Grenoble INP – Ensimag
École Nationale Supérieure d'Informatique et de Mathématiques Appliquées

Final year project report

Performed at Okayama University

Human behavior analysis
Analysis of social relation and interaction in pedestrian groups

Adrien Gregorj
3rd year – Option MMIS Bio

19th February, 2018 – 20th July, 2018

Okayama University
1-1, Tsushima-Naka, 1-Chome
Okayama 700-8530
Japan

Supervisors
Akito Monden
Zeynep Yücel
School Tutor
James Crowley

Contents

1 Context

This internship takes place in an academic context. The home structure is the Graduate School of Natural Science and Technology of Okayama University. In particular, the host laboratory is part of the Division of Industrial Innovation Sciences, in the Department of Computer Science.

Okayama University is a well-ranked university in Japan located in Okayama Prefecture, in the Chūgoku region of the main island of Honshū. The school was founded in 1870 and welcomes around 14000 students (10000 undergraduates, 3000 postgraduates and 1000 doctoral students). It's motto is « Creating and fostering higher knowledge and wisdom ».

Okayama University and Université Grenoble Alpes have a long history of exchange that has been reactivated for the past three years with around a dozen of Grenoble students performing research oriented internship in Okayama every year. As a University Research Administrator, Dr Bernard Chenevier's mission is to develop research collaborations at the international level and, for this purpose, he works to strengthen the bond between these two universities. It is in this context that internship topics are proposed to French students.

The Graduate School of Natural Science and Technology was originally established in April 1987. It focuses on research in fundamental sciences such as global climate change, plant photosynthesis or supernova neutrinos. The Division of Industrial Innovation Sciences works especially on applied engineering in the field of computer science, robotics, material sciences among others.

In the Department of Computer Science, the topics of research are the basic theory and application of information technology, artificial intelligence and computer technology. Examples of research projects include the development of a visualization tool for a processor or human tracking algorithm by means of attention control. Dr Akito Monden is the professor of the Theory of Programming and Artificial Intelligence laboratory, where this internship takes place and Dr Zeynep Yücel is the assistant professor in the same laboratory.

2 Motivation and objectives

Dr Zeynep Yücel has been working for a few years on modeling crowd movements [1][2][3]. This study is an extension to their work in the last couple of years.

The motivations behind constructing crowd or pedestrian motion models are numerous: testing architectural designs for an evacuation process, designing accurate crowd models for movies or video games, etc., and it is a very active field of research as it is requisite in recent domains such as autonomous driving.

One of the main challenges behind pedestrian motion modelling is to take into account the social relation and interactions between social pedestrian groups. Indeed, [3] has demonstrated that the motion patterns of individuals performing gestures (i.e. physical contact, mutual gaze, conversation and arm gestures) are significantly different from the ones of not-interacting individuals. In addition, [4] proved that pedestrian pairs, which are in different social relations, have different motion characteristics. Both [3] and [4] rely on human annotations to model the different motion patterns and evaluate their performance.

Therefore, this project aims to develop methods to automatize the recognition and classification process of pedestrians'. Namely, our goals are twofold: (i) recognition of social relations and (ii) detection of gestures. We study each of these problems in Section 4 and Section 5. For recognition of social relations we devise a probabilistic method with a Bayesian approach, whereas for detecting the gestures of interaction we propose a method inspired from audio signal processing.

Thanks to the methods proposed in Sec 4 and 5, instead of relying on human referees to decide whether or not a set of pedestrians are presumably in a given relationship and interacting (performing gestures) in a given video, this decision can be made automatically, which enables real time processing and reaction of social interaction robots.

3 Background and related work

This section will detail the previous works that have been done in the different domains that this project involves.

3.1 Crowd modeling

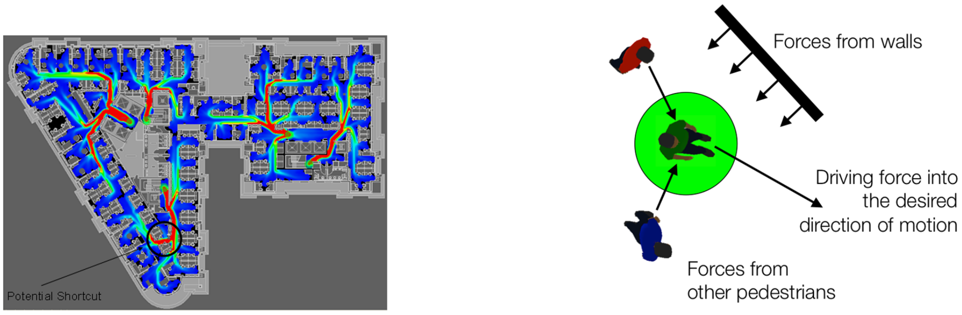
When modeling the crowd, two kinds of models are generally in use:

Macroscopic models: This kind of models [5, 6, 7] is used to describe often high density crowds. They are generally based on dynamic fluid representation of the people and study the density and flow of pedestrians in simulated environments (see Figure 1a). Their applications include planing evacuation processes or large scale events for avoiding stampedes by identifying high density areas.

Microscopic models: These models, where the most prominent one is Social Force Model [8], describe pedestrians as particles with a certain position and velocity and subject to a variety of forces: attractive forces from other particles of the same group, repulsive forces from obstacles like walls or other pedestrians, etc (see Figure 1b). They are used in modeling the interaction at a smaller scale and can be used to develop autonomous agents such as companion robots or smart wheelchairs.

This study focuses on motion of pedestrians in small groups. The most popular model for this topic is the Social Force model (SFM). Many studies based their models on the original SFM and extended it to enhance the simulated behaviors of pedestrians to prevent sudden changes in velocity or direction [11, 12].

Most recently, data driven techniques such as neural networks have been used to compute socially acceptable trajectories [13, 14]. Mainly, networks are trained on real trajectory data to generate missing portions considering the beginning of a trajectory.


Potential Shortcut
Forces from walls
Driving force into the desired direction of motion
Forces from other pedestrians

(a) (b)

Figure 1: (1a) Macroscopic [9] and (1b) microscopic [10] crowd models illustrations.

3.2 Gait analysis

Motion patterns of pedestrians are significantly affected by gait. Length and frequency of steps of walking individuals have been particularly studied to inventory empirical values for panels of people in various environments.

In [15], Hui *et al.* manually measured the gait parameters of pedestrians to study their correlation with gender and age. A particularly vast number of studies have investigated the impact of age on gait parameters [16, 17, 18, 19, 20]. The recurring finding is that older people exhibit slower walking speed and shorter step length than younger ones. Sun *et al.* analysed the impact of surface slope on human gait and discovered that pedestrians' step length decrease during ramp descent [21].

The uniqueness of human gait rythme has been established in various studies [22, 23]. This specificity was found to be a discriminating feature to build pedestrian detectors [24][25] or differentiate pedestrians from other moving vehicles [23]. Moreover, gait analysis has also been used as a way to identify people's gender or age range [26][27].

Few studies present automatized methods to measure the gait parameters of pedestrians. In [28], Niyogi *et al.* detect the characteristic braided pattern of walking in the spatiotemporal volume by computing an autocorrelation sequence of trajectory values (the inverse Fourier transform of the magnitude of the Fourier transform) and finding the peak in this sequence. Saunier *et al.* [29] and Hediyeh *et al.* [30] compute power spectra of speed profiles to extract gait frequency of pedestrians.

3.3 Action recognition

The problem of action detection and recognition has been addressed in many ways as it is at the center of various other topics such as video surveillance or human-computer interaction. Initial methods used dense trajectories, where feature points are tracked across frames and descriptors such as Fisher Vectors are computed to capture the motion information of the subjects in the video [31].

More recently, machine learning algorithms, such as neural networks, have been widely applied to this task. Skeleton-based action recognition using Recurrent Neural Networks

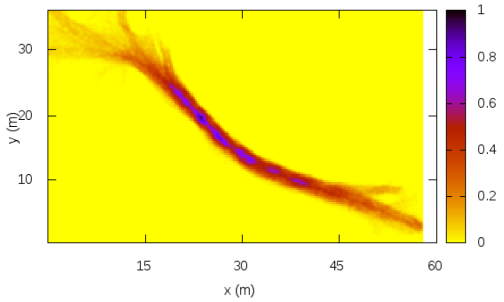


Figure 2: Density map of the environment.

(RNNs) [32, 33] or their extension using Long Short-Term Memory (LSTMs) units [34] has shown significant improvements over the previous methods.

Recently introduced Graph Convolutional Networks (GCNs) [35] generalize the convolution operation on 1D and 2D arrays to graph data structure. In [36], Yan *et al.* propose a Spatio-Temporal Graph Convolutional Network (STGCN) to classify actions in videos. Their study relies on building a graph composed of the skeletons of a given person at multiple successive frames, each joint being linked to its detection on the previous and following frame. Their classifier takes this graph as input to perform action classification.

4 Recognition of social relations

This section focuses on the impact of social relations on pedestrian motion. The definition of the relationships is based on [37], where Bugental proposes a domain-based approach and divides social life into five non-overlapping domains as attachment (*e.g.* family), hierarchical power, mating (*e.g.* couple), reciprocity (*e.g.* friends) and coalitional (*e.g.* colleagues).

In this study, hierarchical relation is eliminated, since it does not apply to the full extent to pedestrians in a public space. For this simplified case, we contain ourselves to the two most distinct social relations as pointed by [4]: mating and coalitional.

4.1 Dataset

The dataset used for studying the impact of social relations on walking patterns was introduced in [38]. This dataset was obtained by setting up a tracking environment in the "ATC" shopping center in Osaka. The system was composed of multiple 3D range sensors, covering an area of about 900 m2. It is composed of a set of trajectory data files containing the coordinates of the pedestrians, their height as well as their velocities.

For annotation purposes, the tracking area was recorded with 16 cameras. A subset of the footage was annotated by human coders to identify the pedestrian social groups and relation (apparent purpose of the visit, apparent gender, apparent relation, apparent age) for 988 dyads. The relation between dyads is distributed as follows: 358 coalitional, 96

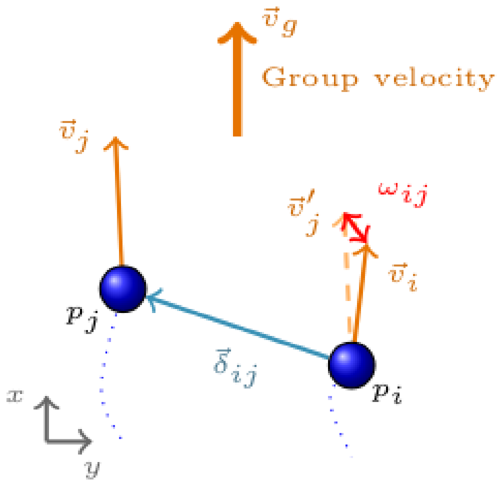


Figure 3: Schema of the observables depicted on a single dyad.

mating, 216 attachment and 318 reciprocal.

4.2 Observables and empirical distributions

The raw trajectory files are exposed to a preprocessing operation. Namely, the data points, which are not uniform over time, are averaged over intervals of 0.5 s to obtain uniform sampling, and to decrease the effect of measurement noise and pedestrian gait. Figure 2 shows the cumulative density map of the environment.

From the preprocessed data, a certain number of observables are computed for each dyad at each sampling instant according to the *group reference frame*. Namely, we adopt a dynamic reference frame such that its x-axis is always aligned with the direction of motion. This enables eliminating the effect of maneuvering for taking curves or avoiding obstacles.

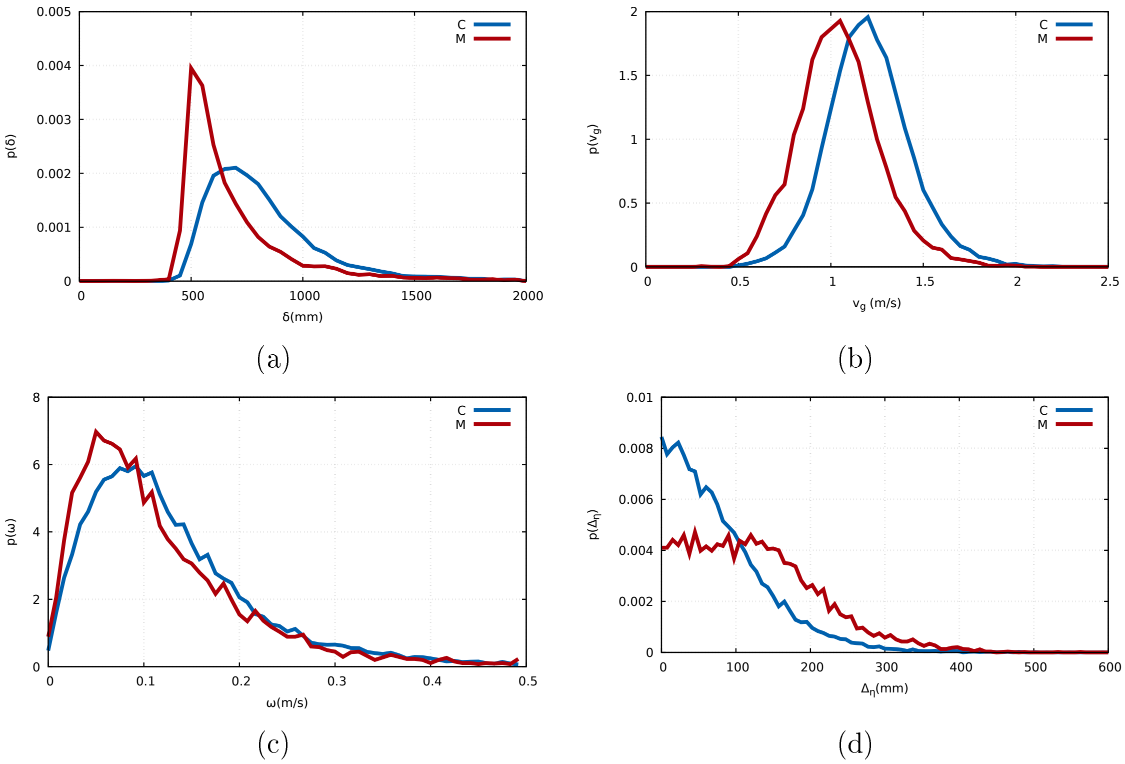
The observables, which are illustrated on Figure 3, are defined explicitly as follows:

1. *Interpersonal distance*, δ_{ij}, is defined as the distance between the peers.

2. *Group velocity*, v_g, is the magnitude of the average instantaneous velocities of the peers,
$$v_{g(i,j)} = \left| \frac{\vec{v}_i + \vec{v}_j}{2} \right| .$$

3. *Absolute difference of velocity vectors*, ω_{ij}, is defined as the magnitude of the difference vector,
$$\omega_{ij} = |\vec{v}_i - \vec{v}_j| .$$

4. *Height difference*, is
$$\Delta_\eta = |\eta_i - \eta_j|,$$
where η_i and η_j stand for the height of the pedestrians p_i and p_j, respectively.



(a)

(b)

(c)

(d)

Figure 4: Empirical distributions of (4a) δ, (4b) v_g, (4c) ω and (4d) Δ_η, of dyads in coalitional (C, blue) and mating (M, red) relation.

After computing the observables, their probability density distributions are approximated with cumulative histograms (see Figure 4). It is clear that the distributions are significantly different. Moreover, an ANOVA analysis confirmed that the observables δ, v_g, ω and Δ_η have a p-value smaller than 10^{-4}. Considering that in the literature ([7]), a value lower than 0.05 is generally admitted to demonstrate statistical significance, it can be concluded that coalitional relation and mating relation are considerably different in terms of the observables.

4.3 Bayesian model

Following the findings of [4], the idea is to build a predictive model based on the *a priori* knowledge of the observable distributions to classify dyads into their apparent social relations as coalitional or mating.

Let $\Sigma(t) = \{\delta, v_g, \omega, \Delta_\eta\}$ be the set of observables at time t. The probability that $\Sigma(t)$ comes from a group in social relation of r, where $r \in \{C, M\}$ can be computed as

$$P_t(r|\Sigma) = \frac{P_t(\Sigma|r)P_t(r)}{P_t(\Sigma)}. \tag{1}$$

The term $P_t(\Sigma|r)$ can be decomposed as the product of probabilities of each single observable, provided that their independence can be assumed. To verify this hypothesis, the Jaccard distance is computed.

Let Θ and Δ be two random variables. Their Jaccard distance is defined as

$$D(\Theta, \Delta) = \frac{H(\Theta, \Delta) - I(\Theta, \Delta)}{H(\Theta, \Delta)}, \tag{2}$$

where $H(\Theta, \Delta)$ and $I(\Theta, \Delta)$ are the joint entropy and mutual information of the variables, respectively, and are described as

$$H(\Theta, \Delta) = -\sum_{i,j} p(\theta_i, \delta_j) \log_2(p(\theta_i, \delta_j)), \tag{3}$$

$$I(\Theta, \Delta) = \sum_{i,j} p(\theta_i, \delta_j) \log_2 \left(\frac{p(\theta_i, \delta_j)}{p(\theta_i)p(\delta_j)} \right). \tag{4}$$

The distance $D(\Theta, \Delta)$ should be 1 for uncorrelated variables and closer to 0 for correlated ones.

Table 1 shows Jaccard distance values between pairs of observables for two sample subsets of dyads coming from coalitional and mating relations. The results being always higher than 0.94, the independence hypothesis can be considered as reasonable.

Therefore, the term $P_t(\Sigma|r)$ can be decomposed as follows,

$$P_t(\Sigma|r) = P_t(\delta|r)P_t(v_g|r)P_t(\omega|r)P_t(\Delta_\eta|r). \tag{5}$$

Table 1: Jaccard distance between observables for subsets of (1a) coalitional and (1b) mating relations.

(a)

	v_g	ω	δ	Δ_η
v_g	0	0.98	0.99	0.98
ω	-	0	0.98	0.97
δ	-	-	0	0.98
Δ_η	-	-	-	0

(b)

	v_g	ω	δ	Δ_η
v_g	0	0.96	0.97	0.96
ω	-	0	0.95	0.94
δ	-	-	0	0.96
Δ_η	-	-	-	0

Without any prior belief, the probability of being in a C or M relation are initialized with unbiased equal probabilities,

$$P_0(r) = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}. \tag{6}$$

As time elapses, we propose updating the prior as in Equation 7, where the parameter α defines the rate of update,

$$P_t(r) = \alpha P_0(r) + (1-\alpha)P_{t-1}(r). \tag{7}$$

In this manner, $P_t(C)$ and $P_t(M)$ are computed at every time instant t and the relation associated with the larger probability is considered as the estimated relation at that instant. For quantifying estimation performance, the mean rate of correct detections over all time instants for all dyads are reported.

4.4 Results

Table 2: Detection performance for varying α (in %).

α	C	M	Total
0	88.5 ± 2.1	73.4 ± 5.0	85.3
0.5	88.1 ± 2.1	79.1 ± 3.8	87.2
1	87.1 ± 2.3	81.3 ± 4.1	86.5

In practice, 30% of the pairs are randomly chosen and their trajectories are used to build the probability density functions in Eq. 5. The remaining 70% are used to test the estimation method. Moreover, repeating this validation procedure 20 times, the mean and standard deviations of performance values are computed to investigate the sensitivity (i.e. dependence) of the observables on training set. By randomly picking 30% of the entire samples and repeating this procedure 20 times, the probability that a particular sample is not used for training falls below 10^{-3} .

From the recognition rates presented in Table 2, it is observed that coalitional relation is recognized with a somewhat higher rate for all values of α, which could be due to the imbalance of samples in the dataset as given in Section 4.1. Moreover, although the overall performance rate for $\alpha = 1$ seems sligtly lower than that of $\alpha = 0$, taking a fixed and unbiased prior is regarded to perform better in the sense that it yields a more balanced estimation performance for the social relations. In addition, due to the very low standard deviation values, the effect of random shuffling is regarded to be minute, which suggests that the observables are stable across samples and the method is resilient to changes in training set.

It is important to note that the algorithm relies on previous annotations made by coders for the training step (construction of the distributions). Yet it proves to be able to generalize the classification process to new data.

4.5 Alternative methods

The proposed method detailed in Section 4 considers each set of observations $\Sigma(t) = \{\delta, v_g, \omega, \Delta_\eta\}$ collected at every sampling instant t and yields a probability of being in a C or M relation. In that sense, it enables a *local* decision, which may gradually evolve.

In this section, we describe a few alternative methods, which work on the entire set of observations for all time instants providing *global* decisions. In other words, for a particular dyad of interest, we obtain distributions of each observable at the end of its trajectory and compare them to the corresponding representative distributions of C and M relations. At this point, we need to measure the *distance* between each pair of probability distribution on a statistical manifold. Therefore, it is important to define proper distance metrics.

In what follows, we define the metrics that we considered. Let two discrete probability distributions be denoted with P and Q,

1. *Kullback-Leibler (KL) divergence* from P to Q is defined as,

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \tag{8}$$

   Equation 8 can be interpreted as the amount of information lost when Q is used to approximate P. Obviously, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. Therefore, in order to have a fair comparison we compute the divergence from the observed distribution to the representative distribution of C and M. In addition, the logarithmic term is likely to suffer from the zero probabilities, if the observation is not sufficiently long. We tackle this issue simply by ignoring the terms with $P(i) = 0$ and $Q(i) = 0$.

2. *Jensen-Shanonn (JS) divergence*, is derived from KL divergence as follows,

$$D_{JS}(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M), \tag{9}$$

where $M = \frac{1}{2}(P + Q)$, and it has the advantage to be symmetric. Therefore, it does not matter whether we compute $D_{JS}(P||Q)$ or $D_{JS}(Q||P)$, but the issue relating the zeros in the logarithmic term pertain.

3. *Earth mover's distance* (EMD) is defined as,

$$EMD(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} d_{i,j}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j}}, \tag{10}$$

where $f_{i,j}$ is the flow between p_i and q_j and $d_{i,j}$ is the ground distance between p_i and q_j. If the distributions are pictured as pile of dirt, the EMD represents the minimum cost to turn one pile into the other (the cost corresponding to the amount of dirt moved times the distance by which it is moved).

4. *Log likelihood* follows,

$$LL(P \sim Q) = -\sum_i p_i \log(q_i). \tag{11}$$

It represents the plausibility that the histogram of P was sampled from a distribution Q. If Q has any zero values we skip those observations, since they make the entire term undefined.

To create a decision algorithm, we compute the previous metrics for each observables. For the Kullback-Leibler divergence, the Jensen-Shanonn divergence and the EMD, we select the relation that minimize these distances. Conversely, for the Log Likelihood, we select the relation that maximize it.

Table 3: Detection performance for the other proposed metrics.

Metric	v_g	ω	δ	Δ_η
Proposed ($\alpha = 1$)	76.6	68.1	77.8	76.4
KL	77.8	37.1	67.7	81.2
JS	77.2	63.4	75.8	77.4
EMD	76.0	63.6	67.6	76.2
LL	76.8	66.3	76.3	74.8

The performance obtained using these alternative methods are detailed in Table 3 (total accuracy for both classes), together with a comparison to the proposed method for same observables. For the sake of brevity, we skip reporting detailed detection performance for each class but we refer the reader to the appendix for these results.

Table 3 shows that the proposed metric has comparable or better performance to the alternative methods. A closer look into the individual detection rates reveal that the alternative methods may sometimes yield a high detection rate due to a bias to sort most observations as C, which has many more samples than M, and thus misleadingly increases the total performance.

In addition, one important inference drawn from Table 3 is that blending all observables into a single decision strategy improves the performance of the proposed method. It may be worthwhile, to test whether this integration may help the alternative methods. However, our expectation is that D_{KL} and D_{JS} suffer more from the zeros in the logarithmic term, since such cases may arise more frequently (*i.e* from any one of the observables), reducing the available amount of data used in the decision process. In addition, LL may have problems concerning machine precision due to the growing number of decimal terms in the product.

4.6 Discussion

Using only trajectory and height information of pairs of pedestrians, we are able to estimate their social relation with over 80% accuracy. Considering the challenge of the problem, the proposed method is regarded to achieve significant accuracy.

To the best of our knowledge, this is the first study on detection of social relation of interacting pedestrians. Therefore, it is not possible to compare our performance to alternative methods. We can propose several improvements or extensions to the current study.

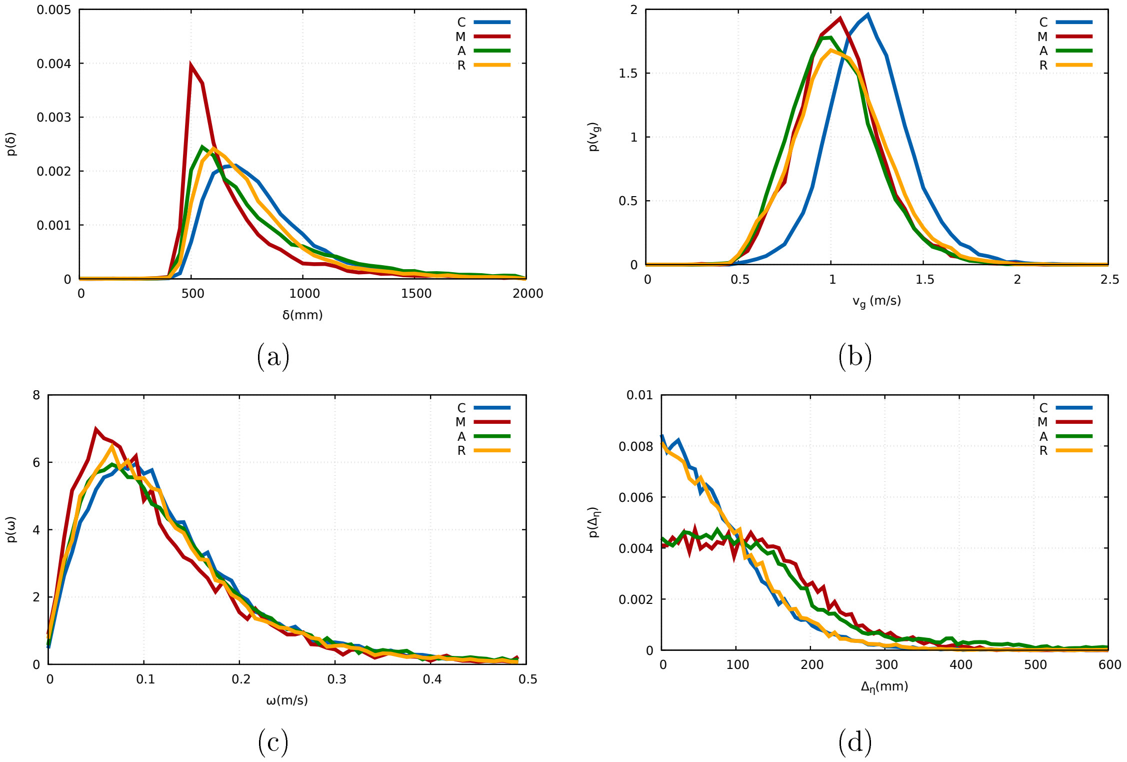
First of all, in this work we considered only the coalitional and mating relations but in the future we could extend this study to discriminate more challenging and less distinct relations such as attachment and reciprocity.

Moreover, other methods to classify the trajectories can be imagined. One promising idea could be to develop a neural network that takes the observables as input and returns the probability of each relation. One basic implementation could be done with a fully connected Multi-Layer Perceptron (a basic fully connected neural network) that takes a fix number of observables (from portions of the trajectory) to classify. However, LSTM neural networks would be more suited for this task, as they take in account the chronological component of the data.

5 Detection of gestures during interactions

In this section, we study detection of gestures of interaction. We first choose a public dataset with an abundant number of pedestrian dyads and annotate it for a set of gestures, as conversation, gaze exchange, arm gestures and physical contact. Arm gestures turn out to have the highest rate of inter-rater agreement, and we decided to develop a method for detecting them.

The idea is that human body is subject to somewhat regular oscillations due to the walking rhythm and any disruption of this regularity (around the wrist, elbow, shoulder



(a)

(b)

(c)

(d)

Figure 5: Empirical distribution of (5a) δ, (5b) v_g, (5c) ω and (5d) Δ_η, of peers for dyads in coalitional (C, blue), mating (M, red), attachment (A, green) and reciprocity (R,orange) relation.

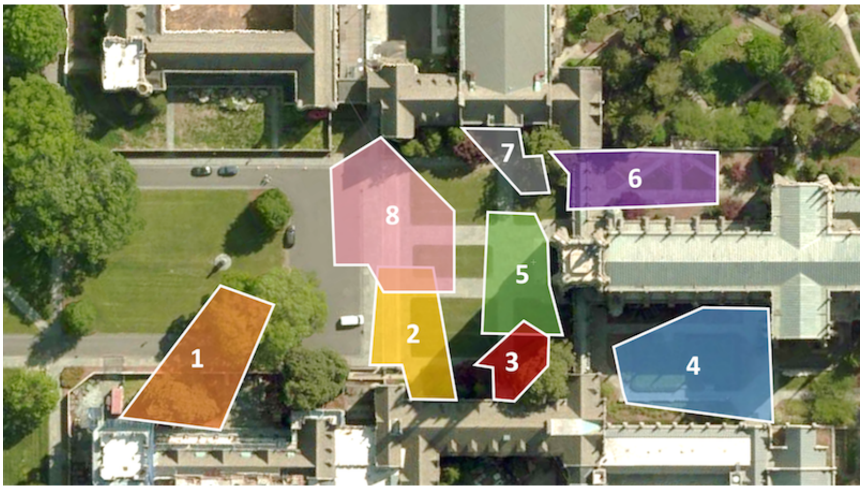


Figure 6: Top view of the Duke University campus with the fields of view of the cameras used to record the DukeMTMC dataset in color.

area) can be considered to arise from a gesture within a social dyad. We propose using a method inspired from pitch detection to pinpoint these disruptions.

5.1 Dataset

We started by searching for a dataset that could be exploited for the gesture detection task. Many video datasets involving pedestrians are publicly available but most of them are not suitable for studying gestures of pedestrian interaction. Some of them are based on bird's eye view camera footage, in which pedestrian interactions can not easily be observed (BIWI dataset [39]), whereas some datasets involve too sparse pedestrian traffic, too short footage or acted and unnatural behaviors (CAVIAR dataset [40], ETHZ [41], PETS 2009 [42]).

The DukeMTMC dataset, introduced in [43], is found to be the most appropriate set for studying gestures of interacting pedestrians. It presents pedestrian traffic filmed at multiple locations in the campus of the Duke University. Figure 6 shows a top view of the fields of view of each camera of the datasets. In addition to a very large quantity of data (85 minutes of 1080p and 60 fps video for 8 cameras, with more than 2,000 identities), it involves several ground truth values.

Trajectories on image coordinate frame and real-world coordinates, as well as the pose estimations of each pedestrian, are provided as ground truth. Trajectories on image plane are obtained by manual annotation. Namely, coders were asked to mark the feet position of pedestrians on certain key frames and the trajectory points between those were linearly interpolated. This gives piece-wise linear trajectories for each individual (see Figure 7). The real-world trajectories are computed using the homography matrices of each corresponding camera (that are also available).

In addition, a subset of this dataset (20 minutes of videos for the camera 1, 2, 4 and 5) was also annotated for groups: the DukeMTMC-Groups dataset, introduced in [44]. In

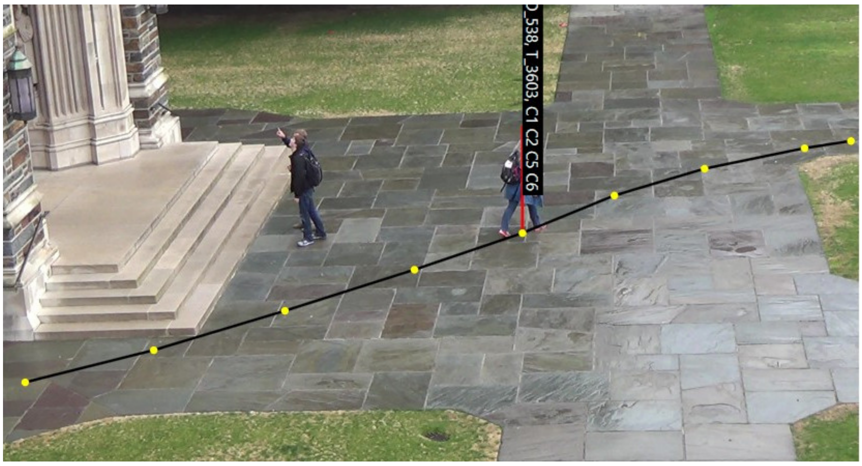


Figure 7: Example of annotations on the DukeMTMC dataset. Yellow dots were manually placed by coders between the feet of the pedestrians and black lines are linear interpolation between those keyframe annotations.

Table 4: Inter-rater agreement analysis

Gesture	Cohen's κ	Krippendorff's α
Speaking	0.45	0.44
Touching	0.49	0.48
Arms gesture	**0.77**	**0.77**
Gazing	0.49	0.48
Intensity	-	0.64

total, 64 groups are identified and tracked for an average of 400 frames per group.

In order to get the ground truth for gesture and intensity of interaction concerning DukeMTMC-Groups dataset, we carried out an additional annotation task. Two coders are asked to watch clips for each group on each camera and annotate the four kinds of gestures: conversation, gaze exchange, arm gestures, physical contact. In addition, they were asked to rank the intensity I of the interaction on a scale from 0 to 3, $I = 0$ being no interaction and $I = 3$ the maximum level of interaction.

We performed inter-rater agreement analysis using the Cohen's kappa coefficient (κ) and Krippendorff's alpha coefficient (α) [45]. The results are displayed in Table 4. Concerning Cohen's κ, values greater than 0.67 are often considered to be statistically reliable, and values at the level of 0.80 are regarded to indicate to a substantial agreement. Since the agreement is best for the arm gestures, we decided to focus on this particular interaction in our analysis. Regarding Krippendorff's α, there does not exist a benchmark value for significance but values around 0.66 are regarded as satisfactory [45].

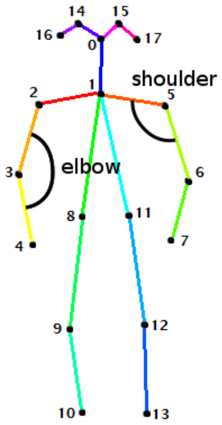


Figure 8: Schema of the OpenPose joints numbering with the elbow and shoulder angles.

5.2 Pose estimation

As mentioned in Section 5.1, pose estimation had already been performed by the research team of the DukeMTMC dataset and is publicly available.

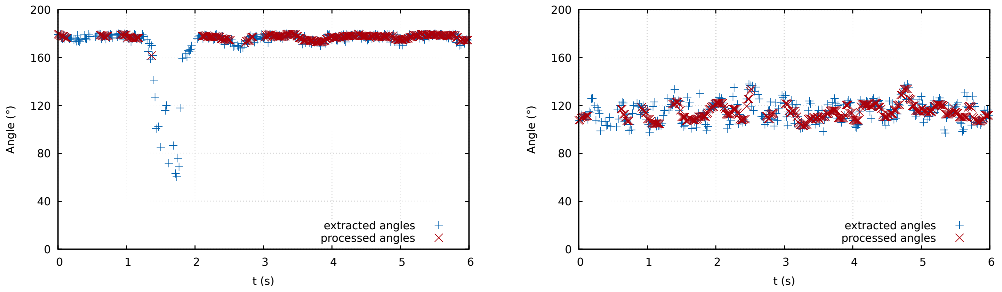
The pose estimations are obtained using the publicly available OpenPose library, which is state of the art for this problem [46]. To identify joints in every frame, Cao *et al.* employ body parts inference, which we briefly explain in what follows.

The model is mainly composed of a multi-stage Convolutional Neural Network that evaluates the confidence maps for body parts (18 different joints: right and left elbow, right and left feet, etc.) but also for what Cao *et al.* call « part affinity fields » (PAF), which correspond to vectors that encode the direction between the two joints for each pixel on a given limb. Multiple iterations enable gradual refinement of estimations so as to avoid false detections (*e.g.* right wrist instead of the left).

A graph can be modeled using the detected body parts as vertices and by scoring the edges using the part affinity fields with the outputs provided by the neural network. The integral of the PAF over the line that joins the two parts gives a cost that encodes the confidence of the joints belonging to the same limb. In order to retrieve actual skeletons from this graph, which is an NP Hard problem, the authors introduced a greedy algorithm that performs well and rapidly. They actually subdivided the problem to work on smaller bi-partite graphs (considering successively all the possible association of joints: right wrist - right elbow, right foot - right knee, etc.) instead of considering all the joints at once.

5.3 Angles extraction

We defined and computed two kinds of angles: elbow and shoulder angles (see Figure 8). Due to the occlusions or misdetections, there is the possibility of having some instability on these values. For eliminating this sort of inconsistencies, a median filtering (with a window of size 9) is carried on the elbow and shoulder angles. In addition, only those pedestrians



(a)

(b)

Figure 9: (9a) Elbow and (9a) shoulder angles before and after preprocessing.

with less than 20% of missing values along their trajectory are considered in the analysis.

Figure 9 shows elbow and shoulder angles for a pedestrian in a non-interacting dyad. It appears that the walking motion (swinging arms or just the overall oscillations due to the body movements) causes a certain periodicity in the signal, which we intend to identify.

After examining the empirical observations and the videos, we decided to focus on elbow angles in our analysis mainly for the two following reasons. Firstly, the region of the shoulder joint is inherently broader than the region of the elbow, which makes estimations of OpenPose vary more from one frame to the next. This noisy behavior of shoulder angle estimations and how it compares to those of elbow angles can be seen on Figure 9. In addition, most gestures encountered in pedestrian interaction (pointing, waving, etc.) are observed to involve more pronounced motions of the elbow rather than the shoulder, which puts greater importance on elbow angles.

5.4 Pitch detection

The problem of retrieving the oscillation of the walking motion from the angle data is very similar to pitch detection problems, where the goal is also to identify a low frequency periodicity from a noisy signal. Different methods exist to perform such tasks. In our analysis, we decided to use the average magnitude difference function (AMDF) introduced in [47] as,

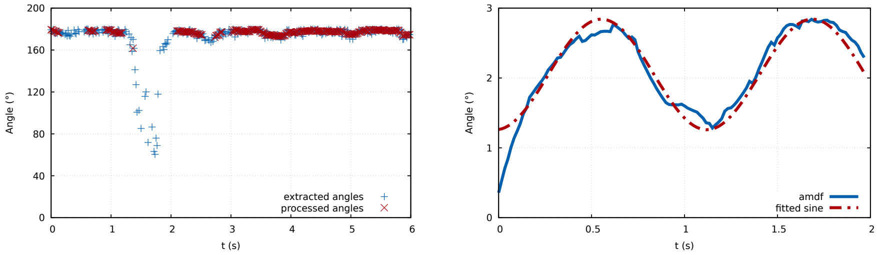
$$D(\tau) = \frac{1}{N-\tau-1} \sum_{n=0}^{N-\tau-1} |x(n) - x(n+\tau)|,$$

where τ is the lag number.

For mere walking action, AMDF should resemble a sine wave and have minimas at lags corresponding to the period of the walking oscillations and its multiples (see Figure 10).

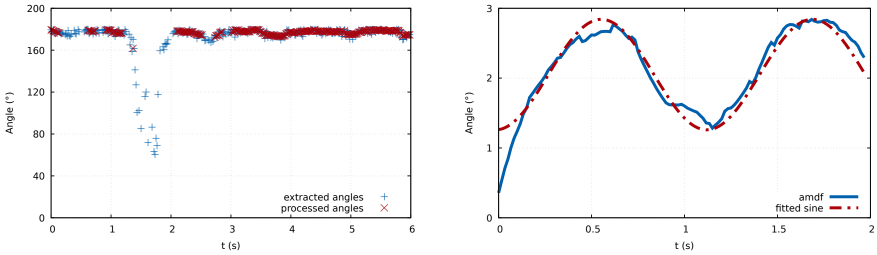
Provided that the input signal contains a prominent periodic component, its AMDF can be approximated with a sinusoidal waveform, $D(\tau) \sim y(\tau)$, such that,

$$y(\tau) = A \sin(\omega\tau + \phi) + c. \tag{12}$$



(a) (b)

Figure 10: (10a) Elbow angles and (10b) AMDF.



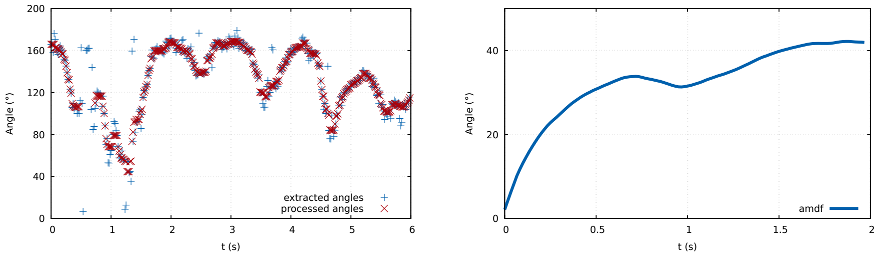
(a) (b)

Figure 11: (11a) Elbow angles and (11b) AMDF and fitted sinus for a no-gesture group.

In Equation 12, A stands for the amplitude of the sinusoidal waveform, ω is its frequency, ϕ is the phase and c is the offset. If periodicity is not pronounced enough, deriving such an approximation would not be possible. In other words, for our specific case, solving for $\{A, \omega, \phi, c\}$ would be possible, only if the limbs are subject to predominant oscillations (possibly due to walking).

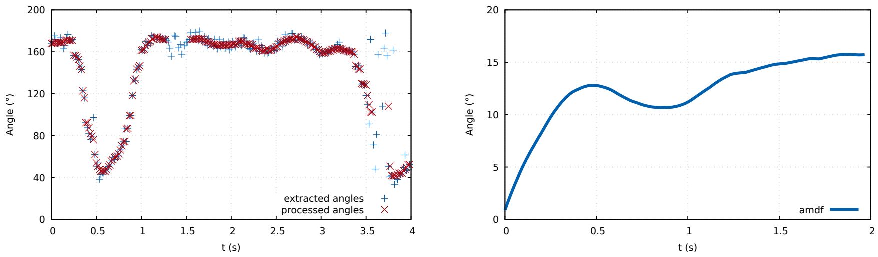
5.5 Results

Running our decision algorithm on all the groups, we obtain the confusion matrix of Table 5. We get a precision of 0.47 and a recall of 0.90, which suggests a model biased toward positive



(a) (b)

Figure 12: (12a) Elbow angles and (12b) AMDF for a strong gesture group.



(a) (b)

Figure 13: (13a) Elbow angles and (13b) AMDF for a weak gesture group.

Table 5: Estimation performance in identification of arm gestures.

		Estimation	
		No gest.	Gest.
Ground truth	No gest.	41	31
	Gest.	3	27

gesture detection. Overall, the accuracy of the model is 0.68.

In Figures 12 and 13, we present two examples, where the algorithm detects the gesture state successfully. In Figure12, the angle waveform is subject to abrupt changes, and the AMDF curve ascertains this. Thus, the dyad is correctly identified as *no gesture*. In Figure13, although the waveform involves such disruptions only for a very short duration, namely less than 2 seconds, the shape of the AMDF reflects the existence of a gesture clearly.

5.6 Discussion

Various sources of error that can have an impact on the performance of the model were identified:

- Pose estimation errors or gaps may arise from occlusions (by object or other peers) and changing view angle due to taking a curve (see Figure 14). Indeed, OpenPose detections relies on the identification of body parts in the picture. If occlusions occurs, the algorithm returns either guessed improbable detections (like for the hips on Figure 14) or no detections (the coordinates are set to $(0, 0)$). While computing the angles, if on joints is missing, the angle is ignored for this frame, but in case of a misplaced joint, a false value will be computed.

- Actions other than gestures may be performed (*e.g.* switching cup from one hand to the other). As a matter of fact, when analyzing the clips that were annotated, it turns out that some pedestrians are engaged in arm gestures that are unrelated to their interaction (*i.e.* they are not pointing at something for instance). The coders

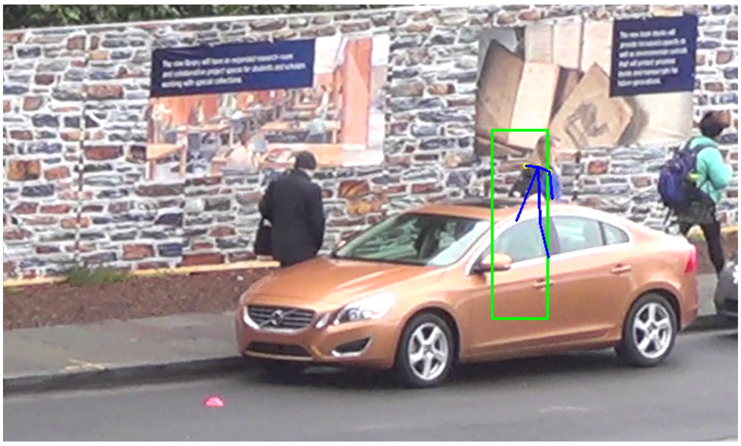


Figure 14: Misdetection of pose due to occlusion.

therefore did not considered these gestures, but the classification algorithm detects them, which increase the number of false-positives.

- Coder's incorrect labels. It is possible that the annotators coded a group as non performing arm gestures, while actually some are being performed. As coders look for gestures during the annotations process and put a positive value only when they notice one, all positive labels can be assumed to be correct while negative ones may be due to failure in noticing the gesture. Consequently, such errors only impact the false-negative side of the confusion matrix.

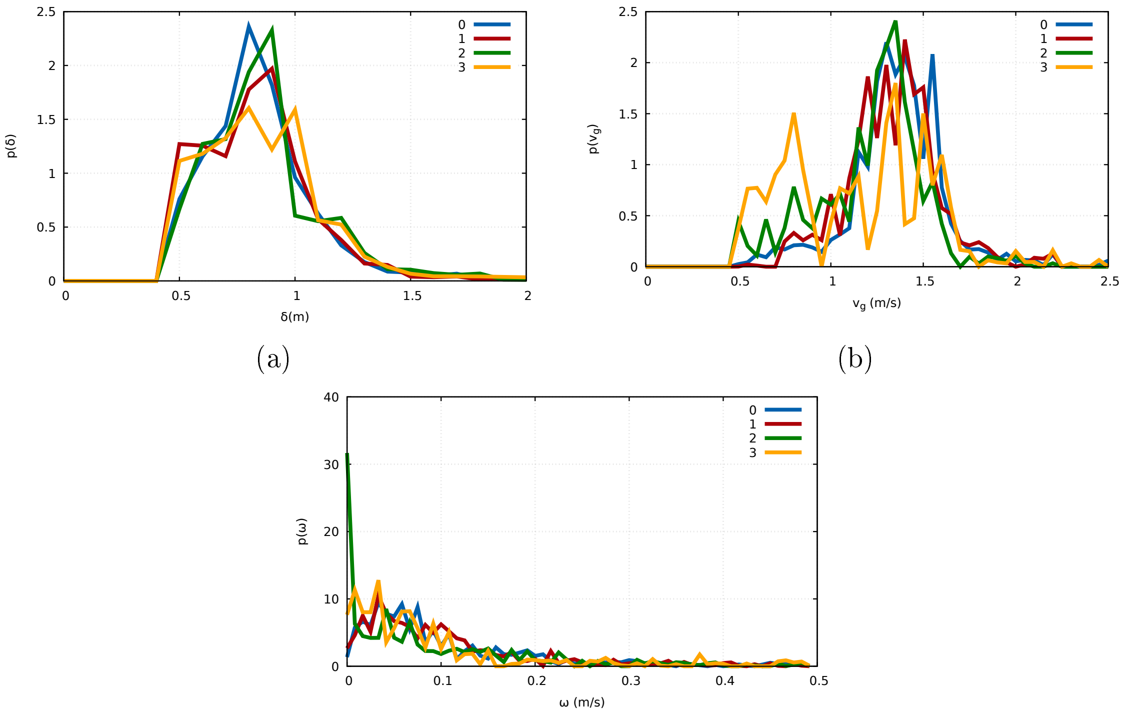
Future work for this study would first require to improve the results of the skeletons estimations to correct missing or false detections. This could be done by performing some interpolation computation to generate skeletons by assuming regular motions of the pedestrians. Another more advanced idea would be to use the recent studies in the field of neural network with graph input data to build a model which would, given a set of detections, generate detections for the missing time steps.

6 Estimation of intensity of interaction

We explored the possibility of estimating intensity of interaction from the observables as described in Section 4. People's height being unavailable in the DukeMTMC dataset, we only computed v_g, ω and δ.

The resulting empirical distributions of the observables for pedestrians interacting at a given intensity are as given in Figure 15.

We expected to see a clearer pattern in the peak location or dispersion of the distributions. For instance, δ for $I = 0$ (Figure 15a), is concentrated around a smaller expected value whereas for $I = 3$ the expected value as well as the tail grow. We would expect the distributions for $I = 2$ and $I = 1$ to follow a gradual pattern between these. In other



(a)

(b)

(c)

Figure 15: Empirical distribution of (15a) δ, (15b) v_g and (15c) ω of peers for dyads interacting at different intensity levels.

words, we would expect δ for $I = 2$ to have a smaller expected value and lighter tail than for $I = 3$. But we observe that this hypothesis is not verified as the order is reversed.

We believe that one reason that can explain these unexpected results comes from the way that the trajectory ground truth is computed in the DukeMTMC dataset. Indeed, as explained in Section 5.1 and illustrated on Figure 7, pedestrians feet positions are annotated at certain frames and then the missing values are linearly interpolated. This leads to three possible sources of error in the trajectories: (i) precision of the annotations limited when clicking between people's feet on an image, (ii) approximation errors when projecting from the image coordinates to the real-world coordinates, and (iii) non-realistic values due to the interpolation. The latest is, in our opinion, the most problematic issue, as it leads to piece-wise constant velocities.

7 Conclusion

To sum up the work exposed in this report, we proposed solutions to the tasks introduced in Section 2, namely (i) recognition of social relations and (ii) detection of gestures.

For (i), we introduced a Bayesian model based on empirical distributions of a set of observables computed from pedestrian trajectories. By computing the probabilities of observables to have been sampled from a given empirical distributions, we are able to classify the social relation of a pair of pedestrians (coalitional or mating) with an accuracy larger than 85%.

We also compared our results for this task with *global* methods (see Section 4.5), based on distributions comparison and showed that the Bayesian model gives the best performance.

For (ii), we proposed a method based on the detection of walking motion pattern from pedestrian arm movements. The idea is that for non-interacting pedestrians, the AMDF of elbow angles movements should resemble a sinusoidal waveform. In case of interaction arm gestures being performed, trying to fit such a sinusoidal function to the computed AMDF should fail. Using this decision algorithm we obtained promising results for what we believe to be a novel approach to pedestrians gesture detection.

We tried to apply the idea used for social relation classification to differentiate the intensity of interactions between pedestrians. Namely, as detailed in Section 6, we computed empirical distributions of observables for trajectories of pedestrians interacting with different level of intensity. The results being inconclusive, we left it out for the time being.

Japan being very involved in the robotic industries, one possible application of this project could be to analyse behaviors pairs of pedestrians to develop robots, that can potentially walk along side with people in the most natural manner.

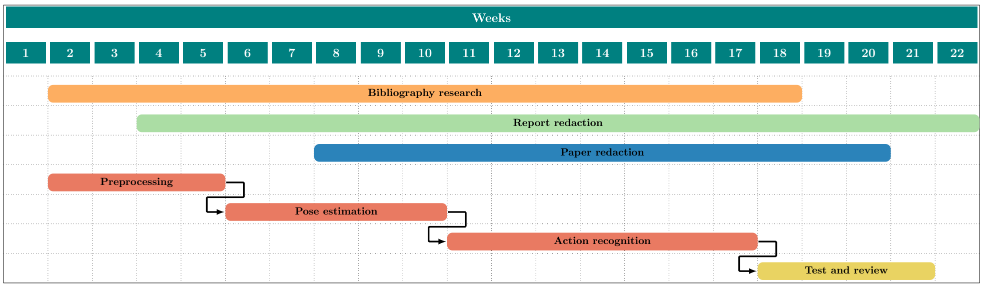


Figure 16: Estimated schedule showing progress of the project.

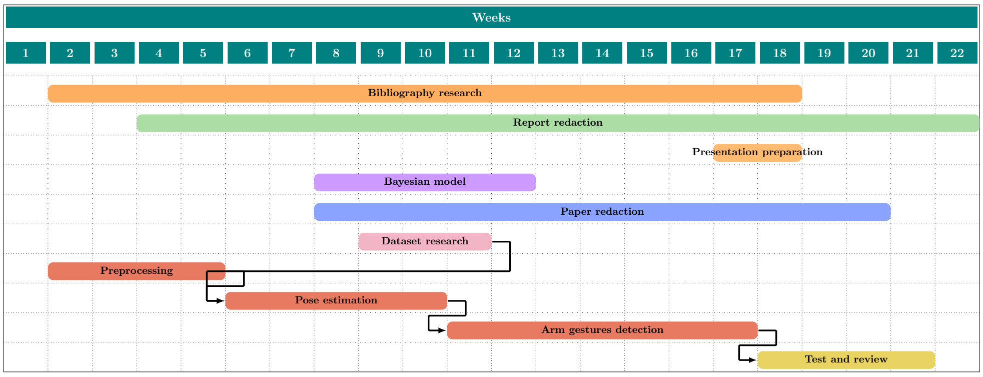


Figure 17: Final schedule of the project.

8 Work organization

As detailed in Section 1, this internship took place in an academic context. It follows that, although the goal to accomplish was well defined, the way to reach it was only drafted. A substantial part of the project consisted in bibliography research and documentation on the various domains that the subject covered.

Figure 16 shows the estimated schedule established six weeks after the beginning of the internship as shown in the pre-report written at that time. Figure 17, on the other hand, shows the actual planning that has been followed during the project.

The major difference is the disappearance of the «Action classification», which has been replaced by «Arms gesture detections». Indeed, for various reasons (limited quantity of data, specificity of the action to detect), the first idea that we had (namely, using a STGCN-like neural network to perform action classification of the pedestrians gesture) turned out to be impractical.

9 Societal and environmental impact

This section will give some ideas of the impact of this project on an societal and environmental level.

9.1 Societal impact

For this project, the possible applications are in the field of robotics. As detailed in Section 2, this could allow to created embodied agents able to follow socially acceptable trajectories. This could be used in a variety of ways, such as guidance or assistantship for disabled and elderly.

Relating to the privacy of the data, for the dataset used in the social relation detection (see Section 4.1), it does not contain video of the pedestrians thus insuring their complete anonymity. For the work on gesture detections, the DukeMTMC dataset (see Section 5.1) is publicly available for research purpose.

9.2 Environmental impact

Regarding the environment, the first part of the project was fulfilled on my personal laptop computer. It was mainly used in battery mode, being charged around two hours per day. According to its technical documentation, this represents a power consumption of approximately 300Wh/day. For the 21 weeks of the internship, this corresponds to a total consumption of around 30kWh.

The project only took place in Okayama University. Except from the round trip from France to Japan, no further travels have been performed, in the context of this internship. Yet, this represents a substantial carbon footprint of 4t of CO2, almost twice the maximum amount that individuals are encouraged not to exceed to prevent global warming. In terms of power consumption, this journey corresponds to more than 8000kWh, which is roughly equivalent to using 250 laptops during the duration of the internship.

Moreover, Okayama being a very flat city, many people (and especially students) use bikes to move around. Likewise, I used a bike to go to work therefore limiting my environmental impact, the international dormitory being very close from the university.

Japan is very involved in environmental procedure such as sorting of waste. In the laboratory office, we had two thrash cans, one for the plastic bottles and another one for the rest of the trash. In the main buildings of the university, there were around 7 different trashes for various kind of waste: burnable, non burnable, cans, etc.

10 Personal feedback

10.1 Work environment

On a typical day of work, I arrive at the laboratory at 9:30AM and leave around 17:30PM (working 7.5 hours/day). I am supposed to share the laboratory with another French student, a Chinese student and 10 Japanese students. The six first weeks of my internship coincided with the end of the school year in Japan, so the laboratory was never actually full and most of the times we were only 4 or 5 working. Dr Akito Monden's office is located in the same floor as the students' office and Dr Yücel's office is located two floors lower.

The overall ambience in the laboratory was great and the integration process went very smoothly, probably mainly because we were all students and because of the Japanese helpfulness and kindness. The furniture in the office also reflect this friendly environments, with bookshelves full of mangas, coding books and retro console and games. A microwave and a fridge are also available directly in the office room. Once a week we all met with the professors to eat together.

10.2 Benefits

10.2.1 Working experience

Overall, this project was a great working experience.

During my previous internships (in 1^{rst} and 2nd year), the work that I performed was never really connected to my orientations ideas. I mainly worked on Web development for industrial companies, and although it was two great work experiences, it did not correspond to the kind of task I pictured myself doing in the future.

On the other hand, during this project, I was able to concretely apply notion that I have adressed during classes at Ensimag. In 3rd year, in the «Pattern recognition and machine learning» lectures, we studied a bayesian model applied to face detection and during the final exam, we worked on developing another bayesian model, applied to digit recognition this time. My work on social relations detection was greatly facilitated by these previous studies.

Moreover, I discovered new domains that I had not studied or only overflown during my studies at Ensimag. For instance, in 2nd year for the «3D Graphic» lecture we quickly introduced microscopic Social Force Model applied to video games or 3D crowd simulation. During this internship, I had the opportunity to deeply improve my knowledge of this kind of models.

On a very technical level, most of the code implemented during this internship was Python code. I was already familiar with this langage that I have been using for a few years, but I took some time to assimilate some good practices that I was not following before. For instance, I used work environment, which are a great way to handle dependencies in Python projects, especially when alternating between different version in different projects.

Finally, I was comforted in my desire to pursue a PhD, probably in the computer vision and machine learning fields of study. The technical freedom associated with research really suits my expectations. A big portion of this internship, as an academic research work, consisted in looking for documentation regarding the state of the art of the diverse steps of the project (pose estimation, action recognition, and pedestrian models) and previous work related to the studied topics. I find this part of the work extremely rewarding on the scientific level as it allows to dive into a wide variety of interesting methods and solutions. Even if most of the papers describe algorithms that could not be applied on this specific project, I enjoyed reading them.

10.2.2 Personal experience

I had the chance to do my internship in Okayama, Japan. This allowed me to discover the Japanese culture, which is very different from what I was used to. Japanese people are extremely helping and always willing to assist in case of difficult situations. Nevertheless, administration is very rigorous and respecting deadlines and regulations is especially important.

Moreover, I had the opportunity to meet a large number of people during this project, not only students from my laboratory but also from the dormitory and other laboratories accros the university. This turned out to be a great human experience and a chance for me to get to know new people from all over the world.

Okayama is a very well situated city in Japan. As a matter of fact, it is close to big cities such as Osaka or Kyoto which are great to visit. The weather is also particularly nice in Okayama prefecture which is rightfully known as the «Land of Sunshine ». I had the chance to be in Japan during the cherry blossom and participate to *Hanami*, the traditional custom during which Japanese people admire the flowering of the cherry trees.

10.3 Difficulties

In relation to the project, the main difficulties that I experienced were linked with environment issues. I started working on my personal Mac Book Pro but many algorithms that I needed to use to test feasibility of considered solutions required an NVidia GPU. Indeed, most machine learning algorithms require heavy computation that is generally accelerated on GPU, very often using the CUDA toolkit that integrates well with traditional machine learning libraries such as Torch or Tensorflow. I started preprocessing the video data until I could use the wanted algorithms on a new computer.

On advantage of the research domain is the freedom of choice when designing a solutions. However a drawback is that there is generally no guarantee that the project will complete. Moreover, the solution might drastically change during if it turns out not to be viable. As detailed in Section 8, I started working on action recognition algorithms and I was planning to use methods based on STGCN [36] to classify the gestures. This idea turned out to be unpracticable due to the specificity of the action that had to be detected as well as the relatively small quantity of data.

10.4 Acknowledgment

I would like to thank all the people that have made this internship possible: Dr. Akito Monden and Dr. Zeynep Yücel for proposing the subject as well as Dr. Bernard Chenevier for giving me the opportunity to postulate. Dr Yücel, in particular, was overseeing my work and was extremely available during the entirety of the project. Dr James L. Crowley was also supervising my work, as an Ensimag representative, and gave me very interesting feedback and help regarding the project.

I would also like to thank all the student from Monden's laboratory for their invaluable kindness and helpfulness.

References

[1] Francesco Zanlungo, Zeynep Yucel, and Takayuki Kanda. The effect of social roles on group behaviour. *Pedestrian and Evacuation Dynamics Conference 2016 Heifei China*, (25):1–10, 2017.

[2] Zeynep Yücel, Francesco Zanlungo, Tetsushi Ikeda, Takahiro Miyashita, and Norihiro Hagita. Deciphering the crowd: Modeling and identification of pedestrian group motion. *Sensors (Switzerland)*, 13(1):875–897, 2013.

[3] Zeynep Yücel, Francesco Zanlungo, and Masahiro Shiomi. Walk the Talk: Gestures in Mobile Interaction. In Abderrahmane Kheddar, Eiichi Yoshida, Shuzhi Sam Ge, Kenji Suzuki, John-John Cabibihan, Friederike Eyssel, and Hongsheng He, editors, *Social Robotics*, pages 220–230, Cham, 2017. Springer International Publishing.

[4] Francesco Zanlungo, Zeynep Yücel, Dražen Brščić, Takayuki Kanda, and Norihiro Hagita. Intrinsic group behaviour: Dependence of pedestrian dyad dynamics on principal social and personal features. *PLoS ONE*, 2017.

[5] Bertrand Maury, Aude Roudneff-Chupin, and Filippo Santambrogio. A macroscopic crowd motion model of gradient flow type. *Arxiv preprint arXiv*, 2010.

[6] Yan Qun Jiang, Peng Zhang, S. C. Wong, and Ru Xun Liu. A higher-order macroscopic model for pedestrian flows. *Physica A: Statistical Mechanics and its Applications*, 2010.

[7] M. Twarogowska, P. Goatin, and R. Duvigneau. Macroscopic modeling and simulations of room evacuation. *Applied Mathematical Modelling*, 2014.

[8] Dirk Helbing and Péter Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 1995.

[9] Systematica. `http://www.systematica.net`.

[10] Winnie Daamen. *Modelling Passenger Flows in Public Transport Facilities.* Delft University Press Science, 2004.

[11] Yuan Gao, Peter B. Luh, Hui Zhang, and Tao Chen. A modified social force model considering relative velocity of pedestrians. *2013 IEEE International Conference on Automation Science and Engineering (CASE)*, 2013.

[12] Yan Qun Jiang, Bo Kui Chen, Bing Hong Wang, Weng Fai Wong, and Bing Yang Cao. Extended social force model with a dynamic navigation field for bidirectional pedestrian flow. *Frontiers of Physics*, 2017.

[13] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social GAN: socially acceptable trajectories with generative adversarial networks. In *CVPR*, volume abs/1803.10892, 2018.

[15] Xiong Hui, Lv Jian, Jiang Xiaobei, and Li Zhenshan. Pedestrian Walking Speed, Step Size, and Step Frequency from the Perspective of Gender and Age: Case Study in Beijing, China. In *Transportation Research Board 86th Annual Meeting*, 2007.

[16] P. A. Hageman and D. J. Blanke. Comparison of gait of young women and elderly women. *Physical Therapy*, 1986.

[17] R. J. Elble, S. Sienko Thomas, C. Higgins, and J. Colliver. Stride-dependent changes in gait of older people. *Journal of Neurology*, 1991.

[18] M. P. Murray, R. C. Kory, and B. H. Clarkson. Walking patterns in healthy old men. *Journal of gerontology*, 1969.

[19] Alan Crowe, Monique M. Samson, Marja J. Hoitsma, and Alexandra A. Van Ginkel. The influence of walking speed on parameters of gait symmetry determined from ground reaction forces. *Human Movement Science*, 1996.

[20] John H. Hollman, Francine M. Kovash, Jared J. Kubik, and Rachel A. Linbo. Age-related differences in spatiotemporal markers of gait stability during dual task walking. *Gait and Posture*, 2007.

[21] Jie Sun, Megan Walters, Noel Svensson, and David Lloyd. The influence of surface slope on human gait characteristics: A study of urban pedestrians walking on an inclined surface. *Ergonomics*, 1996.

[22] Hideo Mori, N. Moghadam Charkari, and Takeshi Matsushita. On-Line Vehicle and Pedestrian Detections Based on Sign Pattern. *IEEE Transactions on Industrial Electronics*, 1994.

[23] S. Yasutomi, H. Mori, and S. Kotani. Finding pedestrians by estimating temporal-frequency and spatial-period of the moving objects. *Robotics and Autonomous Systems*, 1996.

[24] Chiraz BenAbdelkader, Ross Cutler, and Larry Davis. Stride and cadence as a biometric in automatic person identification and verification. In *Proceedings - 5th IEEE International Conference on Automatic Face Gesture Recognition, FGR 2002*, 2002.

[25] Mark S. Nixon, John N. Carter, Michael G. Grant, Layla Gordon, and James B. Hayfron-Acquah. Automatic recognition by gait: Progress and prospects. *Sensor Review*, 2003.

[26] J W Davis. Visual Categorization of Children and Adult Walking Styles. *Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication*, 2001.

[27] Yasushi Makihara, Hidetoshi Mannami, and Yasushi Yagi. Gait analysis of gender and age using a large-scale multi-view gait database. *Computer Vision–ACCV 2010*, 2011.

[28] Sourabh A Niyogi and Edward H Adelson. Analyzing and Recognizing Walking Figures in XYT. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR-94*, 1994.

[29] Nicolas Saunier, Ali El Husseini, Karim Ismail, Catherine Morency, Jean-Michel Auberlet, and Tarek Sayed. Estimation of Frequency and Length of Pedestrian Stride in Urban Environments with Video Sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2011.

[30] Houman Hediyeh, Tarek Sayed, Mohamed H. Zaki, and Greg Mori. Pedestrian gait analysis using automated computer vision techniques. *Transportmetrica A: Transport Science*, 2014.

[31] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[32] Vivek Veeriah, Naifan Zhuang, and Guo-Jun Qi. Differential Recurrent Neural Networks for Action Recognition. *International Conference on Computer Vision (ICCV)*, 2015.

[33] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.

[34] Jun Liu, Amir Shahroudy, Dong Xu, Alex Kot Chichung, and Gang Wang. Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[35] Tomas Kipf. Graph Convolutional Networks. *Blog*, 2017.

[36] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *CoRR*, 2018.

[37] Daphne Blunt Bugental. Acquisition of the Algorithms of Social Life: A Domain-Based Approach. *Psychological Bulletin*, 2000.

[38] Drazen Brscic, Takayuki Kanda, Tetsushi Ikeda, and Takahiro Miyashita. Person Tracking in Large Public Spaces Using 3-D Range Sensors. *IEEE Transactions on Human-Machine Systems*, 2013.

[39] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, 2009.

[40] Caviar test case scenarios. `http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/`.

[41] Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and appearance for mobile scene analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, 2007.

[42] J Ferryman and A Shahrokni. PETS2009: Dataset and challenge. *Pets*, 2009.

[43] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.

[44] Francesco Solera, Simone Calderara, Ergys Ristani, Carlo Tomasi, and Rita Cucchiara. Tracking social groups within and across cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[45] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology.* Sage Publications, 2004.

[46] Zhe Cao, Tomas Simon, Shih En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[47] Andrew Cohen, Richard Freudberg, Andrew Cohen, Richard Freudberg, Myron J. Ross, Harry L. Shaffer, and Harold J. Manley. Average Magnitude Difference Function Pitch Extractor. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1974.

List of Figures

List of Tables

A Detailed performance of alternative methods for recognition of social relation

Table 6: Detailed performance rates for recognition of social relations using KL divergence.

	v_g	ω	δ	Δ_η
M	71.15 ± 7.30	90.58 ± 12.40	78.37 ± 7.83	44.81 ± 7.38
C	81.09 ± 4.50	24.73 ± 15.40	69.40 ± 10.75	92.60 ± 2.04
Total	78.90	39.27	71.38	82.04

Table 7: Detailed performance rates for recognition of social relations using JS divergence.

	v_g	ω	δ	Δ_η
M	72.60 ± 7.27	68.46 ± 5.10	72.12 ± 4.65	55.96 ± 4.55
C	79.13 ± 2.88	64.07 ± 7.98	78.77 ± 5.79	85.00 ± 3.57
Total	77.68	65.04	77.30	78.59

Table 8: Detailed performance rates for recognition of social relations using EMD.

	v_g	ω	δ	Δ_η
M	76.06 ± 6.03	63.37 ± 5.11	74.23 ± 4.19	59.33 ± 5.78
C	76.45 ± 3.19	64.02 ± 4.37	67.51 ± 6.06	82.62 ± 3.37
Total	63.87	76.36	69.00	77.48

B Contact information

Name	Role	Email	Phone number
Adrien Gregorj	Intern	adrien.gregorj@gmail.com	+81 80 4099 5120
Dr Akito Monden	Supervisor	monden@okayama-u.ac.jp	+81 86 251 8180
Dr Zeynep Yücel	Project supervisor	zeynep@okayama-u.ac.jp	+81 86 251 8245
Dr James L. Crowley	Ensimag supervisor	james.crowley@inria.fr	+33 4 76 61 53 96

Table 9: Detailed performance rates for recognition of social relations using Log-likelihood.

	v_g	ω	δ	Δ_η
M	73.94 ± 6.19	63.46 ± 7.30	72.12 ± 4.19	60.38 ± 6.12
C	78.20 ± 2.78	68.11 ± 8.50	78.93 ± 5.03	81.31 ± 4.10
Total	77.26	67.09	77.43	76.69

C French abstract

Ce rapport présente le travail réalisé dans le cadre d'un PFE se déroulant en milieu académique, à l'université d'Okayama, au Japon. Le Dr Zeynep Yücel, superviseure de ce stage, travaille depuis plusieurs années sur la problématique de la modélisation de mouvement de foule et cette étude étend ces travaux. Des résultats dans ce domaine sont particulièrement utiles dans les domaines de la robotique, de la video surveillance ou de la conduite autonome.

Les objectifs du projet sont doubles : (i) la reconnaissance de relations sociales et (ii) la detection de gestes. Pour (i) une méthode probabiliste avec une approche Bayesienne a été développé et pour (ii) une méthode inspirée du traitement de signal audio est proposée.

Pour étudier la reconnaissance de relations sociales, le jeu de données est composé des trajectoires (coordonnées et vitesses) et des tailles des piétons, obtenues préalablement à l'aide d'un environnement de traçage. Un ensemble de données sont calculées pour les trajectoires de paires de piétons : la vitesse du groupe, la différence de vitesse des membres du groupe, la différence de tailles et la distance entre les membres. Le modèle Bayesien développé calcule la probabilité qu'un groupe de piétons soit engagé dans une relation donnée connaissant les variables précédemment introduites.

Ce modèle est également comparé avec des méthodes de comparaison globale de distributions, à savoir la divergence de Kullback-Leibler, la divergence de Jensen-Shanonn, la distance du cantonnier et la log-likelihood. Les meilleurs résultats sont obtenus avec le modèle Bayesien et sont de l'ordre de 80% de précision.

Pour la détection des gestes, le jeu de données DukeMTMC-Groups est utilisé. En plus des annotions de groupes déjà présentes, il a également du être annoté pour relever la présence d'un ensemble de gestes : conversation, échange de regards, contact physique et mouvements des bras. Cependant, seul ce dernier geste est considéré dans cette étude.

L'idée développée ici est que le corps humain est sujet à des oscillations relativement régulières causées par le rythme de marche et que des irrégularités dans ces oscillations (autour des poignets, coudes et épaules) peuvent être considérées comme provenant d'un geste, au sein d'une paire de piétons. Pour detecter et analyser ces irrégularités, une méthode inspirée de la detection de tonalité est utilisée.

Des poses générées par la librairie OpenPose sont utilisées pour calculer l'angle de l'articulation du coude des piétons à chaque image. L'AMDF (average magnitude difference function) du signal est calculée et une fonction sinusoïdale est paramétrée pour approcher la courbe au plus près. Dans le cas où la fonction ne peut pas suffisamment être assimilée à une fonction sinusoïdale, on peut considérer qu'un geste est effectué. En utilisant cet algorithme décisionnel, la précision obtenue est de 68%, et le modèle semble biaisé en faveur de résultats positifs (*i.e.* geste effectué).

Afin d'améliorer ce modèle, les principales sources d'erreurs sont analysées et des pistes de solutions sont proposées (amelioration de la cohérence des poses détectées par exemple).