People Groping by Spatio-Temporal Features of Trajectories

Asami Okada†, Yusuke Moriguchi†, Norimichi Ukita†, and Norihiro Hagita†‡
† Nara Institute od Science and Technology
‡Advanced Telecommunications Research Institute International
e-mail ukita@is.naist.jp

Abstract

This paper proposes a method for detecting people groups from their trajectory data. This grouping is applied to each pair of people. The trajectories of the pair are featured by their spatio-temporal relationships such as a distance and velocities. The features are classified to either of "group" or "non-group" by a discriminative classifier. In contrast to previous features, the proposed features are robust to unsteady behaviors of people and noise of their trajectories. Experimental results using a publicly-available dataset of trajectories demonstrate the effectiveness of the proposed method.

1 Introduction

People tracking is an important issue in machine vision. While several kinds of vision sensors (e.g. video cameras[1, 2] and laser range finders[4, 3]) are applicable to people tracking, the results of all tracking methods are represented by the trajectories of people.

The goal of people tracking is to obtain the trajectory of each individual independently. While this independent tracking is useful for various applications such as surveillance, mutual interactions and relationships between multiple people also give us useful information. For example, the groups of people can be additional clues for people tracking. Navigation and advertisement in a shopping mall depending on the members of each group are also prospective applications.

This paper proposes a method for grouping people in their trajectory data as illustrated in Fig. 1. It is assumed that multiple people are observed simultaneously by a sensor(s), while they are not too crowd to detect groups as shown in Fig. 1 (a). In this paper, it is also assumed that the trajectories are already obtained by existing tracking algorithms. As with previous works[1, 5], spatio-temporal relationships between the trajectories of different people are extracted and classified to either of "group" or "non-group".

Difficulty in people grouping with their trajectories is the unsteady behaviors (e.g. passing and stopping by a crowded spot) and noise of the trajectories. The proposed feature and classification scheme cope with this difficulty by carefully narrowing down and enlarging the types of the features; 1) noisy and ambiguous features are removed and 2) additional useful properties such as cooccurrence of different features are added.

2 Related Work

A number of algorithms have been proposed for people tracking in videos captured by cameras and laser range finders. While people tracking in a dense crowd[6] has been becoming important in machine vision, this paper focuses on relatively-sparse people for detecting groups only from trajectory-based features.

(a) Scene (b) Grouping people trajectories

Figure 1. People in a scene. Their trajectories are divided into groups as illustrated in (b).

The effectiveness of spatial relationships for grouping people and estimating their attributes has been explored also in still images[7]. General differences between the problems in still images and videos are 1) temporal cues are available in the videos and 2) rich appearance features (e.g. age and gender estimation from a face image) are available in the still images while it is difficult to extract such features from the videos because people are imaged smaller in the videos.

As a model for representing interactions between people, a social force model[8] has been widely used. The model is employed in several machine vision problems such as abnormal behavior detection[9] as well as people grouping[1, 10, 5].

In addition to the model, a classification scheme is also crucial for people grouping. A bottom-up hierarchical clustering and a conditional random field are employed in [11] and [12], respectively. In [13], discriminative classification is applied to trajectory data for people grouping.

3 People Grouping by Spatio-Temporal Features of Trajectories

3.1 Basic Spatio-Temporal Features

In the proposed method, each pair of pedestrians is grouped. The trajectories of the pair are featured with their spatio-temporal relationships. Let i and j be the pedestrian IDs of the pair. $\boldsymbol{p}_i$ and $\boldsymbol{v}_i$ denote the position and the velocity of i-th pedestrian, respectively. The following five features (i.e. F1, F2, F3, F4, and F5) proposed in [13] are employed as basic features between i and j at each moment in the proposed method (see Fig. 2):

- (F1) Distance between $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$: $|\boldsymbol{p}_i - \boldsymbol{p}_j|$.
- (F2) Absolute difference in speeds of $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$: $||\boldsymbol{v}_i| - |\boldsymbol{v}_j||$.

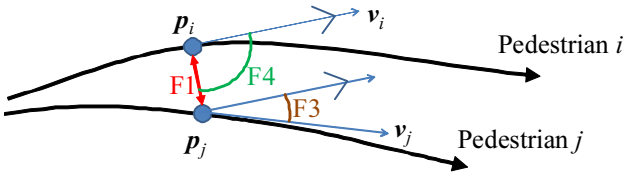


Figure 2. Spatio-temporal features used in [13]. $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ denote respectively the 2D locations of pedestrians i and j at the same time, while $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ denote their velocities.

Figure 3. Distance between each pair in the same group.

- (F3) Absolute difference in directions of $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$: $|\arctan(\boldsymbol{v}_i) - \arctan(\boldsymbol{v}_j)|$.
- (F4) Absolute difference in direction of $\boldsymbol{v}_i$ and relative position between $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$: $|\arctan(\boldsymbol{p}_i - \boldsymbol{p}_j) - \arctan(\boldsymbol{v}_i)|$.
- (F5) Time-overlap ratio: $|\boldsymbol{T}_i \cap \boldsymbol{T}_j| / |\boldsymbol{T}_i \cup \boldsymbol{T}_j|$, where $\boldsymbol{T}_i$ is a set of time steps in which pedestrian i is observed by a sensor(s).

In [13], F5 and the normalized histograms of F1, F2, F3, and F4 are concatenated for obtaining a feature vector. The dimension of this feature is $4d^h + 1$, where d^h is the dimension of each histogram.

3.2 Improving Spatio-Temporal Features

The basic features described in Sec. 3.1 have several problems:

1. Missing cooccurrence in histograms: Since each of F1, F2, F3, and F4, is expressed independently by a histogram, cooccurrence among the different features is not represented.
2. Aeolotropy in F4: F4 is changed depending on whether $\boldsymbol{v}_i$ or $\boldsymbol{v}_j$ is used.
3. Distant pedestrians in a large group in F1: While pedestrians in the same group are expected to be closer, some of F1 features extracted from the large group might be larger. In an example illustrated in Fig. 3, there are three pedestrians whose locations are $\boldsymbol{p}_1$, $\boldsymbol{p}_2$, and $\boldsymbol{p}_3$ in the same group. Since $|\boldsymbol{p}_1 - \boldsymbol{p}_3|$ is larger, it might be closer to a typical distance between pedestrians not in the same group.
4. Unstable directions of velocity in F3 and F4: The directions of velocity might be fluctuated in particular when a pedestrian is standing.

For solving the above problems, the following extensions are implemented in the proposed method:

1. Featurization in each frame: A feature vector (denoted by $\boldsymbol{f}^e_f$) is extracted in each f-th frame so that F1, F2, F3, F4, and F5 are concatenated. $\boldsymbol{f}^e_f$ represents the dependence relationships among F1, F2, F3, F4, and F5. Furthermore, for locally representing a temporal feature, F5 at f-th frame is computed from $(f - T^f)$-th frame to $(f + T^f)$-th frame, where T^f denotes a constant; T^F corresponds to 3 sec in all experiments. Figure 4 shows how to extract the proposed features $\boldsymbol{f}^e_f$ and the original feature (denoted by $\boldsymbol{f}^o$) used in [13].
2. Using two features: $\mathrm{F4}_i$ and $\mathrm{F4}_j$ are computed with $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$, respectively. Both $\mathrm{F4}_i$ and $\mathrm{F4}_j$ are included in a feature vector.
3. Distance between nearest neighbors: If three or more pedestrians are in a group, only a distance to the nearest neighbor pedestrian is regarded as the feature of each pedestrian. If a pedestrian p is paired with two pedestrians, these two pedestrians are also paired with each other through p.

   In an example illustrated in Fig. 3, the nearest neighbors of $\boldsymbol{p}_1$, $\boldsymbol{p}_2$, and $\boldsymbol{p}_3$ are $\boldsymbol{p}_2$, $\boldsymbol{p}_1$, and $\boldsymbol{p}_2$, respectively. In a training step[1], only $|\boldsymbol{p}_1 - \boldsymbol{p}_2|$ and $|\boldsymbol{p}_2 - \boldsymbol{p}_3|$ are trained as features of a group. In a grouping step, on the other hand, $|\boldsymbol{p}_1 - \boldsymbol{p}_3|$ is also evaluated for evaluating whether or not pedestrians 1 and 3 are in the same group. Note that even if they are regarded as people not in the same group because $|\boldsymbol{p}_1 - \boldsymbol{p}_3|$ is larger, they are eventually grouped in the same group if pairs of "1 and 2" and "2 and 3" are grouped.
4. Thresholding in speeds: Features in such frames that the speed of a pedestrian is below a threshold, T^s; $T^s = 0.25$ m/sec in all experiments.

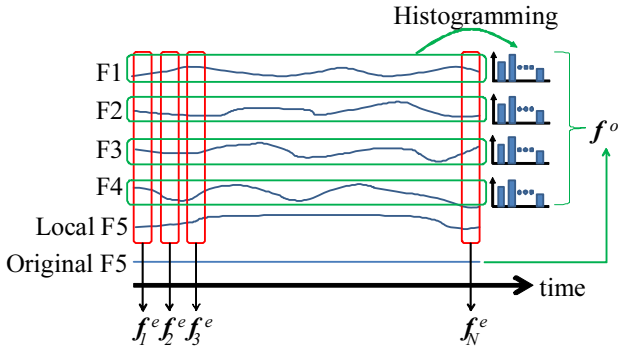


Figure 4. Features used in [13] and the proposed method. Only one feature, $\boldsymbol{f}^o$, is extracted from all frames of a pair of pedestrians by [13]. The proposed features are obtained in each frame, $\boldsymbol{f}^e_1, \cdots, \boldsymbol{f}^e_N$, where N denotes the number of frames in which a pair of pedestrians, i and j, are observed simultaneously: $N = |\boldsymbol{T}_i \cup \boldsymbol{T}_j|$.

While the former two extensions enlarge the features for improving the discriminativity of features, the latter two narrow down the features for suppressing the bad effect due to their noise and ambiguity.

The proposed feature in each frame is defined by a concatenation of the extended F1, F2, F3, $\mathrm{F4}_i$, $\mathrm{F4}_j$,

[1]The training step is described in Sec. 3.4.

Table 1. Percentages of true-positives and false-negatives by different four methods.

	True-positive	False-negative
M1 ([5])	79.5	NA
M2 ([13])	90.7	9.3
M3 (Proposed feature with BoF)	96.2	11.5
M4 (Proposed feature at each frame)	94.4	7.4

and F5 described above. Specifically, since F1, F2, F3, $F4_i$, $F4_j$, and, F5 are scalar values, the proposed feature is a 6D vector.

3.3 Classification of a Set of Features

The proposed feature defined in Sec. 3.2 is classified to "group" or "non-group". One proposed feature is extracted in each frame, while only one feature is extracted from all frames of a pair of pedestrians in the original work[13]. With the proposed feature, therefore, a set of the features are obtained for classifying a pair of pedestrians.

For classifying the set of the features, the following two methods are implemented in the proposed method:

Bag-of-features: In a training step, all features extracted from training data, including "group" and "non-group" features, are clustered (e.g. by using K-means clustering) and then the mean vector of each cluster is obtained. Each set of features extracted from a pair of pedestrians is expressed by a histogram whose bins are represented by the mean vectors.

Frame-by-frame classification: In each frame, a 6D feature vector is classified. If T^w % or more of frames are classified to "group", the corresponding pair of pedestrians is regarded as pedestrians in the same group: $T^w = 66$ in all experiments.

3.4 Discriminative Classification for Grouping

For both bag-of-features and frame-by-frame classification methods, feature classification is required. In [12, 5], the probabilities of "group" and "non-group" are computed from features, while features are classified with a large number of training samples by a discriminative classifier such as the support vector machine[15, 16] (SVM) in [13]. In the proposed method also, features are discriminatively classified by the SVM, in which all positive (i.e. group) and negative (i.e. non-group) samples are employed for training the classifier.

4 Experiments

Experiments were conducted with a dataset of human trajectories[14], which were measured by laser range finders. The trajectories were obtained in a shopping mall (shown in Fig. 1 (a)) where pedestrians walked by, went in various directions, and stopped to browse and chat. This complex situation makes people grouping difficult. The dataset contains the trajectories of 392 pedestrians, including 54 pairs.

Table 1 shows the results of quantitative evaluation. For comparison, the results of the following four methods are shown in the table:

M1: Classification method used in [5].

M2: Classification method used in [13].

M3: Proposed features with bag-of-features based classification.

M4: Proposed features with frame-by-frame classification.

While the results of M1 are taken from its paper[5], M2, M3, and M4 were evaluated under the following conditions. Each result is the mean of 15 trials. At each trial, all trajectories were divided to training and testing data with no duplicates. 10 % of all trajectories were used for training.

The results of the comparative experiments prove the better performance of the proposed feature. It can be seen that one of the proposed methods, M4, overcomes previous methods, M1 and M2, in terms of both true-positive (i.e. correctly detected pairs in the same group) and false-positive (i.e. incorrectly detected pairs NOT in the same group) rates.

Figures 5 and 6 show typical examples of false-positives and false-negatives in M4, respectively. The spatio-temporal features shown in the left-hand and right-hand graphs in Fig. 8 and 9 correspond to the left-hand and right-hand trajectories shown in 5 and 6, respectively.

In Fig. 5, it can be seen that the trajectories of two independent pedestrians got closer in the false-positives. The distance between them is evaluated as F1. For further investigation, the temporal histories of features were shown in Fig. 8. For comparison, those of true-positives were also shown in Fig. 7. The graphs in Fig. 7 and 8 also say that F1 is crucial for classification.

Figures 6 and 9 shows that not only F1 but also other components in the feature affect the results of classification. The trajectories shown in Fig. 6 were regarded as "non-group", while they were close to each other. In the left-hand graph of Fig. 9, a pair was classified to "non-group" at the beginning and the ending of the observation period because F3 was larger. In the right-hand graph, on the other hand, a pair was classified to "non-group" through all frames probably because each component was not greatly different from that of a pair in the same group but every component as a whole was different from that.

Figure 10 shows typical examples in a large group. Although pedestrians ID1 and ID3 were regarded as "non-grouop" by SVM, they were also detected as a pair because "ID1 & ID2" and "ID2 & ID3" were paired by SVM.

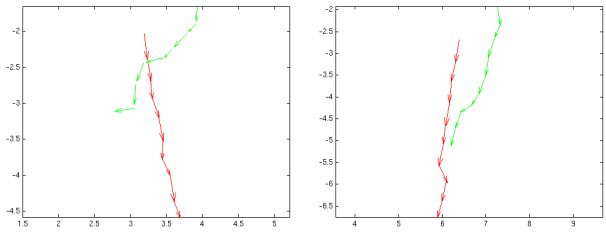


Figure 5. Examples of the trajectories of two pedestrians: false-positives.

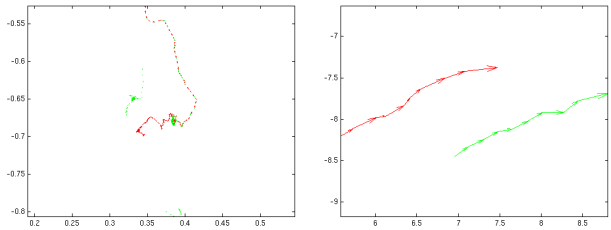


Figure 6. Examples of the trajectories of a pair: false-negatives.

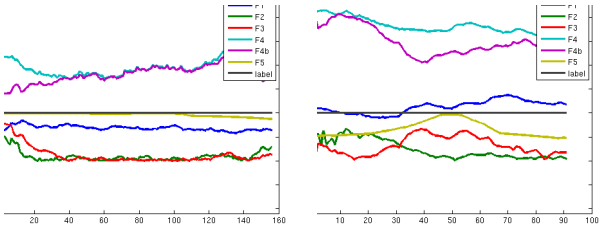


Figure 7. Examples of spatio-temporal features: true-positives. Each graph shows the temporal histories of F1, F2, F3, F4, and, F5 as well as those of classification (denoted by "label").

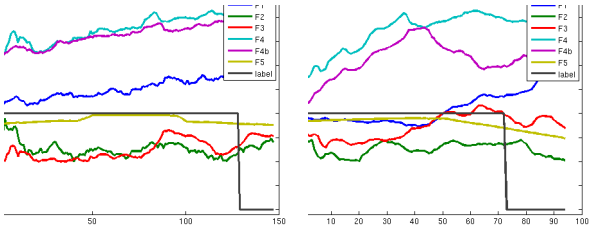


Figure 8. Examples of spatio-temporal features: false-positives. The graphs correspond to Fig. 5.

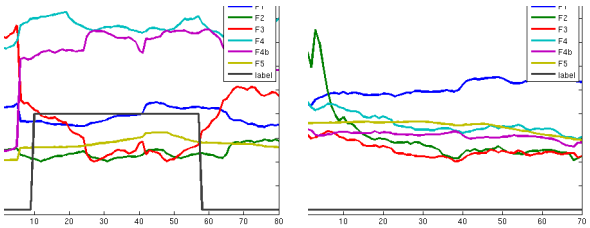
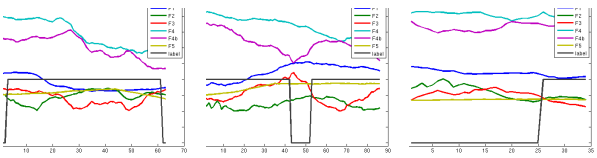


Figure 9. Examples of spatio-temporal features: false-negatives. The graphs correspond to Fig. 6.



ID1 & ID2 ID2 & ID3 ID1 & ID3

Figure 10. Examples of spatio-temporal features in a large group: true-positives. True-positive detections of "ID1 & ID2" and "ID2 & ID3" were propagated to "ID1 & ID3".

5 Concluding Remarks

This paper proposes a method for grouping people by classifying their trajectory data. The proposed feature represents spatio-temporal relationships between a pair of pedestrians at each moment. The trajectories of the pair are expressed by the history of the features and classified to either of "group" or "non-group". Experimental results using a pubic dataset demonstrated the progress of the proposed feature.

Future work includes developing applications of people grouping (e.g. event detection[9] and motion prediction[13]) as well as further extension of the feature for improving robustness to noisy trajectories.

References

[1] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll Never Walk Alone: Modeling Social Behavior for Multi-target Tracking," *ICCV*, 2009.

[2] H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Tracking Multiple People under Global Appearance Constraints," *ICCV*, 2011.

[3] K. O. Arras, O. Mozos, and W. Burgard, "Using boosted features for the detection of people in 2D range data," *ICRA*, 2007.

[4] J. Cui, H. Zhao, and R. Shibasaki, "Fusion of Detection and Matching Based Approaches for Laser Based Multiple People Tracking," *CVPR*, 2006.

[5] Z. Yucel, T. Ikeda, T. Miyashita, and N. Hagita, "Identification of mobile entities based on trajectory and shape information," *IROS*, 2011.

[6] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, "Data-driven Crowd Analysis in Videos," *ICCV*, 2011.

[7] A. C. Gallagher and Tsuhan Chen, "Understanding Images of Groups of People," *CVPR*, 2009.

[8] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical Review E*, Vol.51, No.5, pp.4282-4286, 1995.

[9] R. Mehran, A. Oyama, and M. Shah, "Abnormal Crowd Behavior Detection using Social Force Model," *CVPR*, 2009.

[10] P. Scovanner and M. F. Tappen, "Learning Pedestrian Dynamics from the Real World," *ICCV*, 2009.

[11] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based Analysis of Small Groups in Pedestrian Crowds," *PAMI*, Vol.34, No.5, pp.1003–1016, 2012.

[12] S. Pellegrini, A. Ess, and L. van Gool, "Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings," *ECCV*, 2010.

[13] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and Where are you going?" *CVPR*, 2011.

[14] APT pedestrian behavior analysis. http://www.irc.atr.jp/zeynep/research

[15] B. Scholkopf, K. K. Sung, C. J. C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnikm "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," *IEEE Transactions on Signal Processing*, Vol.45, No.11, pp.2758–2765, 1997.

[16] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, Vol.2, Issue.3, pp.27:1–27:27, 2011.