to detect different groups of supporters in the crowd, analyzing their motion patterns for the first time in this research area.

In the next chapter we moved closer to the interacting subjects, analyzing jointly the behavior of each pair of individuals. We will propose a high level analysis that stems from the sociological theory of the interpersonal distance between a pair of individuals (called *proxemics*), these general rules dictate the relationship between all interacting persons. We are approximating this concepts using three high level features related to the actual distance between the subjects, the intentional distance related to their target in the area and an attention factor that is related to the actual speed and the direction of motion of the interacting subject. Social interaction analysis has much importance in many applications. The primary idea is to use it in a video surveillance scenario in order to detect and possibly prevent, terroristic attacks or crimes in general. However, in the second part of the chapter we have proposed a

performing art installation called THE MUSIC ROOM, where a pair of participants are let into an empty room where a music is automatically composed according with their movements and some basic proxemic rules. This experiment has been proposed during *La notte dei ricercatori 2012*, *ICT days 2013* and at the MART museum of Rovereto in august 2014. During the exposition we have observed a huge amount of interests and engagement by the participants.

In chapter 4 we have generalized the findings of chapter 3 picturing a real world test case. We were driven by the intuition that some types of interactions are performed differently while executed in the real world without the consciousness of being recorded. In our case we centered our analysis on fights. This type of interaction has a further characteristics that is the nonstructural attitude. This feature is another factor that augment the complexity since we are not interested in detecting kicks or fists separately, but the whole behavior as equally dangerous. To the best of our knowledge, this type of observation has never been considered in other works. We are approaching the problem adding a proxemic constraint to the visual feature retrieved to every pair of subjects and we demonstrate that this is enhancing the accuracy of the detection of fighting subjects.

5.1 FUTURE WORK

In summary in this thesis we have studied different approaches to analyze social interactions mixing well known sociological concepts with state of art computer vision methods. This research has to be considered as a launching pad to new and more sophisticated behavioral analysis. Some of the possible directions may be the following:

- The new dataset proposed in chapter 2 has been exploited in a very little part. According with the annotations of the spectators, further experiment could be done spanning from traditional computer vision tasks as *action recognition*, *object detection* to some more sophisticated achievements as *social relation analysis*, *attendance engagement analysis*, *group detection*, etc.

- Many of the works of this thesis have the weakness of relying on ground truth in order to locate persons, this is a not negligible problem, especially in urban/video surveillance scenario where occlusions occurs frequently. It might be challenging to expand this topic and inspect also this critical side of the work.

- Much research effort in computer vision has been lavished to enhance the level of description given by visual features. A further investigation on this topic related to the unconstrained scenario of social interaction in the real world

(using the dataset proposed in chapter 4) would benefit the final results and the overall contribution to this type of research.

PUBLICATIONS

This is the list of the peer reviewed publications I have been working on during my PhD:

1. D. Conigliaro, F. Setti, C. Bassetti, R. Ferrario, M. Cristani, **P. Rota**, N. Conci, and N. Sebe.
   OBSERVING ATTENTION.
   FISU, 2013.

2. F. Morreale, A. De Angeli, R. Masu, **P. Rota**, and N. Conci.
   COLLABORATIVE CREATIVITY: THE MUSIC ROOM.
   Personal and Ubiquitous Computing, 2014, 18.5: 1187-1199

3. F. Morreale, A. De Angeli, R. Masu and **P.Rota**.
   THE MUSIC ROOM.
   In CHI Extended Abstracts on Human Factors in Computing Systems, pages 3099-3102. ACM, 2013.

4. **P.Rota**, N. Conci, and N. Sebe.
   REAL TIME DETECTION OF SOCIAL INTERACTIONS IN SURVEILLANCE VIDEO.
   ECCV Workshops and Demonstrations, pages 111-120, 2012.

5. **P.Rota**, D.-T. Dang-Nguyen, N. Conci, and N. Sebe.
   EXPLOITING VISUAL SEARCH THEORY TO INFER SOCIAL INTERACTIONS.
   IS&T/SPIE Electronic Imaging, pages 86670C-86670C. International Society for Optics and Photonics, 2013.

6. **P.Rota**, H. Ullah, N. Conci, N. Sebe, and F.G.B. De Natale.
   PARTICLES CROSS-INFLUENCE FOR ENTITY GROUPING.
   Eu.Si.P.Co. Eurasip, 2013.

7. B. Zhang, **P.Rota**, and N. Conci.
   RECOGNITION OF TWO-PERSON INTERACTION IN MULTI-VIEW SURVEILLANCE VIDEO VIA PROXEMICS CUES AND SPATIO-TEMPORAL INTEREST POINTS.
   IS&T/SPIE Electronic Imaging, pages 866305-866305. International Society for Optics and Photonics, 2013.

8. B. Zhang, **P.Rota**, N. Conci and F.G.B. De Natale.
HUMAN INTERACTION RECOGNITION IN THE WILD: ANALYZING TRAJECTORY CLUSTERING FROM MULTIPLE INSTANCE.
ICME, 2015. (submitted).

9. **P.Rota**, N. Conci, N. Sebe and J. M. Rehg.
REAL-LIFE VIOLENT SOCIAL INTERACTION DETECTION.
ICIP, 2015. (submitted).

10. D. Conigliaro, **Paolo Rota**, F. Setti, M. Cristani, N. Conci, N. Sebe, C. Bassetti, and R. Ferrario.
THE S-HOCK DATASET: ANALYZING CROWDS AT THE STADIUM.
CVPR, 2015. (submitted).

11. B. Zhang, **P.Rota**, N. Conci and N. Sebe.
HUMAN ACTIVITY RECOGNITION THROUGH MULTI-INSTANCE-LEARNING.
2015. (writing).

BIBLIOGRAPHY

[1] Viper toolkit. http://viper-toolkit.sourceforge.net/, 2014.

[2] J.K. Aggarwal and M.S. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, 2011.

[3] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*, 2007.

[4] T. Amabile. *Creativity in Context: Update to the Social Psychology of Creativity*. Westview Press, 1996.

[5] Mohamed R Amer, Dan Xie, Mingtian Zhao, Sinisa Todorovic, and Song-Chun Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *ECCV*. 2012.

[6] E. L. Andrade, S. Blunsden, and R. B. Fisher. Modelling crowd scenes for event detection. In *ICPR*, 2006.

[7] L.L.E.E. Balkwill, W.F. Thompson, and R. Matsunaga. Recognition of emotion in japanese, western, and hindustani music by japanese listeners1. *Japanese Psychological Research*, 46(4):337–349, 2004.

[8] O. Barinova, V. Lempitsky, and P. Kholi. On detection of multiple object instances using hough transforms. *PAMI*, 34(9):1773–1784, 2012.

[9] BEHAVE. Interactions test case scenarios. http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/index.html. 2007.

[10] T. Blaine and S. Fels. Contexts of collaborative musical experiences. In *Proceedings of the 2003 conference on New interfaces for musical expression*, pages 129–134. National University of Singapore, 2003.

[11] T. Blaine and T. Perkis. The jam-o-drum interactive music system: a study in interaction design. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 165–173. ACM, 2000.

[12] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *ICCV*, 2005.

[13] H. Blumer. Collective behavior. in: A. mcclung lee (ed.) principles of sociology. barnes & noble. In *New Outline of the Principle of Sociology*. 1951.

[14] Jean-Yves Bouguet. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation*, 5, 2001.

[15] Richard E Boyatzis. *Transforming qualitative information: Thematic analysis and code development*. Sage, 1998.

[16] G.R. Bradski. Computer vision face tracking for use in a perceptual user interface. 1998.

[17] F.P. Brooks, AL Hopkins, P.G. Neumann, and WV Wright. An experiment in musical composition. *Electronic Computers, IRE Transactions on*, (3):175–182, 1957.

[18] Simone Calderara and Rita Cucchiara. Understanding dyadic interactions applying proxemic theory on videosurveillance trajectories. In *CVPR Workshop*, 2012.

[19] A. Camurri, I. Lagerlöf, and G. Volpe. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International journal of human-computer studies*, 59(1):213–225, 2003.

[20] A. Camurri, G. Varni, and G. Volpe. Towards analysis of expressive gesture in groups of users: computational models of expressive social interaction. *Gesture in Embodied Communication and Human-Computer Interaction*, pages 122–133, 2010.

[21] A. Camurri, G. Volpe, G. De Poli, and M. Leman. Communicating expressiveness and affect in multimodal interactive systems. *Multimedia, IEEE*, 12(1):43–53, 2005.

[22] Antonio Camurri, C. Canepa, and G. Volpe. Active listening to a virtual orchestra through an expressive gestural interface: The Orchestra Explorer. In *Proceedings of the 7th international conference on New interfaces for musical expression*, pages 56–61. ACM, 2007.

[23] Antonio Camurri, Shuji Hashimoto, Matteo Ricchetti, Andrea Ricci, Kenji Suzuki, and Shuji Hashimotot. EyesWeb : Toward Gesture Music Systems. *Computer Music Journal*, 24(1):57–69, 2013.

[24] CAVIAR. Test case scenario. `http://groups.inf.ed.ac.uk/vision/CAVIAR/`. 2004.

[25] A. B. Chan and N. Vasconcelos. Bayesian poisson regression for crowd counting. In *ICCV*, 2009.

[26] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[27] Ming-Ching Chang, Nils Krahnstoever, Sernam Lim, and Ting Yu. Group level activity recognition in crowded environments across multiple cameras. In *AVSS*, 2010.

[28] Shen-Chi Chen, Chia-Hsiang Wu, Shih-Yao Lin, and Yi-Ping Hung. 2d face alignment and pose estimation based on 3d facial models. In *ICME*, 2012.

[29] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *ICCV Workshops*, 2009.

[30] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning context for collective activity recognition. In *CVPR*, 2011.

[31] D. Cope. *Computer models of musical creativity*. 2005.

[32] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of f-formations. In *Proceedings of British Machine Vision Conference*, 2011.

[33] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, and V. Murino. Towards computational proxemics: Inferring social relations from interpersonal distances. In *SocialCom*, pages 290–297, 2011.

[34] X. Cui, Q. Liu, M. Gao, and D.N. Metaxas. Abnormal detection using interaction energy potentials. In *CVPR*, pages 3161–3167, 2011.

[35] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[36] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[37] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006.

[38] David Doermann and David Mihalcik. Tools and techniques for video performance evaluation. 2000.

[39] Piotr Dollar, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *PAMI*, 36(8):1532–1545, 2014.

[40] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *CVPR*, 2009.

[41] 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.

[42] E. Eriksson, T.R. Hansen, and A. Lykke-Olesen. Movement-based interaction in camera spaces: a conceptual framework. *Personal and Ubiquitous Computing*, 11(8):621–632, 2007.

[43] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis*. Springer, 2003.

[44] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *CVPR*, pages 1226–1233. IEEE, 2012.

[45] 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.

[46] Mario A T Figueiredo and A.K. Jain. Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):381–396, 2002.

[47] T. Fritz, S. Jentschke, N. Gosselin, D. Sammler, I. Peretz, R. Turner, A.D. Friederici, and S. Koelsch. Universal recognition of three basic emotions in music. *Current Biology*, 19(7):573–576, 2009.

[48] A. Gartland-Jones and P. Copley. The suitability of genetic algorithms for musical composition. *Contemporary Music Review*, 22(3):43–55, 2003.

[49] Erving Goffman. *Encounters: Two studies in the sociology of interaction.* 1961.

[50] Erving Goffman. *Behaviour in Public Places*. 1963.

[51] E. Goode. *Collective behavior*. 1992.

[52] D. Gowsikhaa, S. Abirami, and R. Baskaran. Automated human behavior analysis from surveillance videos: a survey. *Artificial Intelligence Review*, 42(4):747–765, 2014.

[53] S. Greenberg, N. Marquardt, T. Ballendat, R. Diaz-Marino, and M. Wang. Proxemic interactions: the new ubicomp? *interactions*, 18(1):42–50, 2011.

[54] E.T. Hall. *The hidden dimension*, volume 6. Doubleday New York, 1966.

[55] E.T. Hall. *The silent language.* Anchor, 1973.

[56] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *CVPR Workshop*, 2012.

[57] L. Hiller and L. Isaacson. Musical composition with a high-speed digital computer, machine models of music, 1992.

[58] Minh Hoai and Andrew Zisserman. Talking heads: Detecting humans and recognizing their interactions. 2014.

[59] S. Hongeng, R. Nevatia, and F. Bremond. Video-based event recognition: activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding*, 96(2):129–162, 2004.

[60] De-An Huang and Kris M Kitani. Action-reaction: Forecasting the dynamics of human interaction. In *Computer Vision–ECCV 2014*. 2014.

[61] J.C.S. Jacques, Junior, S. Raupp Musse, and C.R. Jung. Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine*, 27(5):66–77, 2010.

[62] HC Jetter, Florian Geyer, Tobias Schwarz, and Harald Reiterer. Blended Interaction–Toward a Framework for the Design of Interactive Spaces. In *Proceedings of Workshop on Designing Collaborative Interactive Spaces AVI2012*, 2012.

[63] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *PAMI*, 35(1):221–231, 2013.

[64] S. Jordà, G. Geiger, M. Alonso, and M. Kaltenbrunner. The reactable: exploring the synergy between live music performance and tabletop tangible interfaces. In *Proceedings of the 1st international conference on Tangible and embedded interaction*, pages 139–146. ACM, 2007.

[65] P.N. Juslin and J. Sloboda. *Handbook of Music and Emotion: Theory, Research, Applications: Theory, Research, Applications*. OUP Oxford, 2009.

[66] P.N. Juslin, J.A. Sloboda, et al. *Music and emotion*, volume 315. Oxford University Press Oxford, 2001.

[67] P. KaewTraKulPong and R. Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. In *Proc. European Workshop Advanced Video Based Surveillance Systems*, volume 1. Citeseer, 2001.

[68] Ira Kemelmacher-Shlizerman and Ronen Basri. 3d face reconstruction from a single image using a single reference face shape. *PAMI*, 33(2):394–405, 2011.

[69] Julian FP Kooij, Gwenn Englebienne, and Dariu M Gavrila. A non-parametric hierarchical model to discover behavior dynamics from tracks. In *ECCV*. 2012.

[70] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, 2009.

[71] L. Kratz and K. Nishino. Tracking with local spatio-temporal motion patterns in extremely crowded scenes. In *CVPR*, 2010.

[72] Barbara Krausz and Christian Bauckhage. Loveparade 2010: Automatic video analysis of a crowd disaster. *CVIU*, 116(3):307–319, 2012.

[73] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012.

[74] Tian Lan, Leonid Sigal, and Greg Mori. Social roles in hierarchical models for human activity recognition. In *CVPR*, 2012.

[75] Tian Lan, Yang Wang, Weilong Yang, Stephen N Robinovitch, and Greg Mori. Discriminative latent models for recognizing contextual group activities. *PAMI*, 2012.

[76] I. Laptev. On space-time interest points. *IJCV*, 64(2-3):107–123, 2005.

[77] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, 2009.

[78] R. Legaspi, Y. Hashimoto, K. Moriyama, S. Kurihara, and M. Numao. Music compositional intelligence with an affective flavor. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 216–224. ACM, 2007.

[79] B. Lepri, N. Mana, A. Cappelletti, F. Pianesi, and M. Zancanaro. What is happening now? detection of activities of daily living from simple visual features. *Personal and Ubiquitous Computing*, 14(8):749–766, 2010.

[80] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, 2010.

[81] M. Mancini, G. Castellano, C. Peters, and P. McOwan. Evaluating the communication of emotion via expressive gesture copying behaviour in an embodied humanoid agent. *Affective Computing and Intelligent Interaction*, pages 215–224, 2011.

[82] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *CVPR*, 2009.

[83] Clark McPhail. *The myth of the madding crowd*. 1991.

[84] Ramin Mehran, BrianE. Moore, and Mubarak Shah. A streakline representation of flow in crowded scenes. In *ECCV*. 2010.

[85] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009.

[86] L.B. Meyer. *Emotion and meaning in music*. University of Chicago Press, 1956.

[87] E.R. Miranda and J.A. Biles. Evolutionary computer music. 2007.

[88] F. Morreale, R. Masu, and A. De Angeli. Robin: An algorithmic composer for interactive scenarios. In *Sound and Music Computing*, 2013.

[89] Fabio Morreale, Raul Masu, Antonella De Angeli, and Patrizio Fava. The Effect of Expertise in Evaluating Emotions in Music. *Proceedings of the 3rd International Conference on Music & Emotion*, 2013.

[90] Jeho Nam, Masoud Alghoniemy, and Ahmed H Tewfik. Audio-visual content-based violent scene characterization. In *ICIP*, 1998.

[91] G. Nierhaus. *Algorithmic composition: paradigms of automated music generation*. Springer Verlag Wien, 2009.

[92] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. In *Computer Analysis of Images and Patterns*, 2011.

[93] A.P. Oliveira and A. Cardoso. A musical system for emotional expression. *Knowledge-Based Systems*, 23(8):901–913, 2010.

[94] Nuria M Oliver, Barbara Rosario, and Alex P Pentland. A bayesian computer vision system for modeling human interactions. *PAMI*, 2000.

[95] M.S. O'Modhrain and C. Adviser-Chafe. *Playing by feel: incorporating haptic feedback into computer-based musical instruments*. Stanford University, 2001.

[96] Javier Orozco, Shaogang Gong, and Tao Xiang. Head pose classification in crowded scenes. In *BMVC*, 2009.

[97] M. Jones P. Viola. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[98] F Pachet. The continuator: Musical interaction with style. *Journal of New Music Research*, 2003.

[99] W. Pan, W. Dong, M. Cebrian, T. Kim, J. H. Fowler, and A. S. Pentland. Modeling dynamical influence in human interaction: Using data to make better inferences about influence within social systems. *Signal Processing Magazine, IEEE*, 29(2):77–86, 2012.

[100] Alonso Patron-Perez, Marcin Marszalek, Ian Reid, and Andrew Zisserman. Structured learning of human interactions in tv shows. *PAMI*, 2012.

[101] Alonso Patron-Perez, Marcin Marszalek, Andrew Zisserman, and Ian Reid. High five: Recognising human interactions in tv shows. *BMVC*, 2010.

[102] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, pages 261–268, 2009.

[103] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. *ECCV*, pages 452–465, 2010.

[104] Hamed Pirsiavash, Deva Ramanan, and Charless C Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 2011.

[105] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

[106] Karthir Prabhakar, Sangmin Oh, Ping Wang, Gregory D Abowd, and James M Rehg. Temporal causality for the analysis of visual events. In *CVPR*, 2010.

[107] Karthir Prabhakar and James M Rehg. Categorizing turn-taking interactions. In *ECCV*. 2012.

[108] G.M. Rader. A method for composing simple traditional music by computer. *Communications of the ACM*, 17(11):631–638, 1974.

[109] 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.

[110] et al. Rasamimanana, N. The urban musical game: using sport balls as musical interfaces. In *CHI Extended Abstracts*, pages 1027–1030. ACM, 2012.

[111] N. Robertson and I. Reid. Behaviour understanding in video: a combined method. In *ICCV*, volume 1, 2005.

[112] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert. Density-aware person detection and tracking in crowds. In *ICCV*, 2011.

[113] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert. Data-driven crowd analysis in videos. In *ICCV*, 2011.

[114] Paolo Rota, Nicola Conci, and Nicu Sebe. Real time detection of social interactions in surveillance video. In *ECCV Workshops and Demonstrations*, 2012.

[115] R. Rowe. Interactive music systems: Machine listening and composition, 1993.

[116] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, 2009.

[117] M. S. Ryoo and J. K. Aggarwal. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). `http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html`, 2010.

[118] M. San Biagio, M. Crocco, M. Cristani, and V. Martelli, S. andMurino. Heterogeneous auto-similarities of characteristics (hasc): Exploiting relational information for classification. In *ICCV*, 2013.

[119] P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. *IJCV*, 80(1):72–91, 2008.

[120] Michael Sapienza, Fabio Cuzzolin, and Philip Torr. Learning discriminative space-time actions from weakly labelled videos. *IJCV*, 2012.

[121] R.K. Sawyer. *Explaining Creativity:The Science of Human Innovation: The Science of Human Innovation*. Oxford University Press, USA, 2012.

[122] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *ICPR*, 2004.

[123] P. Scovanner and MF Tappen. Learning pedestrian dynamics from the real world. In *ICCV*, pages 381–388, 2009.

[124] Jianbo Shi and Carlo Tomasi. Good features to track. In *CVPR*, 1994.

[125] I. Simon, D. Morris, and S. Basu. Mysong: automatic accompaniment generation for vocal melodies. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 725–734. ACM, 2008.

[126] B. Solmaz, B. E. Moore, and M. Shah. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *PAMI, IEEE Transactions on*, 34(10):2064–2070, 2012.

[127] M.J. Steedman. A generative grammar for jazz chord sequences. *Music Perception*, pages 52–77, 1984.

[128] Joshua M Susskind, Geoffrey E Hinton, Javier R Movellan, and Adam K Anderson. Generating facial expressions with deep belief nets. *Affective Computing*, pages 421–440, 2008.

[129] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[130] Amir Tamrakar, Saad Ali, Qian Yu, Jingen Liu, Omar Javed, Ajay Divakaran, Hui Cheng, and Harpreet Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *CVPR*, 2012.

[131] Schofield G. Taylor, R. and J. Shearer. esigning from within: humanaquarium. In *Proceedings of CHI 2011*, pages 725–734. ACM, 2011.

[132] P.M. Todd and G.M. Werner. Frankensteinian methods for evolutionary music. *Musical networks: parallel distributed perception and performace*, pages 313–340, 1999.

[133] Diego Tosato, Mauro Spera, Marco Cristani, and Vittorio Murino. Characterizing humans on riemannian manifolds. *PAMI*, 35(8):1972–1984, 2013.

[134] H. Ullah and N. Conci. Crowd motion segmentation and anomaly detection vis multi-label optimization. In *ICPR workshop on Pattern Recognition and Crowd Analysis*, 2012.

[135] K. van Boerdonk, R. Tieben, S. Klooster, and E. van den Hoven. Contact through canvas: an entertaining encounter. *Personal and Ubiquitous Computing*, 13(8):551–567, 2009.

[136] G. Varni, M. Mancini, G. Volpe, and A. Camurri. Sync'n'move: social interaction based on music and gesture. *User Centric Media*, pages 31–38, 2010.

[137] I. Wallis, T. Ingalls, and E. Campana. Computer-generating emotional music: The design of an affective music algorithm. *DAFx-08, Espoo, Finland*, pages 7–12, 2008.

[138] I. Wallis, T. Ingalls, E. Campana, and J. Goodman. A rule-based generative music system controlled by desired valence and arousal, 2011.

[139] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[140] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

[141] A. Wiethoff and S. Gehring. Designing interaction with media façades: a case study. In *Proceedings of the Designing Interactive Systems Conference*, pages 308–317. ACM, 2012.

[142] G. Wiggins, G. Papadopoulos, S. Phon-Amnuaisuk, and A. Tuson. Evolutionary methods for musical composition. *International Journal of Computing Anticipatory Systems*, 1999.

[143] Bo Wu and Ram Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *IJCV*, 75(2):247–266, 2007.

[144] G. Zen, B. Lepri, E. Ricci, and O. Lanz. Space speaks: towards socially and personality aware visual surveillance. In *MPVA'10*. ACM, 2010.

[145] Hua Zhong, Jianbo Shi, and M. Visontai. Detecting unusual activity in video. In *CVPR*, 2004.

[146] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.