People Re-identification across Non-overlapping Cameras using Group Features

Norimichi Ukita†[1] and Yusuke Moriguchi†and Norihiro Hagita†

†*Graduate School of Information Science, Nara Institute of Science and Technology*

Abstract

This paper proposes methods for people re-identification across non-overlapping cameras. We improve the robustness of re-identification by using additional group features acquired from the groups of people detected by each camera. People are grouped by discriminatively classifying the spatio-temporal features of their trajectories into those of grouped people and non-grouped people. Thereafter, three group features are obtained in each group and utilized with other general features of each person (e.g., color histogram, transit time between cameras, etc.) for people re-identification. Our experimental results have demonstrated improvements in people grouping and people re-identification when our proposed methods have been applied to a public dataset.

Key words:
People re-identification, Non-overlapping cameras, People grouping

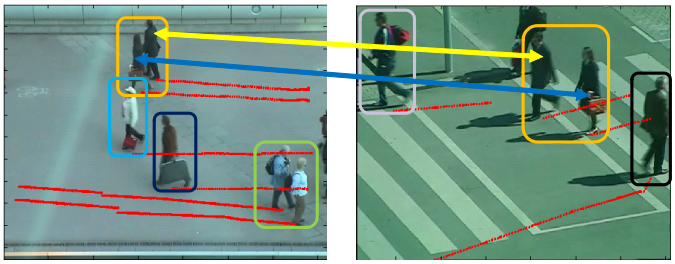


Figure 1: Examples of images captured by surveillance cameras. For wide-area human tracking, identification across fields of view (indicated by solid arrows) is required, as well as tracking within each field of view (indicated by red dotted lines). Each rectangle represents a group.

1. Introduction

Tracking via non-overlapping distributed cameras is crucial for efficient monitoring of people activities over a wide area. Such tracking can be achieved using re-identification methods across multiple cameras following visual tracking in each field of view.

In general, people re-identification in a surveillance scenario should be performed not by recognition using high-level image features, such as face recognition, but by low-level feature matching. This is because image regions of people captured by general surveillance cameras are typically too small for high-level recognition techniques to be applicable.

To improve people re-identification, we propose new features extracted from the group of people detected. The group of people (i.e., each rectangle shown in Fig. 1) is regarded as a set of people who are cohered based on the similarity of their relative positions and trajectories. Under the assumption that people in the same group are, in general, observed together even in different cameras[2], the new features of each group are employed for people re-identification. To improve the accuracy of re-identification, our proposed method combines the proposed group features with traditional image cues that represent the appearance of each person.

The main contributions of this paper are summarized below.

- People grouping using spatio-temporal features of their trajectories: Our proposed spatio-temporal features and classification scheme achieve people grouping such that i) the co-occurrence of different features are expressed and ii) noisy and ambiguous features are removed.
- People re-identification using group features: The group features are represented by the trajectories of people in each group and the number of these people, as well as the image cues of people in the group.

2. Related Work

For people grouping, their trajectories are used in the proposed method. These trajectories can be acquired by visually tracking people in a video. While people tracking in a dense crowd[2, 3] has been increasingly important in computer vision, this paper focuses on scenes with a relatively-sparse number of people for detecting groups only from trajectory-based features.

For grouping people, the proposed feature is represented by spatio-temporal relationships between the trajectories of two pedestrians. The effectiveness of spatial relationships for grouping people has also been explored in still images [4, 5]. Two key differences between using still images and videos are

[2]This assumption is reasonable, for example, in indoor scenes where routes among cameras are limited (e.g. corridors in a building) and nearby cameras between which most pedestrians walk along similar trajectories.

1) temporal cues are available in videos but not in still images and 2) rich appearance features (e.g. age and gender estimation from a face) are available in still images, whereas it is difficult to extract such features from videos, because people are usually imaged smaller in videos.

As a model for representing interactions between people trajectories, the social force model [6], which adjusts a repulsive force between people depending on their social relationship, has been widely used. This model is employed in several machine vision problems such as abnormal behavior detection [7] as well as people grouping [8, 1].

In addition to models, a classification scheme is also crucial for people grouping. Bottom-up hierarchical clustering, iterative clustering using priors of collective behaviors of a group, and conditional random fields (CRFs) have been employed in [9], [14] and, [10], respectively. The classification scheme can also be formulated as a nonlinear optimization problem for finding the optimal trajectories of groups [11, 12]. These approaches are superior in terms of accuracy in group detection as compared with those using simple criteria (e.g. only proximity [13]).

For precise grouping in complex situations, the above-mentioned CRF-based method [10] optimizes its parameters by employing the annotation of people groups (i.e. a group to which each person belongs) in training data. These annotations reveal subtle differences of trajectories between people within and outside a group. With the annotation data, discriminative classification is also possible to improve the accuracy of people grouping [15].

For people re-identification, a number of image features have been proposed, for example, using a set of feature points [18] and histogram of oriented gradients (HOG) [20] based features [19]. Among all such approaches, features using a color histogram are robust against the change in viewpoint of a camera (e.g., a color histogram extracted from an HSV color space [21]).

In addition to the change in viewpoint, variations of illumination among cameras also make re-identification difficult. To cope with this problem, several learning techniques and metrics have been studied, including Adaboost[24], RankSVM[25], multi instance learning boosting[26], and distance learning[27]. While these techniques can improve re-identification across non-overlapping cameras, re-identification with poor appearance features in surveillance videos remains a difficult problem.

Unlike existing image features representing the appearance of people, our focus in this paper is on the features of a group of people. Appearance features of people in a group have been proposed in [34, 35]; these features are extracted not only from a person of interest but also from people who are in the group with that person. Such features are based only on image features and have also been employed for visual tracking (e.g., in [36]). In addition to those image cues, our proposed features also consist of other properties of people in each group (i.e. the trajectories and the number of people in each group).

3. Overview

Figure 2 illustrates the key processes of the proposed method. The entire process consists of two sub-processes, the process performed in each camera and the one that incorporates multiple cameras (these two sub-processes are enclosed by purple rectangles in the figure).

In an image obtained from each camera, several features of each pedestrian used for people re-identification are extracted. To obtain each pedestrian's region and trajectory, the pedestrian is visually detected and tracked.

While the algorithm of group detection has been proposed in [37], new experiments with image sequences are conducted for demonstrating its applicability to a surveillance camera scenario and the proposed people re-identification method.

For people re-identification across multiple cameras, conventional features extracted from the region of each pedestrian are also used, including color histograms (see Section 5.1.1) and transit times between cameras (see Section 5.1.2). In addition to these two conventional features, new group features are employed in the proposed method; these new features are color features of in-group people (see Section 5.2.1), count features of in-group people (see Section 5.2.2), and spatio-temporal trajectory features (see Section 5.2.3). To obtain these new group features, group detection is achieved by employing the trajectories of pedestrians in each group as described in Section 4 below.

4. People Grouping by Spatio-Temporal Features of Trajectories

4.1. Basic Spatio-Temporal Features

In the proposed method, we determine for each pair of pedestrians whether or not they are together in a group. The trajectories of the pair are represented by their spatio-temporal relationships, detailed as follows. Let i and j be the pedestrian IDs, and let $\boldsymbol{p}_i$ and $\boldsymbol{v}_i$ denote the position and velocity, respectively, of the i-th pedestrian. The following five features (i.e. F_1, F_2, F_3, F_4, and F_5) are used in [15] as basic features between i and j at each moment, as illustrated in Fig. 3:

- F_1: Distance between $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$: $|\boldsymbol{p}_i - \boldsymbol{p}_j|$.
- F_2: Absolute difference in speeds of $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$: $||\boldsymbol{v}_i| - |\boldsymbol{v}_j||$.
- F_3: Absolute difference in directions of $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$: $|\arctan(\boldsymbol{v}_i) - \arctan(\boldsymbol{v}_j)|$.
- F_4: Absolute difference in direction of $\boldsymbol{v}_i$ and relative position between $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$: $|\arctan(\boldsymbol{p}_i - \boldsymbol{p}_j) - \arctan(\boldsymbol{v}_i)|$.
- F_5: 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).

While F_1, F_2, F_3, and F_4 are obtained at each frame, their frame IDs are abbreviated for simplicity. In [15], F_5 and the normalized histograms of F_1, F_2, F_3, and F_4 are concatenated to obtain

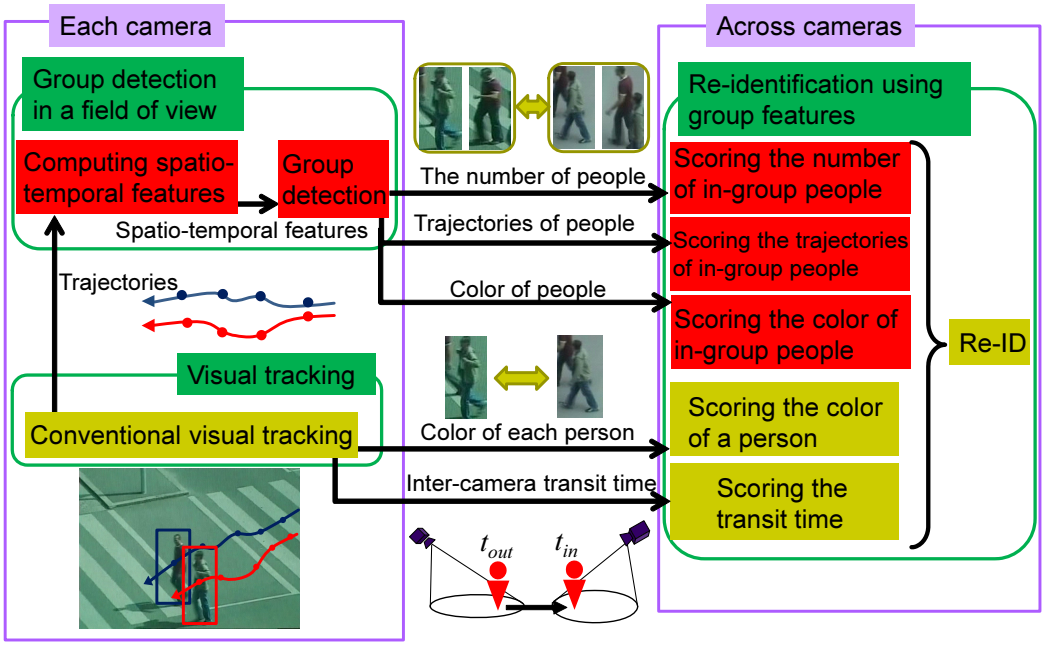


Figure 2: Overview of the proposed people re-identification method. The proposed group features are computed in processes indicated by red rectangles. The three key components of the entire process are colored in green and are as follows: 1) visual tracking in each camera, 2) group detection in each camera, and 3) re-identification across cameras.

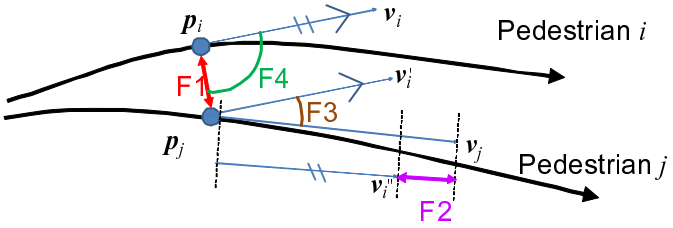


Figure 3: Spatio-temporal features from [15]. $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ respectively denote the 2D locations of pedestrians i and j at the same time; $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ respectively denote their velocities; $\boldsymbol{v}_i$ and $\boldsymbol{v}_i'$ are parallel, and $\|\boldsymbol{v}_i\|$ and $\|\boldsymbol{v}_i''\|$ are equal.

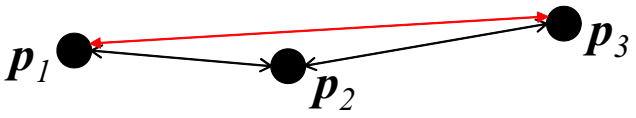


Figure 4: Distances between pairs in a group with three people. Since $\boldsymbol{p}_1$ and $\boldsymbol{p}_3$ (with distance indicated by the red line) are far away from each other, they might be regarded as people in different groups.

feature vector $\boldsymbol{f}$. The dimension of this feature is $4d^h+1$, where d^h is the dimension of each histogram.

Although useful, the basic features described above have several problems, including the ones described below:

1. Missing cooccurrence due to histogramming: Since F_1, F_2, F_3, and F_4 are each expressed independently by a histogram, cooccurrence among the different features at each moment is not represented.
2. Non-robustness of overlap between $\boldsymbol{T}_i$ and $\boldsymbol{T}_j$: When the trajectory of pedestrian i is intermittent because of occlusion or other reasons, $\boldsymbol{T}_i$ is changed. This causes a change in the overlap between $\boldsymbol{T}_i$ and $\boldsymbol{T}_j$ followed by the change in F_5.
3. Distant pedestrians in a large group in F_1: Although pedestrians in the same group are expected to be closer to one another, some of F_1 features extracted from a large group might be larger because members in such a group are separated by a distance. This makes the distributions of group and non-group features overlap. In the example illustrated in Fig. 4, $|\boldsymbol{p}_1 - \boldsymbol{p}_3|$ is larger and might be closer to a typical distance between pedestrians who are not in the same group.

4.2. Improving Spatio-Temporal Features

To solve the problems described in Sec. 4.1, the five features F_1, F_2, F_3, F_4, and F_5 are extended to G_1, G_2, G_3, G_4, and G_5 as follows:

1. Featurization in each frame: A feature vector (denoted by $\boldsymbol{g}_t$) is extracted in each t-th frame so that G_1, G_2, G_3, G_4, and G_5 are concatenated, as illustrated in Fig. 5. Here $\boldsymbol{g}_t$ represents the dependence relationships among all five features at each moment.
2. Temporally-local time-overlap ratio: To suppress negative effects due to intermittent trajectories due to occlusions, a time-overlap ratio, G_5, at frame t is computed only between $t-T_r$ and $t+T_r$, where T_r is a predefined parameter.

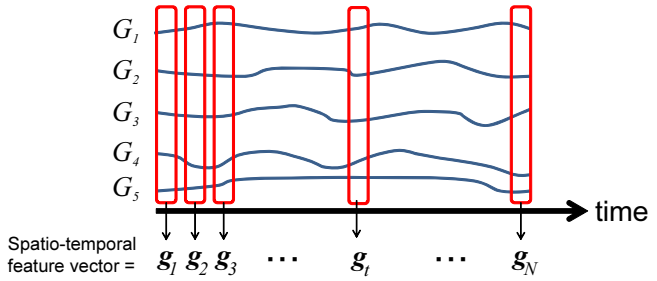


Figure 5: Features used in the proposed method. The proposed features are obtained in each frame, $g_1, \cdots, g_N$, where N denotes the number of frames in which a pair of pedestrians, i and j, are observed simultaneously: $N = |T_i \cap T_j|$.

3. Distance between the nearest neighbors: If three or more pedestrians are in a group in the given training data, only the distance to the nearest neighbor pedestrian is extracted as the feature of the pair in the same group. In the example illustrated in Fig. 4, the nearest neighbors of p_1, p_2, and p_3 are p_2, p_1, and p_2, respectively. In the training step[3], the spatio-temporal features of only two pairs "1 and 2" and "2 and 3" are trained as the features of a group. Note that even if "1 and 3" are regarded as being in different groups because $|p_1 - p_3|$ is larger, they are eventually grouped in the same group if pairs "1 and 2" and "2 and 3" are grouped together.

As with F_1, F_2, F_3, and F_4, the frame IDs of G_1, G_2, G_3, G_4, and G_5 are abbreviated. Their definitions and feature vector g_t are summarized below:

- G_1: Distance between p_i and p_j: $|p_i - p_j|$.
- G_2: Absolute difference in speeds of v_i and v_j: $||v_i| - |v_j||$.
- G_3: Absolute difference in directions of v_i and v_j: $|\arctan(v_i) - \arctan(v_j)|$.
- G_4: Absolute difference in the mean of v_i and v_j and the relative position between p_i and p_j: $\left|\arctan(p_i - p_i) - \arctan(\bar{v})\right|$.
- G_5: Time-overlap ratio: $|T'_i \cap T'_j|/|T'_i \cup T'_j|$, where $T'_i = [k - T_r, \cdots, k + T_r]$
- g_t is a 5D vector defined by the concatenation of G_1, G_2, G_3, G_4, and G_5 at t-th frame:

$$g_t = [G_1, G_2, G_3, G_4, G_5] \quad (1)$$

This feature, g_t, is extracted only from frames where both $|v_i|$ and $|v_j|$ are above threshold T_v. This thresholding allows us to extract the feature robustly to noise in $|v_i|$ and $|v_j|$ computed from the tracking trajectories of objects. Since the computed velocities, $|v_i|$ and $|v_j|$, are noisy and not reliable when the noise level is larger than the walking speed of a person, this thresholding is required for robust feature extraction.

[3]The training step is described in the last paragraph of Sec. 4.3.

4.3. Classification of a Set of Features

The proposed features defined in Sec. 4.2 are classified as "group" or "non-group". One feature vector g_t is extracted in each t-th frame. g_t is classified in each frame. Then the set of features of all frames is classified based on the rate of g_t classified as "group":

$$\frac{N_p}{N_p + N_n}, \quad (2)$$

where N_p and N_n denote the number of g_t classified as "group" and "non-group", respectively. If T_w % or more of frames are classified as "group", the corresponding pair of pedestrians is regarded as pedestrians in the same group. $T_w = 66$, with the value being selected on the basis of preliminary experiments, was used in all experiments.

Assume that all groups of people are known in given training sequences. In the proposed method, features are discriminatively classified by the SVM [16, 17], in which both positive (i.e. group) and negative (i.e. non-group) samples of all training frames are employed for training the classifier.

5. People Re-identification across Non-overlapping Cameras

This section describes people re-identification using group features, which are proposed in Sec. 5.2, extracted from the results of group detection as well as conventional features introduced in Sec. 5.1.

5.1. Conventional Features for People Re-identification

5.1.1. Color Histogram

As used in previous methods [21, 26, 27, 22, 23], the proposed method also computes the similarity of color histograms extracted from the windows of persons of interest. More specifically, the color histogram is extracted from a rectangle with the size of 25 % of the window in order to avoid a negative effect of a background region within the rectangle window. In our experiments, the color histogram was defined as a 24D vector consisting of three 8D vectors that are extracted from R, G, and B channels. Let C_i and C_j denote the color histograms of the i-th and j-th persons, respectively. The similarity score $R_C(i, j)$ is expressed by Bhattacharyya coefficient, which is widely used for non-rigid object tracking [38], between C_i and C_j:

$$R_C(i, j) = \sum_{u=1}^{24} \sqrt{C_i(u) C_j(u)}, \quad (3)$$

where $C_i(u)$ denotes the u-th component of C_i.

5.1.2. Transition Time between Cameras

As proposed in previous methods [28, 29, 30, 31], transit times between cameras, which can be computed from each pedestrian's entrance and exit in the camera's fields of view, are useful for people re-identification. As with [31], the transit

times in each camera pair are expressed by Gaussian distribution $g_T(t; m_T, \sigma_T^2)$. In the re-identification step, transit time is computed from an exit time at one camera and an entrance time at another camera. Let $t_{\mathrm{out},a,i}$ and $t_{\mathrm{in},b,j}$ denote the time when the i-th person exits the FOV of the a-th camera and the time when the j-th person entered the FOV of the b-th camera, respectively. Note that we do not yet know whether or not the i-th and j-th persons are indeed the same person. By substituting transit time $t_{\mathrm{in},b,j} - t_{\mathrm{out},a,i}$ into Gaussian distribution $g_T(t; m_T, \sigma_T^2)$, the score of the transit time, R_T, is obtained as follows:

$$R_T(i, j) = g_T(t_{\mathrm{in},b,j} - t_{\mathrm{out},a,i}; m_T, \sigma_T^2) \quad (4)$$

The greater $R_T(i, j)$, the higher the similarity between the i-th and j-th persons.

5.2. Group Features

The proposed method employs the following three features extracted from each group:

Color of all members in a group: As with conventional color features, a color histogram can be extracted from the regions of the members.

The number of people in a group: It is expected that the number of people in each group is not changed when they are observed in different cameras.

Relationship between the trajectories of people in a group: Independently of fields of view of cameras, the spatio-temporal relationship among people in each group might be uniform; for example, a father is in the center of his children.

Since the latter two features are independent of color, these features might be able to compensate for errors in the color feature. Moreover, the latter two features are undisturbed by any image appearance. This property makes these features robust to a change in viewpoint. Even if the appearance of pedestrians varies between views in terms of orientation and size, these two features are invariant. Therefore, these features are useful for re-identification, if people grouping in each view can be achieved robustly to a change in viewpoint. Since our proposed people grouping is only based on the tracking trajectories of pedestrians, the robustness of the people grouping depends on the performance of visual tracking, which is extensively researched.

Note that, in this work, a person who is alone is regarded as being in a group having only one member. This is because every person must have its group features.

5.2.1. Color Feature of In-group People

The mean of color histograms of all members in a group is regarded as the color histogram of that group. Let $C_{\mathrm{G},i}$ and $C_{\mathrm{G},j}$ denote the color histograms of two groups to which the i-th and j-th persons respectively belong. Similarity score $R_{GC}(i, j)$ between these two groups is expressed by the Bhattacharyya coefficient as follows:

$$R_{GC}(i, j) = \sum_{u=1}^{24} \sqrt{C_{\mathrm{G},i}(u) C_{\mathrm{G},j}(u)} \quad (5)$$

5.2.2. Count Feature of In-group People

The score R_{GH} of the count feature of in-group people is designed so that the score between two persons whose groups are similar in number is higher. Ideally, if the i-th and j-th persons observed by different cameras are the same person, $N_i - N_j$ is 0, where N_i and N_j respectively denote the number of in-group people of i and j. This ideal assumption is often violated because of, for example, unsuccessful group detection and changes in in-group members during inter-camera transitions. These negative effects can be suppressed so that the smaller $(N_i - N_j)$, the greater the similarity score. Such a property can be expressed by the Gaussian distribution as follows:

$$R_{GH}(i, j) = g_{GH}(N_i - N_j; 0, \sigma_{GH}^2) \quad (6)$$

Since Gaussian distribution (6) is defined by constant variance σ_{GH}^2, the probability of successful group detection is considered to be constant independently of the group; however, that probability changes depending on the difficulty in detecting the group of a person of interest. To take into account this difficulty, the variance in score (6) is adjusted on the basis of the discriminativity between the i-th person and other people simultaneously observed in a field of view. More specifically, the variance in score (6) is adjusted so that the lower the confidence in the successful classification of i and n, the larger the variance σ_{GH}^2 (i.e. the lower the weight of score (6), $R_{GH}(i, j)$).

How to obtain the confidence of group detection depends on the method used for group detection. The method proposed in Section 4 classifies a pair to an in-group or non-in-group pair by the SVM. The score of the SVM is positive for a pair in the same group and is larger as the confidence of group detection is higher. The score is, on the other hand, a lower negative value, because the confidence that two persons are not in the same group is higher.

Based on the criteria described above, the score (6), $R_{GH}(i, j)$, is modified as follows:

$$R_{GH}(i, j) = g_{GH}(N_i - N_j; 0, \alpha_h e^{-\beta_h w_i w_j}) \quad (7)$$

$$w_i = \prod_{n}^{M_i} \min(1, |s_{i,n}|), \quad (8)$$

$$w_j = \prod_{m}^{M_j} \min(1, |s_{j,m}|), \quad (9)$$

where 1) M_i is the number of people who are observed simultaneously with the i-th person in a field of view and 2) $s_{i,n}$ denotes the score of the SVM that decides whether or not the i-th and n-th persons are in the same group: $0 \leq \min(1, s_{i,n}) \leq 1$. α_h and β_h are weight variables. If w_i and w_j decrease, the score R_{GH} is not changed significantly.

5.2.3. Trajectory Feature of In-group People

A trajectory feature represents the relationship between the trajectories of nearby persons. Remember that such a feature is defined by a spatio-temporal feature, $\boldsymbol{g}_t$ in (1), used in people grouping. To define a trajectory feature for each person of interest, we have the following difficulties:

- Since the number of total frames in a sequence, in which each pair is observed simultaneously, varies between pairs, the number of spatio-temporal features are also different between the pairs.
- A sequence of spatio-temporal features is obtained for each pair in a group. Since different groups may have different number of in-group people, the number of obtained spatio-temporal features varies between the groups.
- In addition, if the number of in-group people is wrong due to the failure of group detection, this failure may give a negative effect on comparison of spatio-temporal features between detected groups.
- More essentially, only one person also composes its own group. It is impossible to compute the spatio-temporal features of one person.

The proposed method resolves the above difficulties to acquire the trajectory feature of each person by the following steps:

1. Instead of relying on group detection proposed in Sec. 6.1, for each person of interest (denoted by i), a person who has the highest rate (2) for in-group pair detection is found. This person is denoted by i_n. Then the spatio-temporal features between i and i_n is computed.
2. A normalized Bag-of-Words (BoW) feature is obtained in a pair of i and i_n. To obtain this BoW feature, in training, all spatio-temporal 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 computed. Spatio-temporal features extracted from i and i_n are expressed by a histogram (i.e. a BoW feature) whose bins are represented by the mean vectors computed in training.
3. If only i is observed during $\boldsymbol{T}_i$ (i.e., if i_n is not observed), i's BoW feature is determined to be the mean of BoW features between all "non-group" pairs in training data.

The score R_{GT} of the above trajectory features (i.e., BoW features) is defined by the Gaussian distribution of the Bhattacharyya coefficient of the BoW features:

$$R_{GT}(i,j) = g_{GT}(d(i,j);0,\alpha_t \exp(-\beta_t w_{i,i_n})) \quad (10)$$

$$d(i,j) = \sqrt{\sum_{u=1}^{D_B}\left(B_i(u) - B_j(u)\right)}, \quad (11)$$

$$w_{i,i_n} = \min(1, s_{i,i_n}), \quad (12)$$

$$(13)$$

where D_B and $B_i(u)$ denote the dimension of a BoW feature and the u-th component of the BoW feature of i-th person, respectively. α_t and β_t weight variables.

5.3. People Re-identification using Group and Conventional Features

The score R for people re-identification is designed so that R is higher if all five scores (i.e. R_C, R_T, R_{GC}, R_{GH}, and R_{GT}) are greater. R is defined by the product of the five scores as follows:

$$R(i,j) = R_C(i,j)R_T(i,j)R_{GC}(i,j)R_{GH}(i,j)R_{GT}(i,j) \quad (14)$$

$R(i,j)$ is computed for any pair of i-th and j-th persons observed in different cameras. If the $\hat{j}$-th person has the highest score with the i-th person, i is considered to be identified with $\hat{j}$.

6. Experiments

To evaluate the proposed methods, the PRID 2011 dataset [39] was used. This dataset consists of two image sequences captured in nearby locations. Sample frames of two image sequences are shown in Fig. 6. As for annotation data, the 2D trajectory of every pedestrian (i.e. a rectangle enclosing the entire body of a pedestrian and its x and y image coordinates) is given; however, the original annotation data contains 1) trajectories only in such frames that each pedestrian was captured with no occlusion and 2) no group information. The trajectories of all pedestrians were manually completed, and group information (i.e. whether or not each pair of pedestrians are in a group) was also manually provided.

With the revision of the dataset described above, the details of the dataset are as follows:

- Camera a

  Number of frames: 92825 (approximately 62 min)

  Number of pedestrians: 573 including 134 people who belong to groups with two or more people.

  Number of in-group pairs: 134

- Camera b

  Number of frames: 99997 (approximately 67min)

  Number of pedestrians: 884 including 180 people who belong to groups with two or more people.

  Number of in-group pairs: 177

- Relationship between two cameras

  Transition time: 65 seconds on average

  Pedestrians observed in both cameras: 342

6.1. People Grouping

For people grouping, the spatio-temporal features of a pair of pedestrians with their group annotations (i.e. whether they are in a group or not) were given as training data to the SVM.

To use the spatio-temporal features for the proposed group detection method, original image coordinates (i.e. x and y) in

Camera a Camera b

Figure 6: Sample images of two cameras in the PRID 2011 dataset [39].

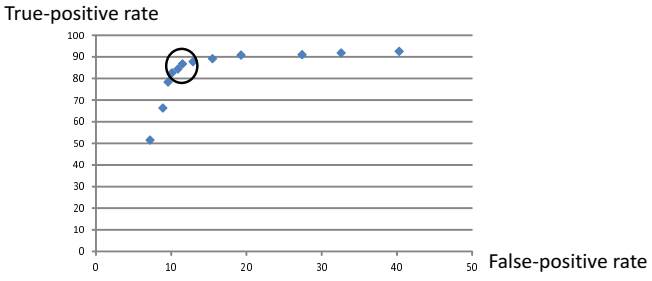


Figure 7: ROC curve depending on two parameters, T_r and T_v.

the dataset are not suitable. This is because the method implicitly assumes that undistorted planar coordinates are used (i.e. that the image plane is parallel to a ground plane), though the image plane of a surveillance camera is generally inclined with respect to the ground plane. To resolve this problem, the image coordinates of pedestrians were rectified by a homography between the image plane and the ground surface.

In addition to the rectification of image coordinates, the scales of image coordinates were normalized between different cameras. This scale normalization was executed so that the mean velocities (in pixels) of pedestrians were the same between cameras. Finally, training data for all cameras can be used to acquire a unified SVM, which can find groups observed in any cameras.

Results of group detection are shown in Table 1; results shown are the mean of 15 trials. For comparison, the previous method [15] was also evaluated. For each trial, all trajectories were divided into training and testing data with no duplicates; 10 % of all trajectories were used for training.

The proposed spatio-temporal feature, $\boldsymbol{g}_t$, requires two parameters, T_r and T_v. To determine these parameters, true-positive and false-positive rates were evaluated among several sets of parameters with training data in each trial. Sample results from one trial are shown in Fig. 7. In each trial, the best set of two parameters were selected so that ((True-positive rate) - (False-positive rate)) was maximized. As enclosed by a circle in Fig. 7, several sets of parameters are close to each other around the best set of parameters. Since the results of people grouping with these parameters sets are similar to each other, the setting of two parameters, T_r and T_v, is not so sensitive.

Experimental results are shown in Table 1. These results show better performance of using the proposed method both in terms of true-positive and false-negative rates.

The more people in a group, the more difficult people grouping becomes. This is because if a group consists of three or more members, all in-group pairs in this group should be detected for complete detection of this group in order to obtain the group features of the members successfully. In the dataset, the numbers of in-group pairs in groups consisting of two members and three or more members are 159 and 41, respectively. The numbers of correctly-detected groups were 141 and 25 (i.e. 87 % and 61 % of all groups), respectively. Since the accuracy of group detection was lower in groups consisting of three or more members, important future work includes improving detection of a large group.

6.2. *People Re-identification*

The following three tests were conducted for evaluating the proposed group detection and re-identification methods:

1. The proposed people re-identification method was evaluated with the ground truth of the groups of people (see Section 6.2.1).
2. The proposed people re-identification method was evaluated with people groups detected by the proposed group detection method (see Section 6.2.2).
3. The proposed people re-identification method was evaluated with people groups detected from trajectories obtained by a visual tracking method [33].

All 342 persons observed both in cameras a and b were used for re-identification evaluation. Re-identification scores, $R(i,1), R(i,2), \cdots, R(i,884)$, were computed for pairs of the i-th person, who was one of the 342 persons, observed in camera a and all 884 persons observed in camera b. To evaluate the re-identification scores, a cumulative match characteristic (CMC) curve is used.

Two parameters, σ^2_{GH} in Eq. (4) and a in Eq. (7), were $\sigma^2_{GH} = 0.222\cdots$ and $a = 1$, which were determined empirically.

6.2.1. *People Re-identification with the Ground Truth of People Groups*

Scores (5,7) were computed with the ground truth of people groups. In this case, score (7) is essentially the same as score(6), so only score (6) was evaluated.

Figure 8 shows results obtained by using the color feature and the proposed three group features. In the CMC curve, the vertical and horizontal axes indicate the identification rate (i.e. the percentage of people whose correct identification appear below each rank of matching score (14)) and the ranks, respectively. From our results, it is observed that the count and trajectory features increased the accuracy of re-identification, as is evident in the difference between "R_C and $R_C R_{GH}$" and "R_C and $R_C R_{GT}$", respectively, in the figure. Further, the color feature of in-group people also contributed to an increase in accuracy, as is evident from the fact that $R_C R_{GC} R_{GT} R_{GH}$.

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

	True-positive	False-negative
Previous method [15]	81.1	13.6
Proposed method	86.7	11.5

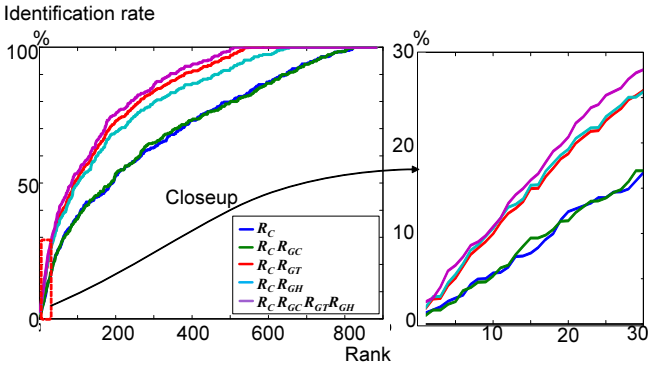


Figure 8: CMC curve of identification using color features and three group features obtained from the ground truth of people tracking and grouping.

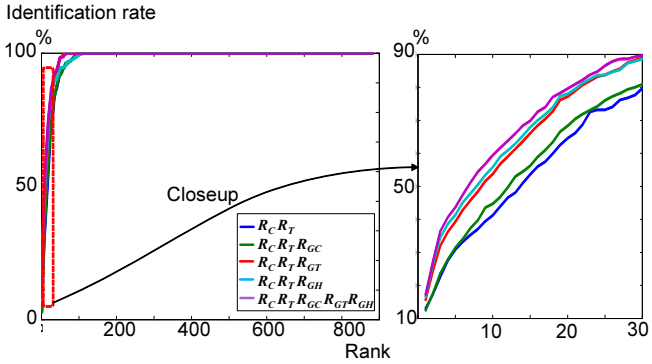


Figure 9: CMC curve of identification using the color and transition interval features as well as three group features obtained from the ground truth of people tracking and grouping.

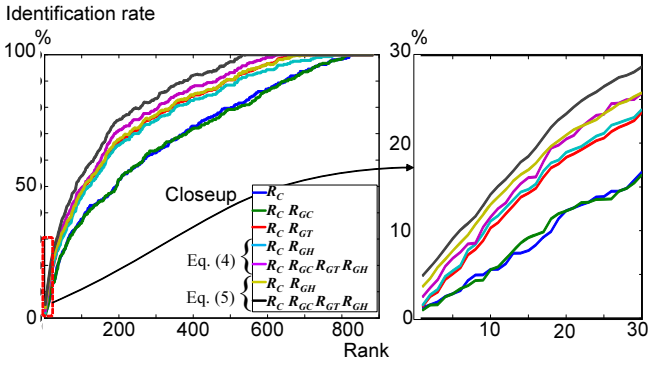


Figure 10: CMC curve of identification using color features and the three group features obtained from the results of people grouping.

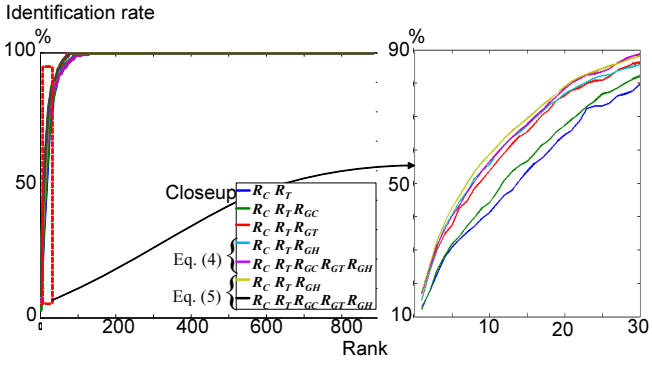


Figure 11: CMC curve of identification using the color and transition interval features, as well as three group features obtained from the results of people grouping.

Figure 9 shows results obtained by using two previous features (i.e. color and transit-time features) and three group features. By comparing the left-hand graphs of Figs. 8 and 9, we note that the transit-time feature is very powerful. In the close-ups (i.e. the right-hand graphs of Figs. 8 and 9), it can be also seen that 1) each of the three proposed group features increased the accuracy of re-identification and 2) the highest accuracy was acquired when all of the three proposed group features were applied.

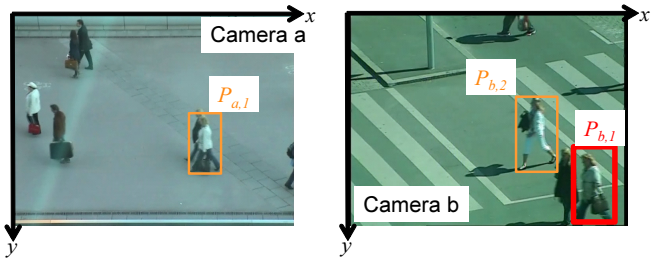
In the results shown in Fig. 9, the numbers of people whose scores of the correct re-identification were ranked within the top ten were 243 and 178 people with and without the group features, respectively. Therefore, the accuracy of re-identification was increased by $\left(\frac{243}{342} - \frac{178}{342}\right) \times 100 \approx 19$ % when applying the group features. The numbers of correctly-identified people (i.e., the numbers of people whose scores of the correct re-identification were ranked in the top one) were 47 and 39 by the proposed method with and without the group features, respectively.

6.2.2. *People Re-identification with Detected Groups of People*

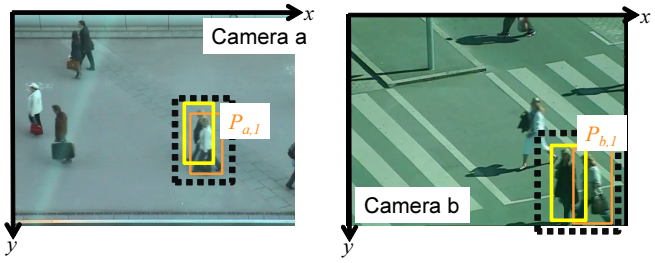
Experiments were conducted with people groups detected by the group detection method proposed in Sec. 4. The groups were detected from the ground truth of people trajectories.

Figure 10 shows results obtained by using the color feature and three proposed group features. In the figure, results with scores (6) and (7) are shown. In Fig. 10, it can be seen that the proposed group features were effective in increasing re-identification accuracy. This effectiveness was, however, lower than that shown in Fig. 8. The drop in effectiveness was caused by errors in group detection. It can be also seen that the drop was suppressed by score (7) rather than score (6).

Figure 11 shows results obtained by using two conventional features (i.e. color and transit-time features) and three group



(a) Without group features



(b) With group features

Figure 12: Example of the effect of the proposed group features. Rectangles with solid and dotted lines indicate detected person and group, respectively. Orange rectangles in cameras a and b were identified with each other.

features. As with the results shown in Fig. 10, the proposed group features successfully improved the accuracy of re-identification.

In Fig. 11, the number of people whose scores of correct re-identification were ranked within the top ten were 239 and 140 people with and without the group features using score (7), respectively. Therefore, the accuracy of re-identification was increased by $\left(\frac{239}{342} - \frac{140}{342}\right) \times 100 \approx 18$ %.

A typical example of the effects of the group features is illustrated in Fig. 12. Without the group features (i.e. Fig. 12 (a)), pedestrian $P_{a,1}$ observed in camera a was identified with $P_{b,2}$ by mistake. This mistake was caused by the change in illumination; more specifically, the clothing color of $P_{b,1}$ who should have been identified with $P_{a,1}$ became darker. With the proposed group features (i.e. Fig. 12 (b)), $P_{a,1}$ was identified correctly with $P_{b,1}$. The numbers of correctly-identified people with and without the group features were 44 and 39, respectively.

The quantitative results of our proposed method are compared with those of other re-identification methods (Table 2). While our proposed method is inferior to the others in the top-one ranking, it outperforms them in the top-ten ranking.

6.2.3. *People Re-identification with People Groups Detected from Tracking Results*

Our final experiments evaluated re-identification in real scenarios where people trajectories and groups were automatically detected. These trajectories were extracted by the method described in [33] after the regions of people were detected by a general human detector using HOG features and the SVM.

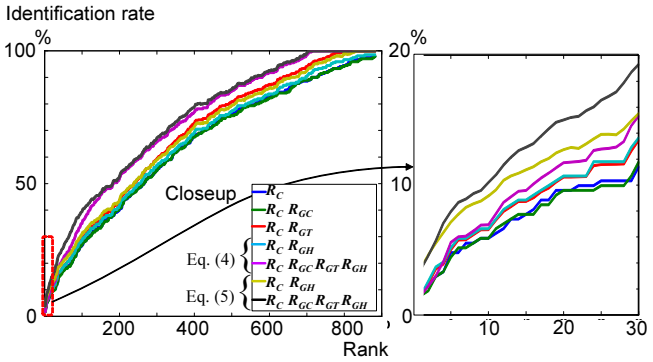


Figure 13: CMC curve of identification using color features and the three group features detected from people tracking results.

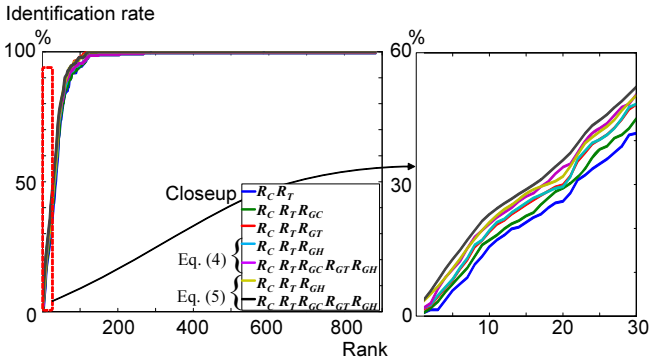


Figure 14: CMC curve of identification using the color and transition interval features, as well as three group features detected from people tracking results.

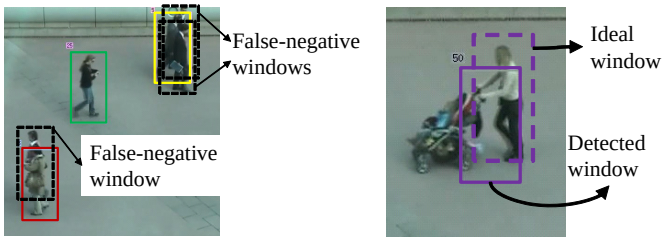
Figures 13 and 14 show results obtained by using the color feature and the three proposed group features (i.e. all features except the transit-time feature) and all features, respectively. The holistic properties of the results were similar to those presented in Figs. 10 and 11, respectively, i.e. that 1) the proposed group features were effective for increasing re-identification accuracy, and 2) score (7) yielded better results in contrast to score (6).

In Fig. 14, the number of people whose scores of correct re-identification were ranked within the top ten were 82 and 55 people with and without the group features using score (7), respectively. Therefore, the accuracy of re-identification was increased by $\left(\frac{86}{342} - \frac{55}{342}\right) \times 100 \approx 10$ %. The numbers of correctly-identified people with and without the group features were 4 and 2, respectively.

Note that the result of re-identification was significantly decreased (i.e. $\frac{86}{342}$) compared with the one using the ground truth of tracking trajectories (i.e. $\frac{239}{342}$), which is shown in Sec. 6.2.2. This decrease may be caused by the poor detection results obtained before people tracking and grouping. Figure 15 shows the examples of the poor detection results. False-negative windows and drift of a detection window are shown in (a) and (b) of the figure, respectively. Compared with the drift, whose negative impact may be suppressed by smoothing tracking trajectories, false-negative windows are critical for all subsequent

Table 2: Comparison between the proposed method and other re-identification methods. All results in this table were obtained by using the ground truth of temporal human windows.

	Hirzer2011 [39]	Hirzer2012 [40]	Wang2014 [41]	Ours
Rank=1	28.1	14.6	28.9	13.7
Rank=10	51.8	42.6	65.5	71.1



(a) Poor detection (b) Poor tracking

Figure 15: Poor human detection and tracking results obtained in our experiments.

processes, including tracking, grouping, and re-identification. While human windows were detected with a simple human detector using the HOG and SVM in our experiments, a more powerful detector (e.g. part-based detector [42]) can detect human windows more precisely. Tracking can be also improved, for example, by recent online discriminative appearance learning [43]. Important future work includes detailed investigation into the relationship between the results of human tracking and our proposed people grouping method.

Results shown in this section have proven that re-identification accuracy was improved by the proposed method even with real tracking trajectories. This conclusion shows the usefulness of the proposed method in real-world applications.

7. Concluding Remarks

This paper proposes methods 1) for detecting people groups by classifying their trajectories and 2) for achieving people re-identification across non-overlapping cameras by employing features obtained from detected groups.

For group detection, the proposed feature represents spatio-temporal relationships between a pair of pedestrians at each moment. The trajectories of the pair are classified as either a "group" or a "non-group".

For people re-identification, three group features are extracted from the group associated with each individual person. One of the group features is based on the color distribution of in-group people, while the other two represent the relationship of in-group people's trajectories and the number of in-group people and its reliability. Our experimental results demonstrated the improvement in people re-identification accuracy using a public dataset; the success rate of people re-identification was improved by 19 %, 18 %, and 10 % with the ground truth of people groups, detected people groups, and extracted people trajectories and groups, respectively.

In group detection, future work includes the further extension of the spatio-temporal features for improving robustness to noisy trajectories. Group features should be also improved. In particular, the color feature of in-group people should be improved by employing more discriminative color representations (e.g. [34],[35]).

We also have other problems in re-identification. For example, for more complex environments with a number of routes among multiple cameras, route prediction using priors over exits and entrances among the cameras is useful as proposed in [31, 13]. The effectiveness of global optimization for multi-object tracking has been also demonstrated in related work; within a field of view [33, 44] and across cameras [13, 45]. All of these tracking techniques should be integrated for more improvement.

References

[1] Z. Yucel, T. Ikeda, T. Miyashita, and N. Hagita, "Identification of mobile entities based on trajectory and shape information," *IROS*, 2011. 2

[2] S. Ali and M. Shah, "Floor Fields for Tracking in High Density Crowd Scenes," *ECCV*, 2008. 1

[3] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, "Data-driven Crowd Analysis in Videos," *ICCV*, 2011. 1

[4] A. C. Gallagher and Tsuhan Chen, "Understanding Images of Groups of People," *CVPR*, 2009. 1

[5] G. Wang, A. Gallagher, J. Luo, and D. Forsyth, "Seeing People in Social Context: Recognizing People and Social Relationships," *ECCV*, 2010. 1

[6] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical Review E*, Vol.51, No.5, pp.4282-4286, 1995. 2

[7] R. Mehran, A. Oyama, and M. Shah, "Abnormal Crowd Behavior Detection using Social Force Model," *CVPR*, 2009. 2

[8] P. Scovanner and M. F. Tappen, "Learning Pedestrian Dynamics from the Real World," *ICCV*, 2009. 2

[9] 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. 2

[10] S. Pellegrini, A. Ess, and L. van Gool, "Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings," *ECCV*, 2010. 2

[11] M.-C. Chang, N. Krahnstoever, and W. Ge, "Probabilistic Group-Level Motion Analysis and Scenario Recognition," *ICCV*, 2011. 2

[12] Z. Qin and C. R. Shelton, "Improving Multi-target Tracking via Social Grouping," *CVPR*, 2012. 2

[13] A. Alahi, V. Ramanathan, and L. Fei-Fei, "Socially-aware Large-scale Crowd Forecasting," *CVPR*, 2014. 2, 10

[14] J. Shao, C. Change Loy, and X. Wang, "Scene-Independent Group Profiling in Crowd," *CVPR*, 2014. 2

[15] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and Where are you going?" *CVPR*, 2011. 2, 3, 7, 8

[16] 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. 4

[17] 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. 4

[18] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux, "Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences," *ACM/IEEE International Conference on Distributed Smart Cameras*, 2008. 2
[19] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and Appearance Context Modeling," *ICCV*, 2007. 2
[20] N. Dalal and B. Berg, "Histograms of Oriented Gradients for Human Detection" *CVPR*, 2005. 2
[21] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person Re-Identification by Symmetry-Driven Accumulation of Local Features," *CVPR*, 2010. 2, 4
[22] O. Javed, K. Shafique and M. Shah, "Appearance Modeling for Tracking in Multiple Non-Overlapping Cameras," *CVPR*, 2005. 4
[23] A. Gilbert and R. Bowden, "Tracking Objects Across Cameras by Incrementally Learning Inter-camera Colour Calibration and Patterns of Activity," *ECCV*, 2006. 4
[24] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," *ECCV*, 2008. 2
[25] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," *BMVC*, 2010. 2
[26] C. H. Kuo, C. Huang and R. Nevatia, "Inter-camera Association of Multi-target Tracks by On-Line Learned Appearance Affinity Models," *ECCV*, 2010. 2, 4
[27] W. S. Zheng, S. Gong and T. Xiang, "Person Reidentification by Probabilistic Relative Distance Comparison," *CVPR*, 2011. 2, 4
[28] V. Kettnaker and R. Zabih, "Bayesian Multi-Camera Surveillance," *CVPR*, 1999. 4
[29] O. Javed, Z. Rasheed, K. Shafique and M. Shah, "Tracking Across Multiple Cameras with Disjoint Views," *ICCV*, 2003. 4
[30] D. Makris, T. Ellis and J. Black, "Bridging the gaps between cameras," *CVPR*, 2004. 4
[31] N. Ukita, "Probabilistic-Topological Calibration of Widely Distributed Cameras," *Machine Vision and Applications*, Vol.18, No.3, pp.249–260, 2007. 4, 10
[32] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and Where are you going?" *CVPR*, 2011.
[33] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-Optimal Greedy Algorithms for Tracking a Variable Number of Objects," *CVPR*, 2011 7, 9, 10
[34] W.-S. Zheng, S, Gong, and T. Xiang, "Associating Groups of People," *BMVC*, 2009. 2, 10
[35] Y. Cai, V. Takala, and M. Pietikinen, "Matching Groups of People by Covariance Descriptor," *ICPR*, 2010. 2, 10
[36] M. Yang, Y. Wu, and S. Lao, "Intelligent Collaborative Tracking by Mining Auxiliary Objects," *CVPR*, 2006. 2
[37] A. Okada, Y. Moriguchi, N. Ukita, and N. Hagita, "People Groping by Spatio-Temporal Features of Trajectories," *IAPR International Conference on Machine Vision Applications*, 2013. 2
[38] V. Ramesh and P. Meer, "Real-Time Tracking of Non-Rigid Objects using Mean Shift" *CVPR*, 2000. 4
[39] M. Hirzer, C. Beleznai, P. M. Roth and H Bischof "Person Re-Identification by Descriptive and Discriminative Classification" *Scandinavian Conference on Image Analysis* , 2011. 6, 7, 10
[40] M. Hirzer, P. M. Roth, M. Kostinger, H. Bischof, "Relaxed pairwise learned metric for person re-identification," *ECCV*, 2012. 10
[41] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person Re-identification by Video Ranking", *ECCV*, 2014. 10
[42] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.32, No.9, pp.1627–1645, 2010. 10
[43] S.-H. Bae and K.-J. Yoon, "Robust Online Multi-Object Tracking based on Tracklet Confidence and Online Discriminative Appearance Learning," *CVPR*, 2014. 10
[44] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," *CVPR*, 2008. 10
[45] C.-H. Kuo, C. Huang, and R. Nevatia, "Inter-camera Association of Multi-target Tracks by On-Line Learned Appearance Affinity Models," *ECCV*, 2010. 10