International Journal of Psychological Research, 2010. Vol. 3. No. 1.
ISSN impresa (printed) 2011-2084
ISSN electrónica (electronic) 2011-2079

Teknomo. K., Estuar, R. E., (2010). Clustering as an EDA Method: The Case of Pedestrian Directional Flow Behavior. *International Journal of Psychological Research, 3* (1), 23-36.

Clustering as an EDA Method: The Case of Pedestrian Directional Flow Behavior.

La clasificación como un método para el análisis de datos exploratorio: el caso de la conducta de flujo direccional de los peatones.

Kardi Teknomo and Ma. Regina E. Estuar
Ateneo de Manila University

ABSTRACT

Given the data of pedestrian trajectories in NTXY format, three clustering methods of K Means, Expectation Maximization (EM) and Affinity Propagation were utilized as Exploratory Data Analysis to find the pattern of pedestrian directional flow behavior. The analysis begins without a prior notion regarding the structure of the pattern and it consequentially infers the structure of directional flow pattern. Significant similarities in patterns for both individual and instantaneous walking angles based on EDA method are reported and explained in case studies.

Key words: Gaussian Mixture, directional flow pattern, pedestrian behavior, trajectory analysis

RESUMEN

Dado que los datos de las trayectorias de los peatones presentan un formato de coordenadas NTXY, tres métodos de clasificación, i.e., K-medias, maximización de valores esperados, y propagación de afinidad, fueron usados como análisis de datos exploratorio (ADE) para hallar los patrones de la conducta de flujo direccional de los peatones. El análisis empieza *sin* ninguna noción a priori sobre la estructura del patrón y tal noción consecuentemente da cuenta de la estructura del patrón de flujo direccional. En este escrito se hace uso del método de ADE para reportar y explicar similitudes significativas entre los patrones de ángulos de marcha de los peatones.

Palabras clave: Mezcla de Gauss, patrón de flujo direccional, conducta peatonal, análisis de trayectorias.

Article received/ Artículo recibido: December 15, 2009/Diciembre 15, 2009, Article accepted/Artículo aceptado: March 15, 2010/Marzo 15/2010
Dirección correspondencia/Mail Address:
Kardi Teknomo, School of Science and Engineering, Department of Information Systems & Computer Science, Ateneo de Manila University, Katipunan Avenue, Loyola Heights Quezon City, Philippines, Email: teknomo@gmail.com
Ma. Regina E. Estuar, School of Science and Engineering, Department of Information Systems & Computer Science, Ateneo de Manila University, Katipunan Avenue, Loyola Heights, Quezon City, Philippines, Email: restuar@ateneo.edu

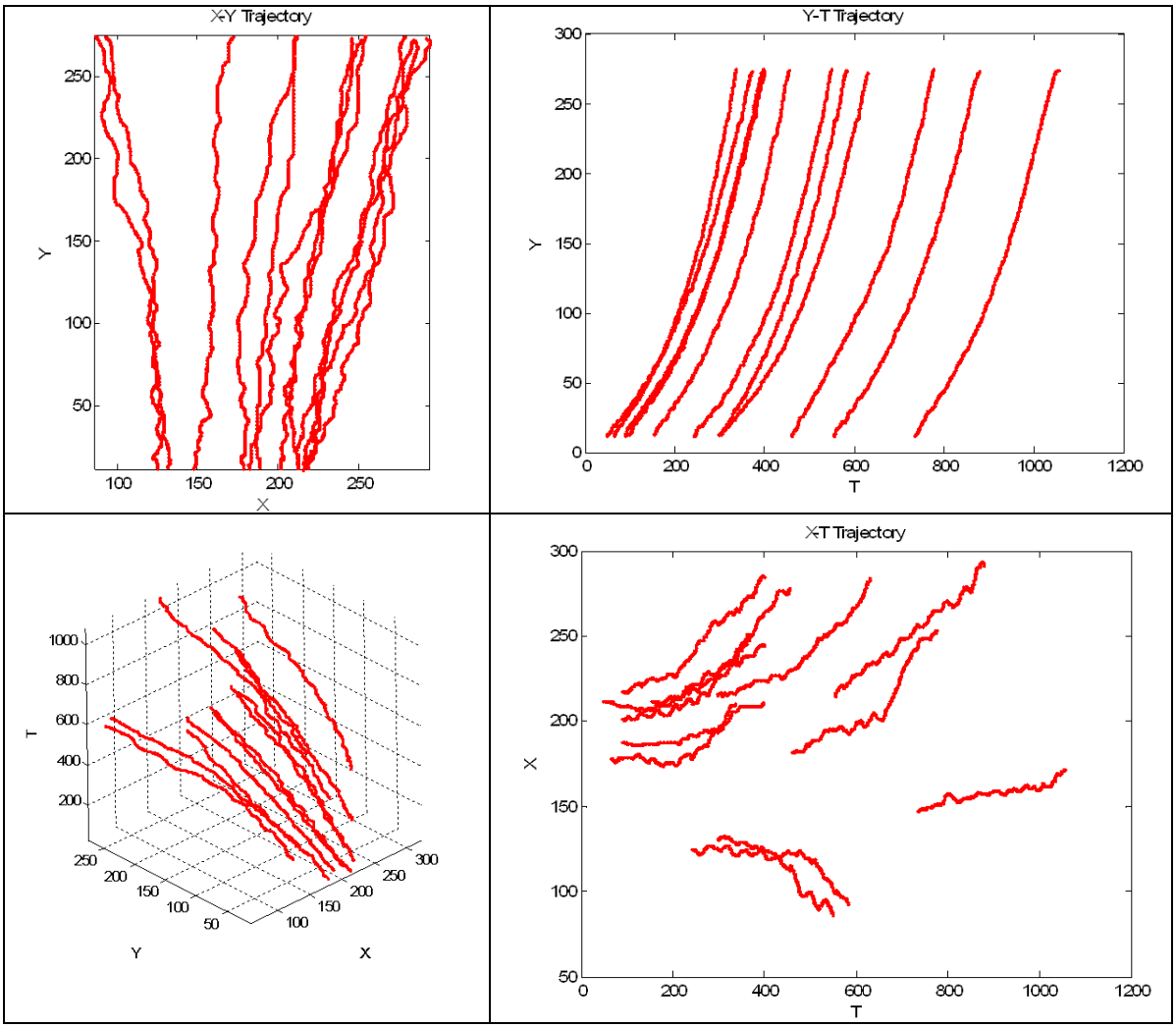
INTERNATIONAL JOURNAL OF PSYCHOLOGICAL RESEARCH esta incluida en PSERINFO, CENTRO DE INFORMACION PSICOLOGICA DE COLOMBIA, OPEN JOURNAL SYSTEM, BIBLIOTECA VIRTUAL DE PSICOLOGIA (ULAPSY-BIREME), DIALNET y GOOGLE SCHOLARS. Algunos de sus articulos aparecen en SOCIAL SCIENCE RESEARCH NETWORK y está en proceso de inclusion en diversas fuentes y bases de datos internacionales.
INTERNATIONAL JOURNAL OF PSYCHOLOGICAL RESEARCH is included in PSERINFO, CENTRO DE INFORMACIÓN PSICOLÓGICA DE COLOMBIA, OPEN JOURNAL SYSTEM, BIBLIOTECA VIRTUAL DE PSICOLOGIA (ULAPSY-BIREME), DIALNET and GOOGLE SCHOLARS. Some of its articles are in SOCIAL SCIENCE RESEARCH NETWORK, and it is in the process of inclusion in a variety of sources and international databases.

International Journal of Psychological Research, 2010. Vol. 3. No. 1.
ISSN impresa (printed) 2011-2084
ISSN electrónica (electronic) 2011-2079

Various methods have been used to understand pedestrian behavior. Most studies have utilized conventional methods such as surveys (Chebat, Gélinas-Chebat, & Therrien, 2005; Elliot, 2004), standard observation methods (Rosenbloom, Nemrodov, & Barkan, 2004), semi-structured interviews (Seedata, MacKenzie, & Mohan, 2005) in everyday and simulated pedestrian scenarios (te Veldea, van der Kamp, Barela, & Savelsbergha, 2005). These methods are merited because it provides us with a rich data set to understand pedestrian intentions and decision making. However, some studies have proposed to include a microscopic view which record motion behavior of pedestrians so that pedestrian flow performance may be quantified (Hoogendoorn & Daamen, 2006; Teknomo, 2002). Using a microscopic lens, pedestrian studies measure motion variables such as walking distance, displacement, speed, velocity, and acceleration. This is made possible by extending data gathering methods to include placement of video cameras in strategic positions to record walking behavior. Captured video data is then translated to include not only pedestrian counting but also include exact location of a pedestrian recorded at a particular time.

In contrast to counting the flow, the basic data structure of microscopic pedestrian is trajectory. By trajectory, we mean the path of a moving object, which in this case, would be the pedestrian. Clearly, trajectory data is simple because it consists of the location of the pedestrian over time, however, the data still contains very rich information that we can harvest through trajectory analysis.

Figure 1. Example of traditional exploratory analysis of trajectories



Teknomo. K., Estuar, R. E., (2010). Clustering as an EDA Method: The Case of Pedestrian Directional Flow Behavior. *International Journal of Psychological Research, 3* (1), 23-36.

The traditional approach to explore trajectory data is to show them in X-Y, X-T and Y-T or even plot the points in 3D as illustrated in Figure 1. In the study of microscopic pedestrian movements, X-Y represents the position of a pedestrian in the defined coordinate system. X-T represents the X position of the pedestrian over time T. The X-T figure shows that the pedestrian walks forward with casual horizontal movements. Similarly, the Y-T presents the Y position of the pedestrian over time T. The Y-T figure can be interpreted as the pedestrian continuously moving forward.

Studies in trajectory analysis have used similar approaches with emphasis on the visual representation of time and space movements (Brillinger, Preisler, Ager, & Kie, 2004; Rinzivillo, Pedreschi, Nanni, Gianotti, Andreinko & Andreinko, 2008; Storck, in press). The traditional approach in exploratory data analysis for hypothesis generation, however, does not take into consideration exploring more aggregate behavior than mere showing the points of the data. Within a particular data set, for example, patterns may emerge from clustering which, in this case, can be used to determine differences in the way people walk. In this paper, we choose one important pedestrian behavior which is directional flow.

This paper extends the analysis of pedestrian behavior by not only providing computations on selected pedestrian flow performance indices but also incorporate analysis based on patterns that can be derived from these indices. In particular, our research problem can be stated as follows: Given a set of trajectory data of microscopic pedestrians, how can we obtain directional flow pattern as a cluster of group walking behavior? A *directional flow pattern* can be defined as a relative weight of pedestrian flow in each direction over a time period of observation. Through directional flow pattern, the three indices can be observed, namely, the angle at which pedestrians are moving, the directional flow patterns that can be inferred from the trajectories, and the relative weight of flow for each direction.

For example, given the trajectory NTXY dataset we would like to recognize whether the pedestrians move in one direction or more directions. Angles are measured relative to a certain coordinate system that we will explain in detail in the later section. Total directional flow and the relative weight of each directional flow is the output of the cluster analysis that can be used in many practical applications. In a busy train station, for example, directional flow pattern can give clue on how many people are going in a certain direction relative to the other direction. Those numbers will be useful for design and planning of the facilities. We define a novel theoretical framework to obtain directional flow pattern from trajectory data and how such system can be automated.

We investigated three clustering algorithms to determine what patterns will emerge from the aggregated flow performance data. For each algorithm, we attempt to answer: at which directional angle the pedestrians moving, how many directional flow patterns we can infer from the aggregated trajectories data and what is the relative weight of flow for each direction (i.e. which direction has more people compare to the other directions). We extend our analysis further by comparing patterns among the three clustering techniques through case studies from pedestrian experiment and real world data that we present it after the framework.

Understanding Exploratory Data Analysis

As early as 1977, exploratory data analysis has been presented as an alternative method in understanding patterns that emerge from a particular data set. In his book, Tukey (1977) presents various methods of visualizing quantitative data using graphs, charts and plots. Patterns are analyzed based on graphical representations of the data set. In a similar manner, other studies have used exploratory data analysis to look for patterns in understanding a wide range of behaviors such as project cycles in businesses (de Mast & Trip, 2008), complex models (Gelman, 2004), mortality (Young, Graham, & Blakely, 2006), and crime (Murray, Mc Guffog, Western, & Mullins, 2001). All of these studies and other similar endeavors have the premise that exploratory data analysis is used in hypothesis formulation more than hypothesis confirmation. As a method, undefined latent constructs emerge from patterns resulting from exploratory data analysis. Based on pattern results, a series of hypothesis are then formulated and may be subjected to validation using confirmatory methods.

Clustering as an EDA method of Trajectory Analysis

Another lens in looking at exploratory data analysis is the visualization of results. Statistical computation is done with the data set to derive clusters or groups based on similarity or nearness of the data. Description or labels are then added based on the characteristics of each cluster. In this paper we attempt to use finite Gaussian mixtures by comparing three partitioning clustering techniques as an exploratory data analysis method in understanding microscopic pedestrian behavior. The three clustering algorithm are K means which represents the basic partitioning method, EM algorithm which extends the basic partitioning method into maximum likelihood estimation and Affinity Propagation, a relatively new partitioning algorithms that claims (Frey & Dueck, 2007) to have promising potential to compete with the two basic techniques. The summary of the three algorithms are as follows.

Describing K means. The aim of K-means clustering algorithm as proposed by Lloyd (1982) is to partition the dataset into a predefine number of subsets k. Let $r_{ij} \in \{0,1\}, j = 1,..k$ be a binary indicator that data point $\mathbf{u}_i$ is assigned to cluster j if $r_{ij} = 1$. The objective of K means clustering algorithm is to find the indicator values $\{r_{ij}\}$ and the centroid of each cluster $\{\boldsymbol{\mu}_j\}$ such that they are optimal in term of distance criterion that minimize within cluster sum of square $\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\left\|\mathbf{x}_i - \boldsymbol{\mu}_j\right\|^2$. In other words, the data point is assigned to the closest cluster center using

$$r_{ic} = \begin{cases} 1 \;\; if \;\; c = \arg\min_j \left\|\mathbf{x}_i - \boldsymbol{\mu}_j\right\|^2 \\ 0 \;\; otherwise \end{cases} \quad (1)$$

The solution for the centroid is given as

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{n} r_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} r_{ij}} \quad (2)$$

Assuming the dataset consists of a finite mixture of Gaussians, we can also compute the covariance and weight for each cluster.

$$\mathbf{C}_j = \sum_{i=1}^{n} r_{ij}\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)^T \quad (3)$$

$$w_j = \frac{\sum_{i=1}^{n} r_{ij}}{n} \quad (4)$$

In other words, given a particular data set, K means clustering algorithm partitions objects into K clusters based on the nearest mean. Characteristics of each cluster are then used to label each group.

Describing Expectation Maximization Algorithm. EM algorithm (abbreviation of Expectation-Maximization algorithm) is an iterative procedure to estimate the maximum likelihood of mixture density distribution. The original theory of EM algorithm (Dempster, Laird, & Rubin, 1977) was to obtain mixture probability density distribution from data samples. Similar to K means, we can also assume that the dataset consists of a finite mixture of Normal distribution of d dimensions with covariance matrix $\boldsymbol{\Sigma}_j = \sigma_j^2\mathbf{I}$.

$$f(\mathbf{x}) = \sum_{j=1}^{m} w_j \frac{1}{\left(2\pi\sigma_j^2\right)^{\frac{d}{2}}} \exp\left(\frac{-\left\|\mathbf{x} - \mu_j\right\|^2}{2\sigma_j^2}\right) \quad (5)$$

The EM algorithm for Gaussian Mixture distribution consists of two steps:

1. Expectation (E) step: compute expected probability π_{ij} that a component distribution j generates that data point i.

$$\pi_{ij}(s) = \frac{w_j}{\lambda}\left(\frac{1}{\sigma_j^d}\exp\left(-\frac{\left\|\mathbf{x}_i - \boldsymbol{\mu}_j(s)\right\|^2}{2\sigma_j^2(s)}\right)\right), \text{ where}$$

$$\lambda = \sum_{l=1}^{k}\frac{w_j}{\sigma_l^d}\exp\left(-\frac{\left\|\mathbf{x} - \boldsymbol{\mu}_l(s)\right\|^2}{2\sigma_l^2(s)}\right) \quad (6)$$

2. Maximization (M) step: maximizes the expected complete data log-likelihood through the estimation of mean, variance and weight of each component of the mixture

$$w_j(s+1) = \frac{1}{n}\sum_{i=1}^{n}\pi_{ij} \quad (7)$$

$$\boldsymbol{\mu}_j(s+1) = \frac{\sum_{i=1}^{n}\pi_{ij}\mathbf{x}_i}{\sum_{i=1}^{n}\pi_{ij}} \quad (8)$$

$$\sigma_j(s+1) = \frac{\sum_{i=1}^{n}\pi_{ij}\left\|\mathbf{x}_i - \mu_j(s+1)\right\|^2}{\sum_{i=1}^{n}\pi_{ij}} \quad (9)$$

The iteration is performed until convergence condition is achieved.

Describing Affinity propagation. Affinity propagation is an unsupervised learning method proposed by Frey and Dueck (2007). Similar to other unsupervised learning such as k-means and Expectation Maximization (EM), affinity propagation takes similarity matrix as input and produces clusters. However, Affinity Propagation does not require number of centers k to be inputted. The algorithm of Affinity Propagation will automatically decide the number of centers as its output. The promise of affinity propagation is that it can be used to determine data centers even for a small data set.

The algorithm of affinity propagation is based on random linkages between data points to create a kind of

International Journal of Psychological Research, 2010. Vol. 3. No. 1.
ISSN impresa (printed) 2011-2084
ISSN electrónica (electronic) 2011-2079

random graph. In each iteration, each data point send messages to other data points. This message is called responsibility because each data point gives responsibility to other data points to become centers. Then, upon receiving the responsibility message, each data point is also send back availability message to other data points to indicate a degree whether that data point is available to become a center or not. Though the convergence is not guaranteed, on most cases, the affinity propagation algorithm will converge. Readers who are interested with the formulations can refer to the online papers and code by Frey and Dueck (2007).

In our work, we extend the results of affinity propagation into finite of Gaussian mixture, formulated similar to the K-means (equation (2) to (4)).

WALKING DIRECTION FROM TRAJECTORIES

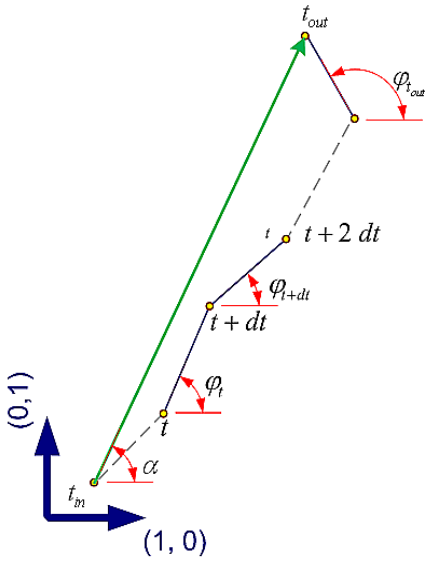
In this section, we explain the definition of directions as flow performance indices from a set of trajectory data. A given trajectory dataset can be summarized into a matrix with four columns. We give name the matrix with this format as NTXY dataset following the tradition in microscopic pedestrian as first proposed by Teknomo, Takeyama, and Inamura (2000). It consists of N = pedestrian ID number, T = time, X and Y as 2D coordinate, either in the image or in the real world. The dataset is assumed to have equally spaced time steps dt (which was obtained after smoothing and re-sampling of the real world data), where outliers have been eliminated at the cost of losing parts of the data. Thus, the trajectory dataset consists of observations in discrete time to serve as an approximation to the corresponding properties of the continuous trajectories.

Each row in the NTXY dataset can be written in compact form as a vector $\mathbf{x}_{i,t}$ to represent coordinate of pedestrian i at time stamp t. As the pedestrian ID is clear from the context, we can further simplify the notation of the coordinate of a pedestrian at time stamp t into $\mathbf{x}_t$. For each individual pedestrian, the first and last recorded time is expressed by t_{in} and t_{out} respectively. The total number of rows in NTXY dataset for pedestrian i can be denoted by $\rho^{(i)}$. Since the pedestrian ID is clear from the context, we drop the subscript for simplicity, thus the total number of observations for each pedestrian is denoted by ρ.

We want to know the movement direction of pedestrian relative to a reference coordinate system. In this case, the movement direction is set based on that reference. The reference coordinate is set arbitrary on the location scene. Then, the movement direction of the trajectory is bounded to that basis coordinate. With respect to aggregation method, we can classify the directional indices into two categories:

- Individual walking angle α: a summary of the instantaneous indices for each individual.
- Instantaneous walking angle φ_t: computed for each individual given a certain time stamp.

Figure 2. *Illustrates the measurement of the two angles. Individual walking angle and instantaneous walking angles*



Before we give the formulation of the two indices, first we need to define the reference coordinate system. Without losing generality, we can define the absolute reference $\mathbf{y}_1 = (1, 0)$ and $\mathbf{y}_2 = (0, 1)$ as two unit vectors spans the XY plane. Since the two vectors are orthogonal ($\mathbf{y}_2 \cdot \mathbf{y}_1 = 0$) they span standard basis vectors.

Individual walking direction uses only two end points (start and end) to determine the direction as illustrated in Figure 2. We define *individual walking direction* as a unit vector that connects the first and last recorded coordinates of a pedestrian trajectory.

$$\mathbf{g} = \frac{\mathbf{x}_{t_{out}} - \mathbf{x}_{t_{in}}}{\left\| \mathbf{x}_{t_{out}} - \mathbf{x}_{t_{in}} \right\|} \qquad (10)$$

Based on the individual walking direction, we can define an individual walking angle relative to the standard basis reference coordinate. *Individual walking angle* is given in term of angle in which the cosine is obtained by

International Journal of Psychological Research, 2010. Vol. 3. No. 1.
ISSN impresa (printed) 2011-2084
ISSN electrónica (electronic) 2011-2079

Teknomo. K., Estuar, R. E., (2010). Clustering as an EDA Method: The Case of Pedestrian Directional Flow Behavior. *International Journal of Psychological Research, 3* (1), 23-36.

the dot product of individual moving direction and the basis vector. As our agreement, we use the first basis vector as the reference axis.

$$\alpha = \arccos\left(\mathbf{y}_1 \cdot \mathbf{g}\right) \tag{11}$$

Alternatively, the individual walking angle can also be obtained from both basis vectors because

$$\tan\alpha = \frac{\mathbf{g}\cdot\mathbf{y}_2}{\mathbf{g}\cdot\mathbf{y}_1} \tag{12}$$

Using only two end points to determine the angle can be misleading if the pedestrian turn direction during his/her course. In contrast to the individual heading angle which only uses the two end points of the trajectory (at the initial time and the end time) to determine the angle, instantaneous walking direction is calculated for each time stamp and hence includes more information on the pedestrian movement. Precisely, *instantaneous walking direction* is defined as a unit vector that represents a heading direction of a pedestrian at a certain time stamp t

$$\mathbf{e}_t = \frac{\mathbf{x}_{t+dt} - \mathbf{x}_t}{\left\|\mathbf{x}_{t+dt} - \mathbf{x}_t\right\|} \tag{13}$$

In continuous trajectory, the instantaneous walking direction is a tangent vector of the trajectory curve at time t. Note that instantaneous walking direction has unit length which implies that $\mathbf{e}_{t+dt}\cdot\mathbf{e}_{t+dt} = \mathbf{e}_t \cdot \mathbf{e}_t = 1.$

Based on the instantaneous walking direction, we can define *instantaneous walking angle* as arc cosine of a dot product between the first basis vector as absolute reference and the instantaneous walking direction.

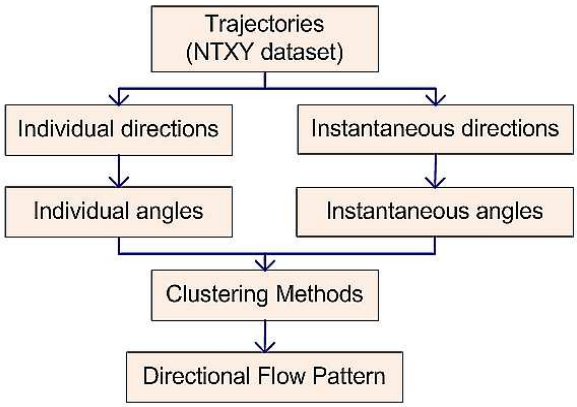
$$\varphi_t = \arccos\left(\mathbf{y}_1 \cdot \mathbf{e}_t\right) \tag{14}$$

The value of instantaneous angle changes for each time step and for each individual. For our analysis, however, we gather these angles irrespective of the individual pedestrian.

DIRECTIONAL FLOW PATTERN FRAMEWORK

Our objective is to find directional flow pattern from any NTXY dataset. Given a set of trajectory data, we can compute individual and instantaneous walking direction. As the directions are defined as unit vectors, we can obtain angle relative to any local coordinate system as explained in the previous section. Once the angles are computed, we use clustering algorithm to get the directional flow pattern. Figure 3 illustrates the framework of our method.

Figure 3. Framework to obtain directional flow pattern



To evaluate the validity of our method, annotated trajectory data were collected manually from several video scenes that we consider these data as the ground truth data. For each video scene, we compare the distribution of individual walking angles and instantaneous walking angles between the ground truth data and the estimated data from the clustering methods. Remarkable results were obtained through case studies presented in the next section that the clustering methods produce the estimated directional flow pattern resemble the original distributions. The clustering methods do not only yield the angles correctly, they also obtain the variation and relative weight of each angle that represent the relative flow between directions.

CASE STUDIES IN PEDESTRIAN ANALYSIS

In this paper we seek to answer the question: Is there a way to obtain directional flow patterns from a trajectory dataset? In this section, we demonstrate our concept and framework in several case studies. The purposes of these case studies are several folds:

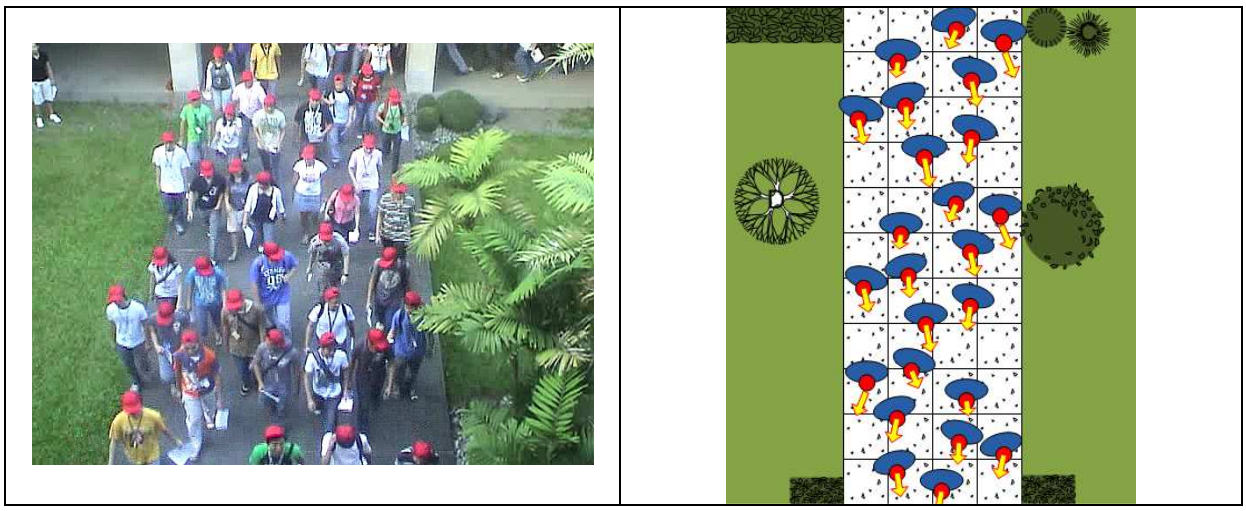
1. To illustrate the process of computation according our framework described in the previous section
2. To show the validity of our method that we can obtain directional flow pattern from the trajectory data.
3. To evaluate which clustering methods are actually suitable for our purpose.

We present two case studies based on the following scenes: One directional pedestrian experiment and two directional real world pedestrian crossing.

Case study 1: One Directional Pedestrian Experiment Dataset

The data set used in this case study was part of a series of experiments on pedestrian movement in Ateneo de Manila University, Philippines. The experiment was designed to capture natural walking behavior of a group of students from a defined origin to a defined destination. A total of 128 students were instructed to walk naturally at normal speed in one direction from North to South. Figure 4 presents a graphical representation of the experiment.

Figure 4. *Set up of Walking Experiment (left: image, right: scheme)*



For the purpose of automatic pedestrian detection, participants were requested to wear red hats. Two video cameras were placed on the second floor to capture the top view of the behavior being observed. The captured data was then translated into an NTXY table where N indicates the pedestrian number, X and Y indicate the position of the person at a particular time T. Specifically, our data consisted of trajectories of pedestrians clicked from images of video sequence. Figure 5 shows an example of the resulting NTXY data set. The actual data set consists of 79,446 rows.

Figure 5. *Sample NTXY data set*

N	T	X	Y
1	12	155	13
1	13	155	14
1	14	155	14
1	15	155	14
1	16	155	16
1	17	155	15
1	18	155	17
1	19	155	17
1	20	151	14
1	21	153	18

In Figure 5, N represents the pedestrian number, T represents the time in seconds, and X-Y represents the position of the pedestrian. 1-12-155-13 means that pedestrian 1 on the 12th second was located at coordinate (155,13).

Preprocessing of Data

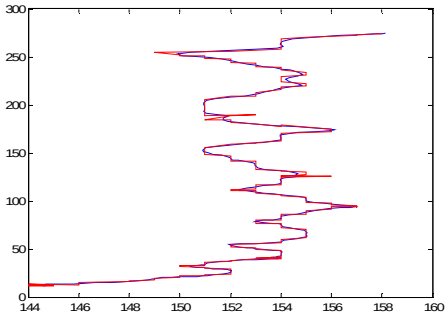
Smoothing was applied on the original data set because the dataset is assumed to have equally spaced time steps. The original pedestrian trajectories are continuous but due to sampling in the image (at rate of 2 Hertz), we lost the smoothness of the data (i.e. the X and Y coordinate the sample data set are repeated over time). Lack of smoothness will cause problem in computation of instantaneous angle because it may incur division by zero. The smoothing is done using cubic smoothing spline in Matlab at smoothing factor 0.5. Figure 6 shows the smoothened data set of the same data we showed in Figure 5. For a consecutive time intervals, the XY coordinate are now more accurately represented.

The reader may suspect that smoothing will change the trajectory data. Figure 7 illustrate the original trajectory and the smoothed trajectory of a pedestrian in the image coordinate.

Figure 6. *Smoothened data set*

1.0000	12.0000	154.9886	13.1308
1.0000	13.0000	155.0036	13.6479
1.0000	14.0000	155.0294	14.0929
1.0000	15.0000	155.0695	14.6128
1.0000	16.0000	155.0956	15.2493
1.0000	17.0000	155.0122	15.7450
1.0000	18.0000	154.6466	16.1170
1.0000	19.0000	153.8607	16.1579
1.0000	20.0000	152.9401	16.2648
1.0000	21.0000	152.6654	17.1664

Figure 7. *Sample trajectory for one pedestrian*



Red line indicates the original trajectory and blue line is the smoothened line. It shows that the smoothed trajectory is indeed very close visually to the original trajectory, thus smoothing does not change the trajectory data. In fact, the smoothing parameters were selected to minimize the average differences between the original trajectories and the smoothened trajectories.

Visualizing Pedestrian Trajectories

Aggregating trajectories for all 128 pedestrians, Figure 8 shows 128 trajectories plotted in XY, XT and YT as well as XYT.

From the X direction over time, it shows the pedestrians are moving in the spreading position over time. The Y position does not change over time because actually the pedestrians are moving from one end of the side walk to the other end.

Directional Flow from Instantaneous Angles

After each clustering of the instantaneous angles, we computed the weight, variance and angles of finite Gaussian mixture and plot them. Figure 9 presents the actual distribution of instantaneous angles and estimated distributions from the clustering methods.

Figure 8. *Traditional trajectories analysis of Philippines Data*

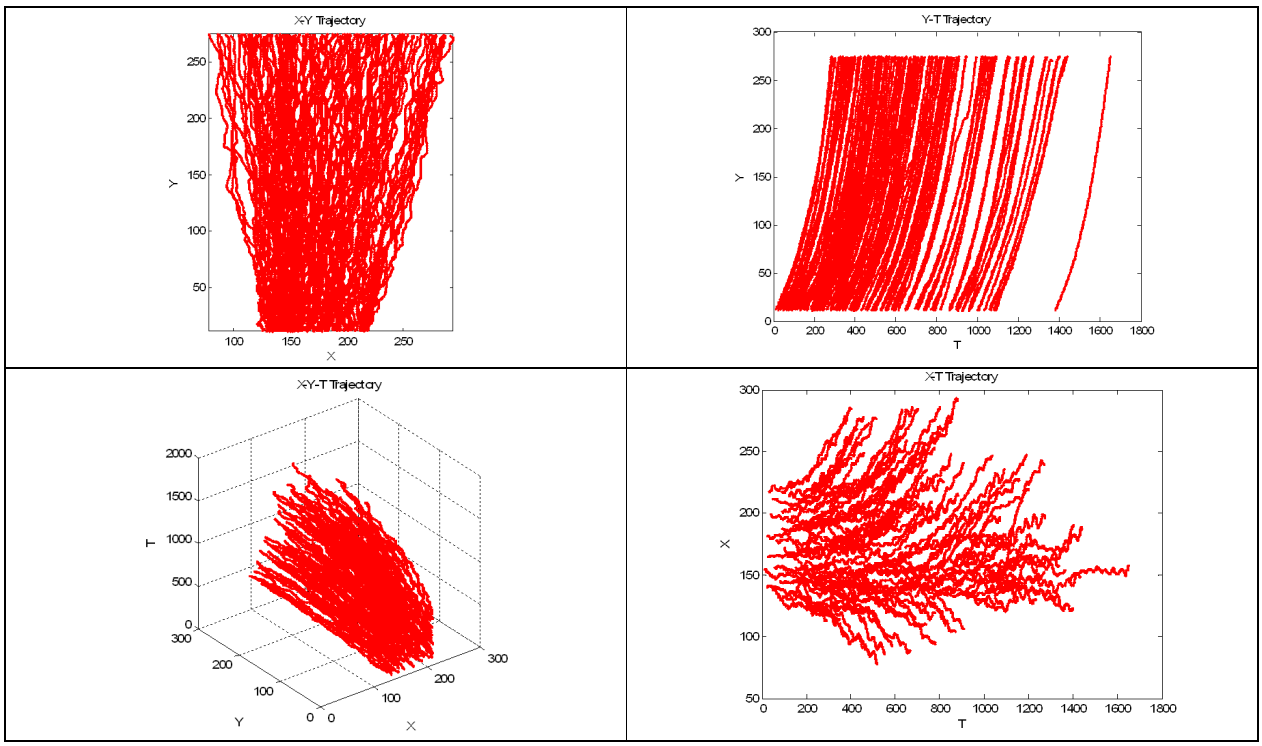
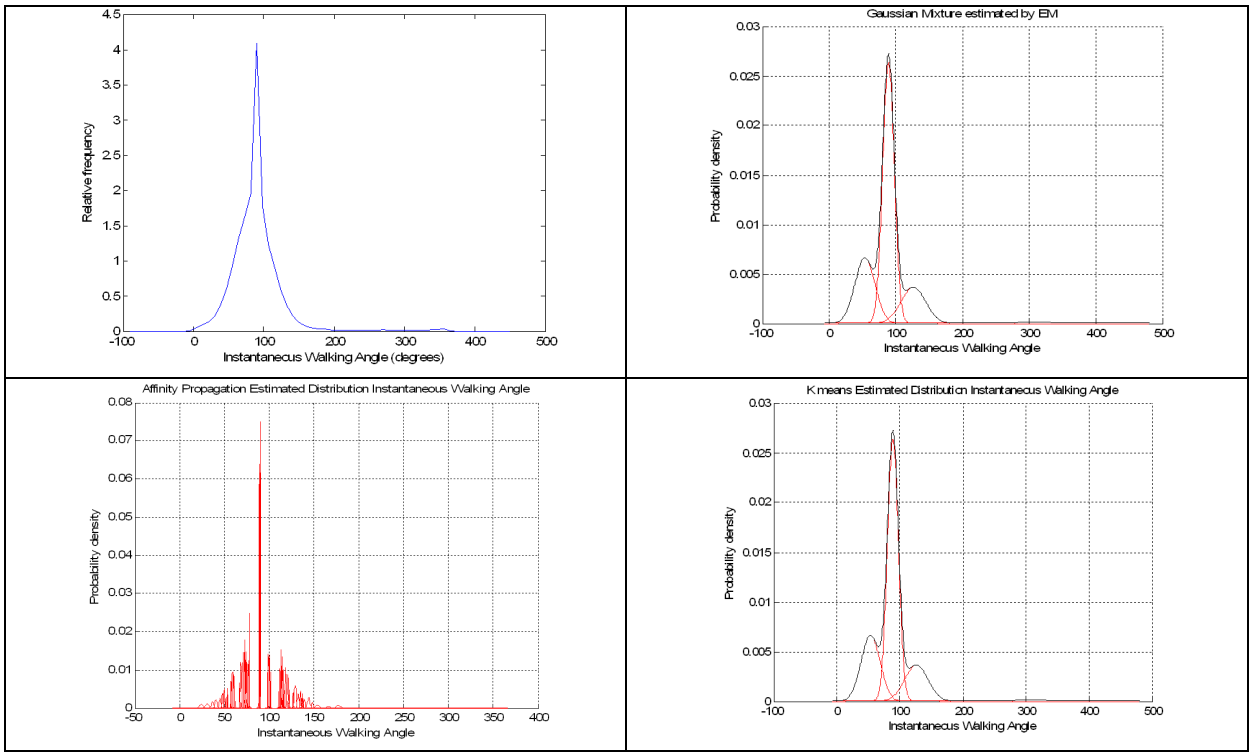


Figure 9. Distribution and Clustering of Instantaneous Walking Angle of Philippines Data



The actual instantaneous angles (79,318 data points) for the 128 pedestrian trajectories are spread from 0.00° to 359.99° with median and means of 89° and standard deviation of 35°.

The 79,000 data points were clustered with the starting K=4 for EM and K Means. The final graphical result for EM and K means show convergence to only one cluster. The average computational time for EM was 4.24 seconds and for K-Means was 14.29 seconds. Similar convergence also happened for Affinity propagation. Note, however, that the Affinity Propagation results of Instantaneous Angles produced about 1,422 clusters after more than 4,435 seconds for only 6 trajectories sampled at random (about 3,000 data points). The reduced number of samples for affinity propagation was due to huge memory and virtual memory requirement for the computation. Using more than 10GB virtual memory, only 4% of the 128 original trajectories were computed.

More detailed results of the mean and variance of instantaneous angles as well as the clustering weights (i.e. basically, the directional flow) are presented in Table 1. Notice that the resulting clusters are the same for both EM and K means. The remarkable results using clustering of instantaneous angles is that we can obtain the correct number of cluster (one cluster in this case) even if it begins from high number of initial clusters. Using affinity propagation, we have even more than a thousand clusters. Regardless the clustering algorithm, all clusters basically have small negligible weights and they are very near to each other that they are actually the same cluster.

Table 1. *Clustering of Instantaneous Angles using EM and K means.*

Algorithm		Cluster 1	Cluster 2	Cluster 3	Cluster 4
K	Mean	126.36°	89.21°	308.05°	53.85°
Means	Variance	335.8	74.9	1859.5	225.4
	Weight	16.72%	57.26%	1.12%	24.90%
EM	Mean	126.36°	89.21°	308.05°	53.85°
	Variance	335.8	74.9	1861.6	225.4
	Weight	16.72%	57.26%	1.12%	24.9%

Directional Flow from Individual Angles

Instantaneous angles are computed for each time step whereas individual angles are computed based on the first and last point of each trajectory. Deriving similar results from individual angles provided us with a more efficient alternative method of computing for pedestrian flow.

Figure 10 presents the actual distribution of 128 individual angles and estimated distributions from the three clustering methods.

Similar to previous procedure, starting with 4 clusters for EM and K Means, the clusters converged to a single cluster of finite Gaussian mixtures for 0.13 seconds. The Affinity Propagation produced 12 clusters for 11.17 seconds. Results also showed that the distributions were actually very close to each other that they actually belong to the same cluster. Thus, we obtain remarkable results for individual angles similar to the instantaneous angles.

Table 2 shows that the detailed mean and variance of the resulting clusters for EM and K means. Notice the similarity between the results of EM and K means.

Algorithm		Cluster 1	Cluster 2	Cluster 3	Cluster 4
K Means	Mean	85.72°	91.41°	77.40°	97.05°
	Variance	3.4483	2.2883	7.6356	2.9006
	Weight	33.59%	21.88%	30.47%	14.06%
EM	Mean	85.34°	90.83°	77.29°	96.76°
	Variance	3.3673	2.58	7.59	3.55
	Weight	31.25%	23.44%	29.69%	15.63%

Table 2. *Clustering of Individual Angles using EM and K means*

However, similar to the results of instantaneous walking angle, affinity propagation resulted to 12 clusters.

Comparing Individual Angles and Instantaneous Angles

The values of the individual angles are between 67.89° and 100.52° with average (mean and median) of 86° and standard deviation of 7.14°. For our experimental data, the individual angles produced more precise distributions than the instantaneous angles. Table 3 presents the descriptive statistics for individual angle and instantaneous angle with mean individual angle at 86.02 ° and mean instantaneous angle at 89.07 °.

Table 3. *Descriptive Statistics of Individual Angle and Instantaneous Angle of Philippine Dataset*

	N	Min Angle	Max Angle	Mean Angle	StdDev Angle	Median Angle
Individual Angle	128	67.89	100.52	86.02	7.14	86.30
Instantaneous Angle	79318	.0074	359.99	7.14	35.42	89.17

The results lead to asking whether the instantaneous angles are actually different statistically from the individual angles. Analysis of variance was used to test for significant differences among sample means from individual angles and instantaneous angles. Results showed that there is no significant difference among all the sample means (F (1, 79444) = .95, p =.33) which implies that individual angles are not statistically different, and therefore, similar to instantaneous angles.

Figure 11 shows the box plots for both instantaneous angle and individual angle with the median of 86.30 ° for individual angle and 89.17 ° for instantaneous angle.

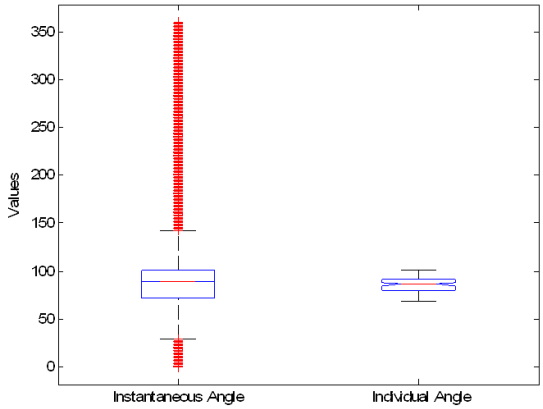


Figure11. *Box Plot Comparison of Instantaneous Angle and Individual Angle of Philippine Dataset*

The box plot shows that the data is slightly skewed. The analysis of variance and box plot confirm our hypothesis that for our limited data the individual angle is actually more precise version of the instantaneous angle.

Case study 2: Two Directional Real World Pedestrian Crossing

Our next case study is based on another ground truth data that we have collected from Pedestrian crossing in Sendai Japan (Teknomo, 2002). This data has been used by pedestrian researchers across different continents (e.g. Bierlaire, Antonini, & Weber, 2003) thus can be considered as one of the standard test data. The scene is illustrated in Figure 12.

Total of 143 pedestrian trajectories were collected manually at 0.5 Hertz using our software.

The traditional plots of trajectories in XY, XT and YT as well as three dimensional XYT are shown in Figure 13.

It is clear from the plots that the trajectories are coming in two directions. The XY plot also shows the range of variation of the trajectories. What is unclear is the relative weight and actual direction.

Figure 10. *Distribution and Clustering of Individual Walking Angle of Philippines Data*

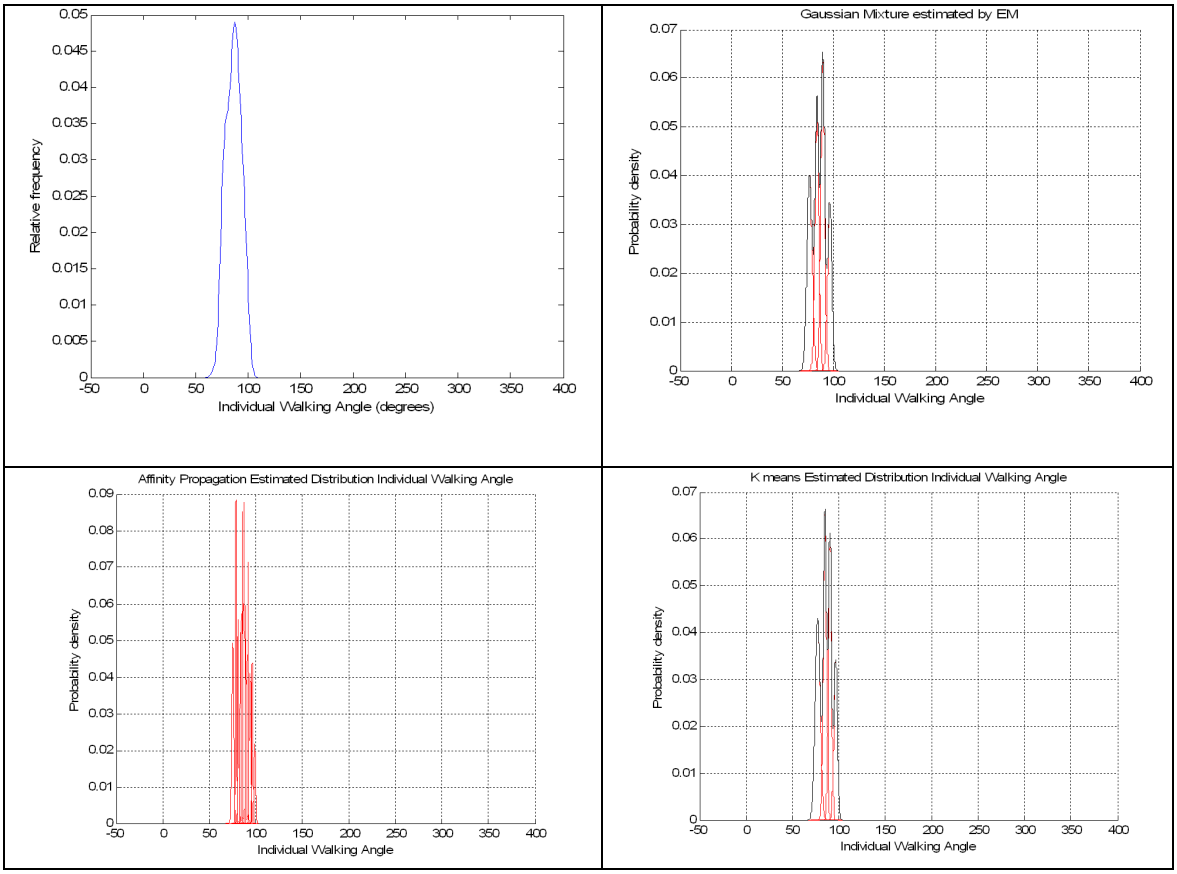
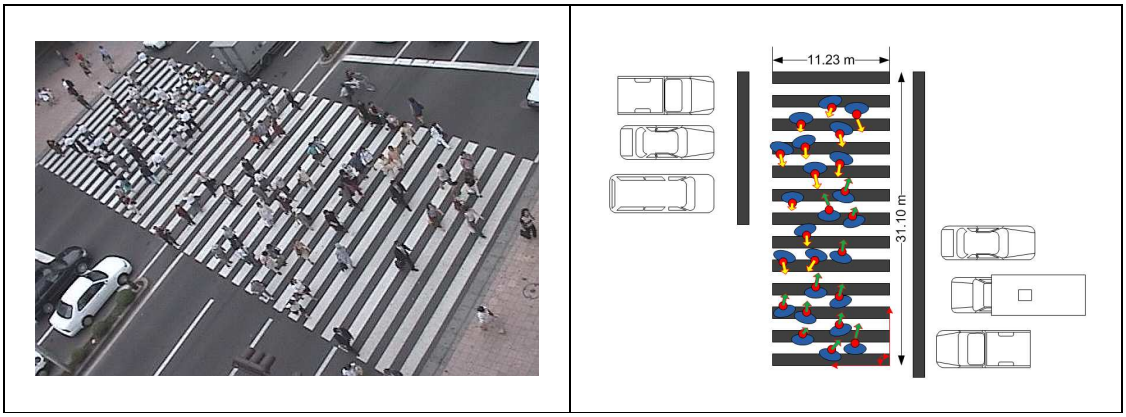


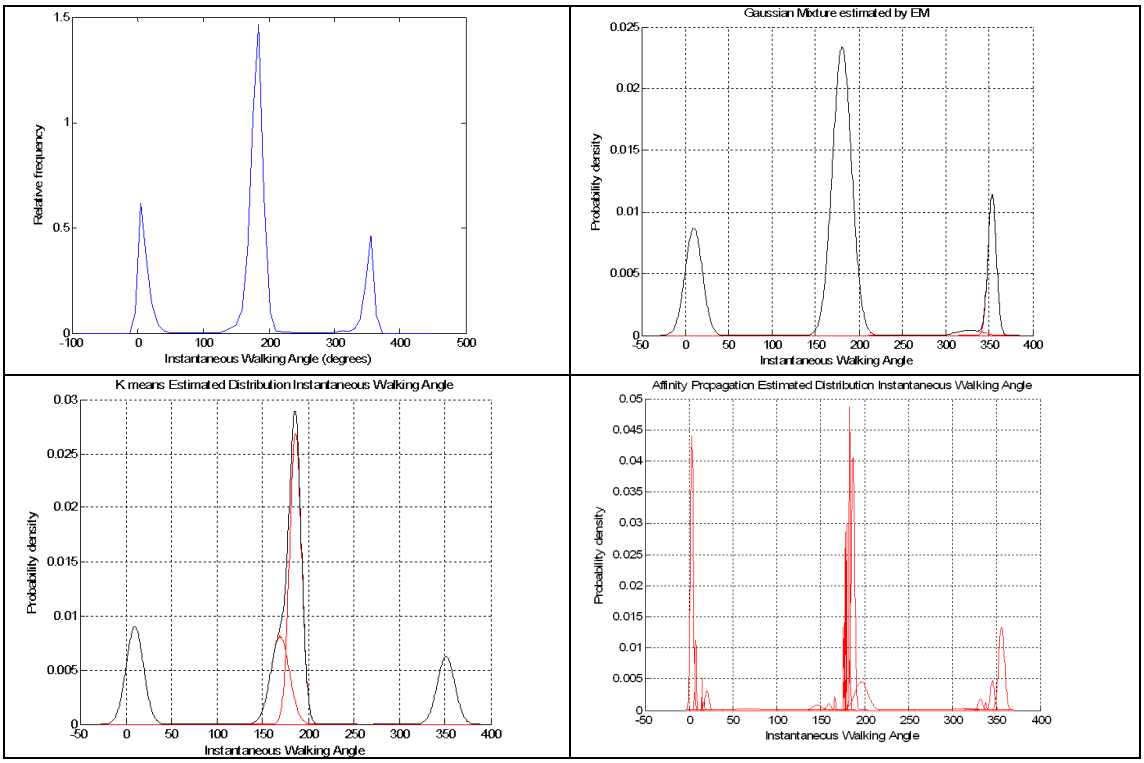
Figure 12. *Setting of Japan data*



Directional Flow from Instantaneous Angles

Figure 14 shows the results of directional flow after clustering instantaneous angles.

Figure 14. *Actual and Estimated Directional Flow Pattern from Japan data based on instantaneous angles*.



The actual instantaneous walking angle has two distributions of zero and 180°. The three apparent distributions in Figure 14 are due to angle computation that 360° is actually the same as zero degree. For EM and K Means algorithm, we started with K equal to 4 and eventually the angles converged into two separated distributions (at least visually, as it is an exploratory data analysis method). The height of the Gaussian represented relative weights of each directional flow and incorrect angles finally tapered off. For affinity propagation, the number of initial distribution is not set by the user. The algorithm of affinity propagation found about 1,207 apparent clusters. However, when we plot those clusters into Gaussian mixtures, we can see visually that they consisted of two angles (zero and 180 degrees). Note that the affinity propagation only used 10% of total trajectories through random sampling to accommodate huge memory requirement of the computation.

Directional Flow from Individual Angles

Figure 15 shows the directional flow of the Japan data set from clustering individual angles.

Individual angles are derived from the starting and ending points. The Gaussian mixture distributions of the three clustering methods produced striking similarities with the actual distribution of walking angles. For EM and K Means, starting with 4 clusters, the results converged into two angles (zero and 180 degrees) while the Affinity Propagation automatically produced hundreds of clusters which can eventually be grouped into two angles.

Comparison of Individual Angles and Instantaneous Angles for Japan Data Set

For both instantaneous and individual walking angles, the values are between zero and close to 360° with an average mean of 166 °, median of almost 180°, and standard deviation of 104°. Similar to the Philippine dataset, the individual angles for the Japan dataset produced more precise distributions than the instantaneous angles. Table 4 presents the descriptive statistics for individual angle and instantaneous angle.

Figure 15. Actual and Estimated Directional Flow Pattern from Japan data based on Individual Angles.

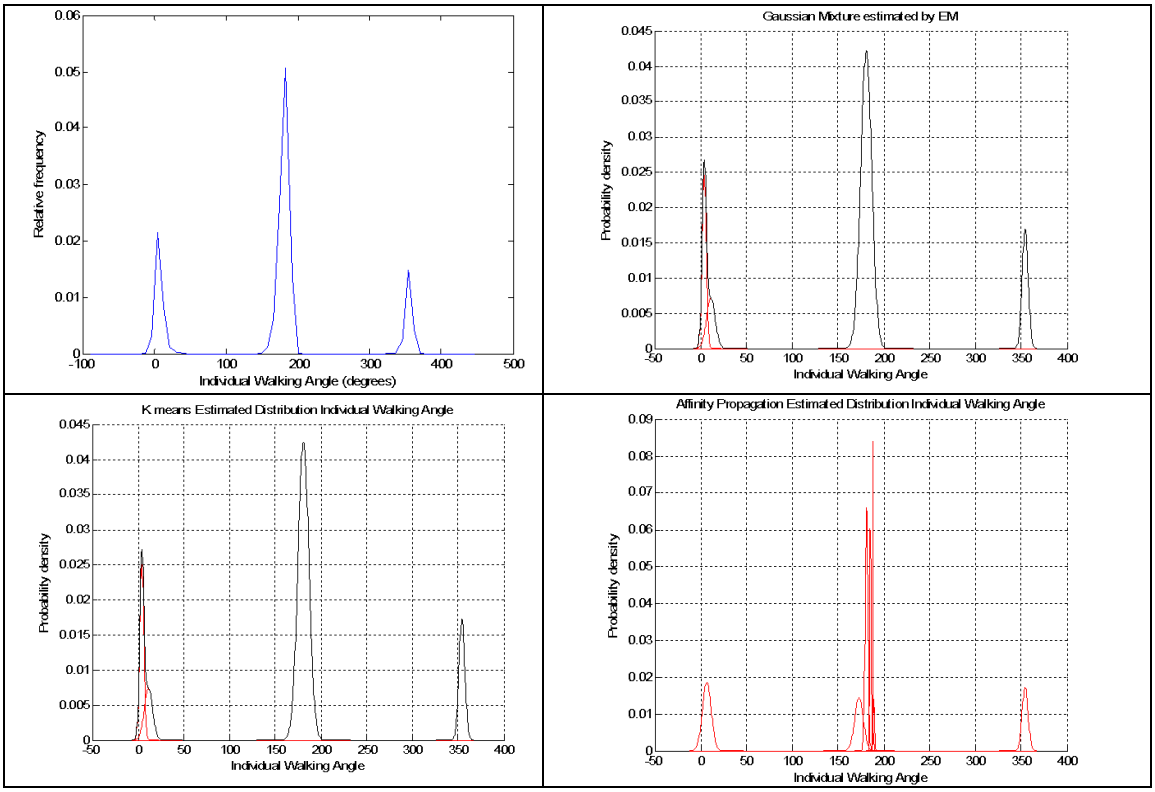


Table 4. *Descriptive Statistics of Individual Angle and Instantaneous Angle of Japan Dataset*

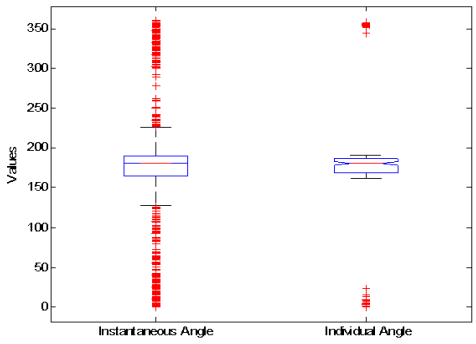
	N	Min Angle	Max Angle	Mean Angle	StdDev Angle	Median Angle
Individual Angle	143	.18	358.00	166.11	104.38	180.74
Instantaneous Angle	15598	.0017	359.99	169.43	102.40	180.83

Analysis of variance was used to test for significant differences among sample means from individual angles and instantaneous angles. Results showed that there is no significant difference among all the sample means (F (1, 15739)= .15, p =.70) which implies that individual angles are similar to instantaneous angles.

Figure 16 shows the box plots for both instantaneous angle and individual angle with the median of 86.30 ^{o} for individual angle and 89.17 ^{o} for instantaneous angle.

The box plot shows that the data is slightly skewed. Similar to the Philippine data set, the analysis of variance and box plot in the Japan data set confirm our hypothesis that for our limited data the individual angle is actually more precise version of the instantaneous angle.

Figure 16. *Box Plot Comparison of Instantaneous Angle and Individual Angle of Philippine Dataset*



Conclusions and Further Studies

We have presented our framework to transform trajectory data into angles. We have also investigated three clustering algorithm, namely, k Means, Expectation Maximization and Affinity Propagation to explore and visualize trajectory data of pedestrians that are aggregated into instantaneous angle and individual angles flow performance. Our findings show that the three clustering

methods produce similar results among each other which also match the actual distributions. Compared to the actual distribution, however, the finite Gaussian mixture produced by the clustering methods yield more precise numerical values to answer at which angle are the pedestrians moving, how many directional flow patterns we can infer from the trajectory and what is the relative weight of flow for each direction.

Instantaneous angles and Individual angles are also similar to each other. In other words, counting the angles simply from the start and end point of trajectories produced similar results (at least of our limited two case studies) and consequently, produced similar distributions of directional flow counted from angles computed for each time step. Of course the number of computation of individual angles is much lesser than instantaneous angles. This result implies individual angle superior than instantaneous angle. However, we should point out that both our case studies are limited for people who do not turn around.

Lastly, comparing the computational time, K-Means and EM are comparable and is recommended for these types of study. For large amounts of data set, however, affinity propagation takes too much computational time and requires large memory for processing the data.

REFERENCES

Bierlaire, M., Antonini, G., & Weber, M. (2003). Behavioral dynamics for pedestrians. In K. Axhausen, *Moving through nets: The physical and social dimensions of travel.* Elsevier.

Brillinger, D., Preisler, H., Haiganoush, K., Ager, A., & Kie, J. (2004). An exploratory data analysis (EDA) of the paths of moving animals. *Journal of Statistical Planning and Inference 122* , 43-63.

Chebat, J., Gélinas-Chebat, C., & Therrien, K. (2005). Lost in a mall, the effects of gender, familiarity with the shopping mall and the shopping values on shoppers' way finding processes. *Journal of Business Research , 58* (11), 1590– 1598.

de Mast, J., & Trip, A. (2008). Exploratory data analysis in quality improvement projects. *Journal of Quality Technology , 39* (4), 301-311.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B , 39* (1), 1-38.

Elliot, M. (2004). *Predicting adolescents' pedestrian behavior.* Wokingham: Transport Research Laboratory.

Frey, B., & Dueck, D. (2007). Clustering by passing messages between data points. *Science , 315* (5814), 972-976.

Gelman, A. (2004). Exploratory data analysis in complex models. *Journal of Computational and Graphical Statistics , 13* (4), 755-779.

Hoogendoorn, S. P., & Daamen, W. (2006). Microscopic parameter identification of pedestrian models and its implications for pedestrian flow modeling. *Journal of the Transportation Research Board , 1982*, 57-64.

Lloyd., S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory , 28* (2), 129–137.

Murray, A., Mc Guffog, I., Western, J., & Mullins, P. (2001). Exploratory spatial data analysis techniques for examining urban crime: Implications for evaluating treatment. *British Journal of Criminology , 41* (2), 309-329.

Rinzivillo, S. P., Nanni, M., Giannotti, F., Andrienko, N., & Andrienko, G. (2008). Visually driven analysis of movement data by progressive clustering. *Information Visualisation , 7* (3-4), 225-239.

Rosenbloom, T., Nemrodov, D., & Barkan, H. (2004). For heaven's sake follow the rules: Pedestrians behavior in an ultra-orthodox and a non-orthodox city. *Transportation Research Part F: Traffic Psychology and Behavior , 7* (6), 395–404.

Seedata, M., MacKenzie, S., & Mohan, D. (2005). The phenomenology of being a female pedestrian in an African and an Asian city: A qualitative investigation. *Transportation Research Part F: Traffic Psychology and Behavior , 9* (2), 139–153.

Storck, J. (in press). Exploring improvement trajectories with dynamic process cost modelling : A case from the steel industry. *International Journal of Production Research* .

te Veldea, A., van der Kamp, J., Barela, J., & Savelsbergha, G. (2005). Visual timing and adaptive behavior in a road-crossing simulation study. *Accident Analysis and Prevention , 37* (3), 399–406.

Teknomo, K. (2002). *Microscopic Pedestrian Flow Characteristics: Development of an Image Processing Data Collection and Simulation Model.* Sendai: Tohoku University Japan.

Teknomo, K., Takeyama, Y., & Inamura, H. (2000). Data Collection Method for Pedestrian Movement Variables. *Dimensi Teknik Sipil - Journal of Civil Engineering Science and Technology , 2* (1), 43-48.

Tukey, J. (1977). *Exploratory data analysis.* Addison-Wesley.

Young, J., Graham, P., & Blakely, T. (2006). Modeling the relation between socioeconomic status and mortality in a mixture of majority and minority ethnic groups. *American Journal of Epidemiology, 164* (3), 282-291.