IPSJ Transactions on Computer Vision and Applications Vol.4 87–99 (July 2012)

[DOI: 10.2197/ipsjtcva.4.87]

Research Paper

Group Context-aware Person Identification in Video Sequences

Haruyuki Iwama[1,a) Yasushi Makihara[1] Yasushi Yagi[1]

Received: August 13, 2011, Accepted: April 23, 2012, Released: July 25, 2012

Abstract: The importance of person identification techniques is increasing for visual surveillance applications. In social living scenarios, people often act in groups composed of friends, family, and co-workers, and this is a useful cue for person identification. This paper describes a method for person identification in video sequences based on this group cue. In the proposed approach, the relationships between the people in an input sequence are modeled using a graphical model. The identity of each person is then propagated to their neighbors in the form of message passing in a graph via belief propagation, depending on each person's group affiliation information and their characteristics, such as spatial distance and velocity vector difference, so that the members of the same group with similar characteristics enhance each other's identities as group members. The proposed method is evaluated through gait-based person identification experiments using both simulated and real input sequences. Experimental results show that the identification performance is considerably improved when compared with that of the straightforward method based on the gait feature alone.

Keywords: group context, CRF model, person identification, belief propagation

1. Introduction

Person identification techniques are becoming increasingly important for visual surveillance and monitoring. Methods of person identification based on a variety of biometric-based cues such as the person's face [45] and gait [17], [25] have been developed, mostly from the viewpoints of discrimination capability and stability. In all of these techniques, however, the identification performance often decreases due to changes in the condition of individuals and their surroundings, and misidentification may consequently occur, particularly in real environments. Also, as the number of individuals increases, the misidentification rate generally increases due to the growth in ambiguity.

An example of misidentification in a straightforward gait-based identification framework is shown in **Fig. 1**. Because the gait feature of probe #1 has changed slightly from that in gallery #a (the same subject), particularly in the arm swing, the feature similarities between probe #1 and gallery #a are smaller than those between the probe and other galleries (e.g., #x and #y).

However, it is useful to take into account the characteristics of human activities to provide context for person identification. In social living situations, people often act in groups, as shown in **Fig. 2**, which are composed using social relationships in most cases, such as family, friends, and co-workers. It is assumed, therefore, that a person is likely to be observed close to other persons of the same group in a video sequence. This observation serves as a contextual cue to improve the identification performance for individuals, i.e., the identity of each person can be inferred not only from their biometric cues alone, but also from the identities of other people in their neighborhood and their group affiliations.

This kind of group context can be used in many places, such as amusement or theme parks, airports, factories, and schools, where many tasks based on person identification techniques are performed. Examples of these tasks include the detection of a lost child in an amusement park, the detection of intruders who enter the amusement park, airport, or factory without passing regular entrance procedures, and the safety confirmation (or attendance checking) of children at the entrance to the school (in particular, there is a rule for going to school in a group composed of community children for almost all Japanese elementary schools). The group context is also useful for person re-identification across multiple non-overlapping network cameras.

Recently, some works have integrated such kind of group context with face-based person identification in photo collection to improve the identification performance [8], [20], [24]. In these methods, the person-to-person relations are modeled in terms of co-occurrence among persons in photos as group prior. Differently from the photo collection, however, a group is often observed with non-group members at a time in the video sequences of surveillance camera and the spatial relations among them are dynamically changed with time. Therefore, the identity of individual should be inferred not only from the viewpoint of co-occurrence among persons, but also from that of behavioral differences among persons through the sequence.

In this paper, we propose a group context-aware framework for person identification in video sequences that unifies the group context with the individual biometric cues. In terms of the group (inter-person) context for person identification, the proposed method take the behavioral differences such as spatial dis-

[1] Osaka University, Ibaraki, Osaka 567–0047, Japan

[a)] iwama@am.sanken.osaka-u.ac.jp

tance and the differences of walking speed and direction among persons through the sequence into account, and this is a primal contribution of this work.

Our key observation is as follows. We assume the group walking situation in a video sequence which includes two different groups and an unregistered person as shown in Fig. 2 and consider the identity of the probe #1 within the group context. We assume that the gallery subjects #a, #b, #c, #d, and #e belong to the group A and the #v, #w, and #x belong to the other group X, as shown in **Fig. 3**. Also, we assume that the identity of probe #6 is not matched with any of gallery subjects and probe #2, probe #3, probe #4, and probe #5 are confidently inferred to be #b, #c, #v, and #w, respectively, while the identity of probe #1 is mis-inferred to be #x (a member of group X) as shown in Fig. 1. If only co-occurrence is used as group context, the identity of probe #1 can be inferred from not only the identities of probes #2 and #3 as #a, but also those of probes #4 and #5 as #x based on their group affiliation information. Consequently, the identity of probe #1 possibly remains to be mis-inferred as #x in this case. In addition, the identity of probe #6 (unregistered person) which appears in the scene from the middle of the sequence is possibly mis-inferred from the identities of other subjects.

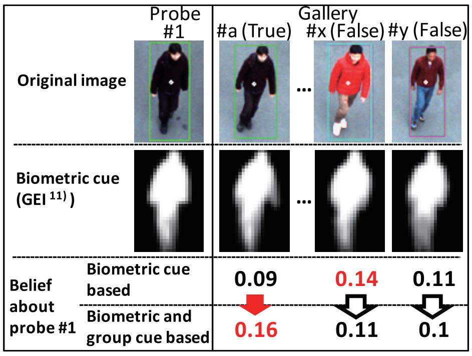


Fig. 1 Biometric cues and belief.

On the other hand, focusing on the behavioral relations among probe subjects through a sequence, we see that probe #6 obviously walks at a distant from all the other subjects, and thus probe #6 can be regarded as an independent subject from all the other subjects. At the same time, we also see that there exists an apparent difference of walking speed between the group of #1, #2 and #3 and the group of #4 and #5. Accordingly, the weights of the inference cues from the probes #2 and #3 come to be able to be distinguished from those of the inference cues from the probes #4 and #5. The identity of probe #1 as #a is then definitely enhanced and the mis-identification of #1 can be recovered as a result.

We realize this idea in the form of a message passing in a graph, where each node corresponds to each probe subject and each edge corresponds to the relationship between each pair of probe subjects. In the iteration of the message passing process, the identity confidence for each probe subject is propagated to the identities of the surrounding probe subjects based on their biometric cues and group information, so that the same group members with sim-

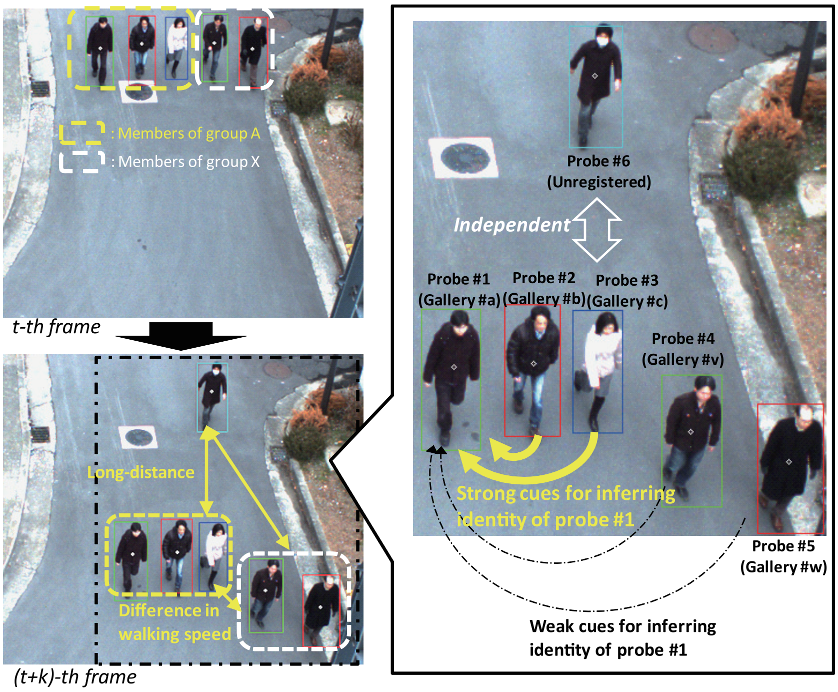


Fig. 2 An example of group walking in video sequence.

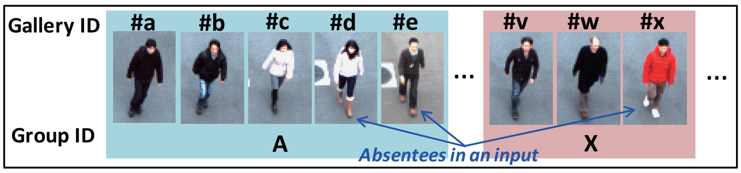


Fig. 3 An example of group affiliation.

ilar characteristics (spatial proximity and similar velocity vector) enhance each other's identities.

The remainder of this paper is organized as follows. Section 2 introduces related work. Section 3 describes our problem formulation, and the detailed implementation is described in Section 4. Section 5 presents experimental testing of the effectiveness of the proposed method and our discussions are presented in Section 6. Finally, conclusions are drawn and future work is proposed in Section 7.

2. Related Work

In recent years, many researchers have paid considerable attention to the use of context in traditional computer vision problems, such as object detection and categorization, action recognition, and person identification, to improve performance. In this section, we review such context-based approaches briefly.

Object detection/recognition: In the task of object detection, context is mainly used to limit the area in which objects are likely to appear, to reduce false positives. Torralba et al. [34] exploited a global image feature called *gist*, which was a low level representation of an image. Hoiem et al. [16] used the 3D geometrical information of the scene, such as the surfaces, the camera viewpoint, and object positions and sizes as context. While these approaches focused on global scene information, some works instead focused on local information [29], [46]. In Ref. [29], the spatial relations between an object of interest and its surroundings are modeled as a visual context feature composed of geometrical and textural features, and are used to extract prior instances of the object's presence from a scene. In this method, object co-occurrence and bottom-up saliency were also used for context. Heitz and Koller [15] modeled the spatial relationships between an object ("thing") and the surrounding regions ("stuff"), which were the results of unsupervised image clustering, as the *TAS model* ("thing" and "stuff" model). The effect of the use of context in object detection is empirically evaluated in Ref. [7].

In recent works in object recognition [10], [27], [30], inter-object relationships, such as co-occurrence, relative location, and scale, are used as context to resolve object appearance ambiguities. Besides those given above, a number of context-based techniques have been discussed and summarized in Refs. [1], [9], [22].

Action/Interaction recognition: Many of these works have indicated that modeling of human-object relationships is useful for the understanding of human actions/interactions [12], [13], [37], [40], [41]. Wu et al. [37] proposed an object-use based action recognition framework, in which the relationships between an action and the object-use events data during that action were used as context, and the relationships were learned automatically using RFID sensors and a common-sense knowledge database. Yao and Fei-Fei [40], [41] proposed two types of approach; one is based on a model of the spatial relationships between human poses (positions of body parts) and objects [40], while the other is based on a structured appearance feature called "*Grouplet*" [41], for recognition of human-object interactions. Marszalek et al. [23] used action-scene relationships as context, which were derived automatically from training videos using video scripts, and in Ref. [18], both scene and object features are integrated with the action features. In Ref. [19], human-human interactions are the focus, and it was shown that spatio-temporal observations of the surrounding people which represent the actions of the surroundings helped with action recognition. In a similar manner, Choi et al. [35] used the spatio-temporal distribution of multiple people, which included their relative motion and locations, to classify collective activities, such as "queueing" and "talking."

Person identification: The automatic annotation, organization, and retrieval of still images, in particular in personal digital photo collections, have been active research topics in recent years. In these tasks, face-based person identification is crucially important and many context-aware methods have been developed. As mentioned in Ref. [32], there are three types of context information: appearance-based, metadata-based, and logic-based context information. In Refs. [32], [43], appearance-based context, such as body parts and clothes, are combined with facial features. Stone et al. [33] used metadata-based context derived from the social network *Facebook*. Gallagher and Chen [8] used co-occurrence between each person as a logic-based context, which indicated how often a pair of faces appeared together in images. In some works [20], [24], such co-occurrence of persons is also integrated together with other types of contexts such as events (time stamp) and locations which are rather peculiar to the field of photo collection.

In a scenario of person re-identification across multiple non-overlapping cameras, Zheng et al. [36] and Cai et al. [4] proposed a solution to the problem of associating groups of people between the different camera views and demonstrated that group information helped to resolve the ambiguities in individual appearances. For person identification, they simply combined group cues with individual cues in the form of a weighted sum of each score. They considered a group as a small number of people walking in close proximity in spatial domain, and quantified the group cue by measurement of the spatial appearance features. In these methods, although their group representations are designed to be invariant to positional changes of the group members between the different camera views, the fluctuations in the numbers of observed members, which are caused by absentees, isolation of group members, or the proximities of non-group members, lead to significant changes in the spatial appearance of the group. Accordingly, the effectiveness of the group cue is degraded. For instance, if a certain group is composed of 5 members in the gallery image and only 3 members of the group are observed in a probe image, the observed group tends to be matched with other groups composed of 3 members by mistake. Furthermore, these methods do not consider the behavioral relations among persons such as velocity vector difference through the walking.

Our work is inspired by the related work described above and we propose a unified framework for the person identification problem in video sequences, in which group context is integrated with individual biometric observations by using CRF model. Though, the CRF-based framework is similar to the existing context-assisted person identification schemes formulated by MRF/CRF model such as Ref. [8], the major difference of this work is that we use the behavioral relations as group context including spatial distance and velocity vector difference among persons through the video sequences, while existing frameworks used co-occurrences among persons as group context. This also differentiate the proposed method from other group context-based person re-identification methods such as Refs. [4], [36]. Though, similar kinds of behavioral relations are utilized for the problem of trajectory prediction of pedestrians in some works [28], [39] and these are also related to our work, we apply such kind of context to person identification problem, and this is a primal contribution of this paper.

3. Group Context-aware Person Identification in Video Sequences

We regard the person identification problem as a many-to-many matching problem for a given image sequence. The task we consider is assignment of a registered person's label to each person that is observed in an input sequence. In this work, *group* is not only explicitly-defined as a unit of people that is composed on the basis of social relations, such as family, friends, and co-workers, but is also implicitly-defined as the result of manual or automatic clustering. Then, the following prerequisites are assumed.

- Each registered person belongs to one of the predefined groups.
- Group affiliation and biometric cues of each registered person are given as gallery data in advance.
- Segmentation and tracking of each subject in an input sequence are obtained in advance.
- Each registered person appears at most once and is likely to appear with group members, in detail, in close vicinity and with similar velocity in an input sequence.

Also, the following conditions are considered:

- Unregistered persons also appear in an input sequence randomly.
- Absence and isolation of a registered person in an input sequence are allowed.

Note that for a registered person who does not belong to any group, an expedient group whose only member is that person is defined, while the "*unregistered*" label is only assigned to actual unregistered persons.

3.1 Problem Formulation

In the labeling task, we must take account of the relationship between the observed characteristics of each probe, such as spatial position and velocity, and the group affiliation of each label in addition to the biometric cue for each probe. The preferred label assignment, therefore, is one where the same group members are likely to appear in a group, and the biometric cues of each probe are given substantial consideration.

We use a pair-wise CRF (conditional random field) model in a manner similar to Refs. [5], [6] for our labeling problem. Let each node in a graph represent a person who appears in an input sequence, and the label for the i-th node x_i represents the index of the registered person or "*unregistered*" label. The label set is defined as $\mathbf{L} = \{l_1, l_2, \cdots, l_n, l_{un}\}$, where l_k $(k = 1, 2, \cdots, n)$ is the label of the k-th registered person and l_{un} is the label for an unregistered person. A mapping from an individual label to a group is then defined as $g(l_k) \in \mathbf{G}$, where $\mathbf{G} = \{G_1, G_2, \cdots, G_{n_G}, G_{un}\}$ is a group identifier set, G_k $(k = 1, 2, \cdots, n_G)$ is a group identifier for each registered person, and G_{un} is a identifier for an expedient group for any unregistered person.

The graphical representation is shown in **Fig. 4**. In this example, there are seven probe subjects in an input sequence, and each node is connected to neighbor nodes which correspond to the persons within a set spatial distance $d_{\max}$, which is set to 3 [m] in this work, in the input sequence. Also, as described later in detail, all of the nodes are connected by a factor node, which controls the exclusion of each label.

We then let $\mathbf{x}$ be the label assignment for all the nodes and $\mathbf{y}$ be the set of biometric cues for all the nodes, and then the conditional probability of an assignment $\mathbf{x}$ is formulated as,

$$P(\mathbf{x}|\mathbf{y}) \propto \left\{ \prod_i \phi_i(x_i) \prod_{j \in N(i)} \psi_{i,j}(x_i, x_j) \right\} E(\mathbf{x}), \qquad (1)$$

where $\phi_i(x_i)$ is the local evidence term for node i, $\psi_{i,j}(x_i, x_j)$ is the compatibility term between node i and node j, and $N(i)$ represents a neighbor node set around the node i. $E(\mathbf{x})$ is a label exclusion term, which becomes zero if any registered person label is used more than once and is otherwise one (the label for unregistered persons l_{un} can be used more than once).

The local evidence ϕ_i is defined based on the observed biometric cues for each person. The compatibility $\psi_{i,j}$ corresponds to the group context. The magnitude of the compatibility, therefore, depends on a pair of group identifiers for the label that is assigned to the i-th person and the j-th person and their spatial distance and velocity vector difference, which are defined in Section 4.2 in detail.

3.2 Approximate Solution via Loopy Belief Propagation

LBP (Loopy belief propagation) [42] is used as an approximate

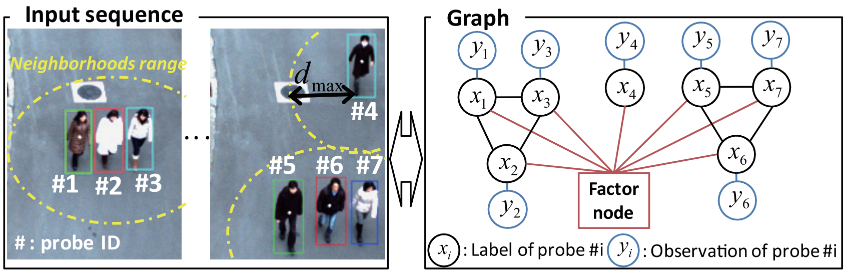

Input sequence
Neighborhoods range
#1 #2 #3
# : probe ID
$d_{\max}$
#4
#5 #6 #7
Graph
y_1 y_3 y_4 y_5 y_7
x_1 x_3 x_4 x_5 x_7
x_2 x_6
y_2 y_6
Factor node
x_i : Label of probe #i
y_i : Observation of probe #i

Fig. 4 Our graphical representation: An example of the input sequence (left) and the corresponding graph (right).

solver to find the assignment $\mathbf{x}$ that maximizes the probability $P(\mathbf{x}|\mathbf{y})$. Ignoring the exclusion term $E(\mathbf{x})$ at this stage, the message $m_{ij}(x_j)$ from node i to node j for each label is defined as,

$$m_{ij}(x_j) \propto \sum_{x_i} \psi_{i,j}(x_i, x_j)\phi_i(x_i) \prod_{k\in N(i)\setminus j} m_{ki}(x_i). \tag{2}$$

The belief $b_i(x_i)$ at the node i for each label is found as a marginal probability by gathering messages from its neighbor nodes and from the local evidence,

$$b_i(x_i) = k\phi_i(x_i) \prod_{j\in N(i)} m_{ji}(x_i), \tag{3}$$

where k is a normalization constant (summation of belief is normalized to 1). The label assignment of the node i is,

$$x_i^* = \arg\max_l b_i(x_i = l). \tag{4}$$

Note that each message is initialized to 1, normalized local evidence is given as the initial belief value, and that the upper limit of iteration of LBP was set to 10 in this work.

3.3 Handling the Exclusion Term

The label exclusion term $E(\mathbf{x})$ is defined such that it forbids the use of a registered person's label more than once, i.e., to suppress the use of the label l_k if another node already has high belief about l_k. Since the label exclusion term is a global function, we can represent it using a factor node that is connected to all of the nodes. In terms of the message passing scheme, the message from a factor node f to a node i is,

$$m_{fi}(x_i = l) \approx \prod_{t\in S\setminus i} \left(1 - m_{tf}(x_t = l)\right), \tag{5}$$

where S is the set of all nodes and m_{tf} is defined as,

$$m_{tf}(x_t = l) = \left(b_t(x_t = l)\right)^{\alpha}, \tag{6}$$

where α is the message attenuation parameter, and is set to 2 in this work.

Actually, label exclusion via the above message does not completely control the one-time use of the label of a registered person, because the belief of each node for a certain label does not always become 1.0 after message passing. To complete the exclusion control, we therefore execute the *Greedy Algorithm* in terms of the belief score for finalization of the label assignment after the convergence of LBP.

4. Implementation

4.1 Local Evidence

4.1.1 Label of Registered Person

An observed biometric feature of each person, such as their face or gait, is a crucial clue in itself for person identification, as numerous previous works have demonstrated. We therefore use such a feature as the local evidence for the label of a registered person and it define as,

$$\phi_i(x_i = l_k) \propto p(x_i = l_k|\mathbf{y}_i), \tag{7}$$

where $\mathbf{y}_i$ is the observed feature vector of the i-th person and l_k is the label of the k-th registered person. Actually, we regard the prior $p(x_i = l_k)$ as constant for all k, then Eq. (7) can be described as,

$$\phi_i(x_i = l_k) \propto p(\mathbf{y}_i|x_i = l_k). \tag{8}$$

The probabilistic observation models of the feature vector for each label of each registered person are constructed from gallery feature vectors such as the Gaussian distribution model in advance.

However, since the gallery feature vector of each registered person cannot be captured a number of times, but at most once or twice in most cases, such as real surveillance scenarios, it is difficult to construct the probabilistic model properly in practice. For instance, in the case where only one gallery feature vector is given, it makes no sense to construct a Gaussian distribution as it is. In such a case, therefore, we regard the variation of each feature vector element to be common for all elements and for all persons, and we set the probability model to be,

$$p\,(\mathbf{y}_i|x_i = l_k) \propto \exp\left(-\frac{{D_{i,k}}^2}{2}\right) \tag{9}$$

$$D_{i,k} = \frac{|\mathbf{y}_i - \bar{\mathbf{y}}_k|}{\sigma}, \tag{10}$$

where $\bar{\mathbf{y}}_k$ is the average vector of the gallery of the k-th registered person and σ is standard deviation of the feature vector element, which is given as a hyper-parameter.

4.1.2 Label of Unregistered Person

For the label of an unregistered person, the model cannot be constructed, because the feature vector which represents the "*unregistered person*" can be never captured as gallery data. We thus give a constant value C_{un} as the local evidence for the label l_{un} instead,

$$\phi_i(x_i = l_{un}) = C_{un}. \tag{11}$$

4.2 Compatibility

The compatibility score for a pair of labels is required to be high only if the group affiliations of the two labels are the same and the corresponding persons appear in close proximity and with similar velocities in an input sequence.

We quantify this using two terms: the distance term E_d and the velocity term E_v. One is based on the spatial distance between the two persons and the other is based on the velocity vector difference between them in the world coordinates. Compatibility for a pair of labels, l_s and l_t, is then defined as,

$$\psi_{i,j}(x_i = l_s, x_j = l_t) \propto \begin{cases} C & (l_s = l_{un} \text{ or } l_t = l_{un}) \\ (1-\delta_{l_s,l_t})\left(E_d\left(d_{i,j}\right) E_v\left(v_{i,j}\right) \delta_{g(l_s),g(l_t)} + C\right) & (\text{otherwise}) \end{cases}, \tag{12}$$

where δ is the *Kronecker delta*, $d_{i,j}$ and $v_{i,j}$ are the spatial distance and the velocity vector difference between the i-th person and the j-th person in the world coordinates, and C is a constant value. The distance term $E_d(d_{i,j})$ and the velocity term $E_v(v_{i,j})$ are designed as,

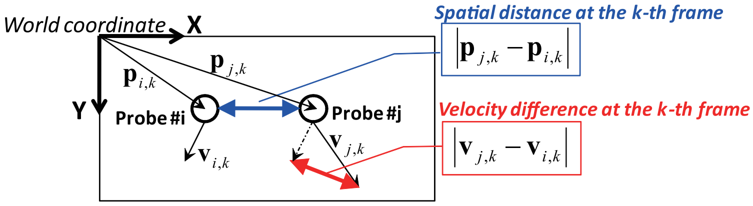

World coordinate
X
Y
$\mathbf{p}_{i,k}$
$\mathbf{p}_{j,k}$
Probe #i
Probe #j
$\mathbf{v}_{i,k}$
$\mathbf{v}_{j,k}$
Spatial distance at the k-th frame
$|\mathbf{p}_{j,k} - \mathbf{p}_{i,k}|$
Velocity difference at the k-th frame
$|\mathbf{v}_{j,k} - \mathbf{v}_{i,k}|$

Fig. 5 Spatial distance and velocity vector difference between a pair of probe subjects (#i and #j) at the *k*-th frame.

$$E_d\left(d_{i,j}\right) = -\frac{\left(d_{i,j} - d_{\max}\right)}{d_{\max} - d_{\min}} \tag{13}$$

$$E_v\left(v_{i,j}\right) = -\frac{\left(v_{i,j} - v_{\max}\right)}{v_{\max} - v_{\min}}, \tag{14}$$

where $d_{\max}$ and $d_{\min}$ are the upper and lower limits of the spatial distance ($d_{\max}$ is equal to the one described in Section 3.1), and $v_{\max}$ and $v_{\min}$ are these limits for the velocity difference. In all of our experiments, the parameters are set as $C = 0.1$, $d_{\max} = 3$ [m], $d_{\min} = 0.5$ [m], $v_{\max} = 1$ [km/h], and $v_{\min} = 0$ [km/h].

To use the spatial distance and the velocity information in the world coordinates, we need to estimate them from an input video sequence. One of the most reasonable ways to do this is a method based on ground constraints. If the homography correspondences between the ground plane in the world coordinates and the image plane are calibrated in advance, the foot's position trajectory on the ground plane can be estimated from the bottom coordinate of the corresponding person in the images. Subsequently, we can derive the spatial distance $d_{i,j}$ and the velocity vector difference $v_{i,j}$ between the i-th and the j-th person as follows,

$$d_{i,j} = \max_{t_s \le k \le t_e} \left|\mathbf{p}_{i,k} - \mathbf{p}_{j,k}\right| \tag{15}$$

$$v_{i,j} = \max_{t_s \le k \le t_e} \left|\mathbf{v}_{i,k} - \mathbf{v}_{j,k}\right|, \tag{16}$$

where $\mathbf{p}_{i,k}$ is the smoothed 2D position in world coordinates of the i-th person at the k-th frame, $\mathbf{v}_{i,k}$ is the smoothed 2D velocity vector of the i-th person at the k-th frame (both are illustrated in **Fig. 5**), and t_s and t_e are the first and last frame identifiers for the frames where the i-th person and the j-th person appear together in an input video. Note that, in this case, if a pair of persons does not appear together in any frame, they are not considered to be in the same neighborhood as each other.

4.3 Seed Node Selection

While the ambiguity of a biometric-based identity is solved by messages, it is desirable that a node with confident local evidence for a certain label is then unchanged by messages, to avoid unreasonable belief variation.

For this purpose, we fix the labels of the nodes to such persons with confident local evidence at the first stage. We denote this label-fixed node and the fixed label as the *seed node* and *seed label*, respectively. The seed node is decided using the thresholding mahalanobis distance (Eq. (10)) with threshold T_s. More specifically, when only the k-th node has a lower mahalanobis distance than T_s about a certain label l, the k-th node and the label l are regarded as the seed node and the seed label. We then set the belief of the other nodes about the label l to 0 and set the messages to the k-th node from the other nodes and the belief of the k-th node as

$$m_{ik}(x_k = l_j) = b_k(x_k = l_j) = \delta_{l_j,l}, \tag{17}$$

where δ is the *Kronecker delta*. Also, we set the message from the seed node (the k-th node) to the other node as,

$$m_{ki}(x_i = l_j) \propto \psi_{k,i}(x_k = l, x_i = l_j). \tag{18}$$

Note that the local evidence for the seed label l is regarded as 1 ($\phi_k(x_k = l_k) = \delta_{l,l_k}$) in this equation. In addition, in the message passing process, if the belief of a node about a certain label reaches a predefined criterion, which is set to 0.9 in this work, we set the node to be the seed node at that stage.

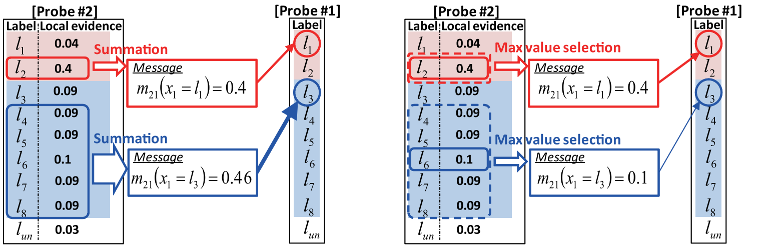
4.4 Relaxation of a Biased Message Caused by an Imbalance in the Number of Group Members

In the presence of an imbalance in the number of group members, the message magnitude is biased by this imbalance. We illustrate this with examples of the gallery set and the situation in an input video as shown in **Fig. 6**.

Consider the messages from probe #2 (the true label is l_2) to probe #1 (the true label is l_1) at the first iteration. As long as the local evidence of probe #2 about the label l_2 is higher than the local evidence about the other labels, the message to enhance the belief for the label l_1 at probe #1 is preferred, because probe #1 and probe #2 belong to the same group G_1 in this situation. For simplicity, suppose that the compatibility between probes #1 and #2 is approximated to $\psi_{2,1}(x_2 = l_s, x_1 = l_t) = (1 - \delta_{l_s,l_t})\,\delta_{g(l_s),g(l_t)}$, where δ is the *Kronecker delta*. The message about the label l_k is then described as,

$$m_{21}(x_1 = l_k) = \sum_{l \in L_{g(l_k)} \backslash l_k} \phi_2\,(x_2 = l), \tag{19}$$

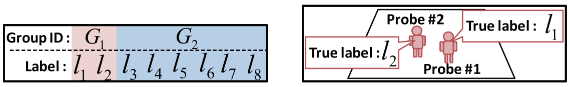
where L_G is a label set of group G members, defined as $L_G = \{l | g(l) = G\}$. Consequently, the magnitude of the message depends not only on the local evidence $\phi_2\,(x_2 = l)$, but also on the number of group members $|L_{g(l_k)}|$. This may cause an undesired reversal of the message magnitude when the local evidence of probe #2 is given as shown in **Fig. 7** (a). In this case, because the number of group G_2 members is higher than that of the group G_1 members, the summation of the local evidence for the labels of group G_2 becomes higher than that for the labels of group G_1, despite the fact that the local evidence for the label l_2 is the highest, and that the evidence about each label of group G_2 is low. As a result, the message about the label l_3 becomes higher than that about the label l_1, as illustrated in Fig. 7 (a).



(a) Reversal effect in the standard form

(b) Exclusion of within-group labels via the max selection

Fig. 7 An example of the reversal effect of the message magnitude caused by the bias for the number of group members in the standard message form, and the concept of exclusion of within-group labels via the max selection as a solution to the problem.



(a) Gallery set

(b) Input situation

Fig. 6 An example of the gallery and the input situation.

To avoid such undesirable message effects, we propose an alternative message form based on the exclusion of within-group labels via a max selection scheme in message formula (Eq. (2)) as,

$$m_{ij}^{max}\left(x_j = l_k\right) \propto \sum_{g\in\mathbf{G}} \max_{l\in L_g} \psi_{i,j}\left(x_i = l, x_j = l_k\right)\phi_i\left(x_i = l\right) \prod_{k\in N(i)\setminus j} m_{ki}\left(x_i = l\right) \quad (20)$$

In this form, the number of group members no longer influences the message magnitude, because we exclude all of the labels of the group G_k other than the within-group maximum in marginalization of the message, as illustrated in Fig. 7 (b). The intuitive interpretation of this form is that we model a person-to-group relationship in this message form, rather than a person-to-person relationship, i.e., from the standpoint of probe #1, the magnitude of the message from probe #2 is based not on "who is the probe #2," but "to what group does the probe #2 belong."

5. Experiment

In this experiment, the effectiveness of the proposed method was examined first using real video sequences, and the performance for a massive data set was then explored using simulation data sets. We chose gait as the biometric cue and used GEI [14] (22 pixels × 32 pixels) as the gait feature, because it achieved the best performance in Ref. [26]. The group affiliation of each gallery is manually assigned in these experiments. The performance of the proposed method was compared with straightforward local evidence-based labeling via the *Greedy Algorithm*. We evaluated the labeling accuracy R_l as $R_l = \frac{N_t}{N_p}$, where N_p and N_t were the probe number and the correctly labeled probe number, respectively.

5.1 Experiment with Real Image Data

We conducted the experiments for two types of real image sequences, one is captured at our campus for preliminary performance evaluation, and the other is obtained from the surveillance cameras installed in a Japanese elementary school.

5.1.1 Preprocessing

We obtained the blob information of each subject in image sequences as follows. First, the foreground regions are extracted via graph-cut-based segmentation [2] in conjunction with background subtraction. Second, each blob is extracted from the foreground regions based on connectivity and the blob statistics, such as area, gravity position, and bounding box are then obtained for each blob. In this process, blobs of different persons may be merged in case where a person is closely-attached to the other person. To avoid such merge, we set the upper limits for the height and width of bounding box respectively, and we split the blob based on the limits if necessary. For example, if the blob has larger height than its upper limit, we count the number of foreground pixels for each height and split the blob at the height with the minimum pixel count within a certain height range.

As for tracking, each bounding box in the current frame is corresponded to the nearest bounding box in the next frame, and the foot's position trajectory of each individual is obtained as a result [*1]. Finally, the gait feature of each individual is extracted from the corresponding blob sequence. The bounding box and trajectory contain errors in some degree, and these also decrease the quality of gait feature.

Note that we omitted the occlusion situation among persons in this experiment, because we focus on the evaluation of the effectiveness of the proposed inference algorithm.

5.1.2 Preliminary Evaluation

Gallery and probe data set: We used an input sequence (640 pixels × 480 pixels/15 fps/bmp format) which includes 18 probe subjects, as shown in **Fig. 8**. In this sequence, the walking directions of all subjects are almost the same. Then, we arranged the gallery set, which includes 20 subjects, as shown in **Fig. 9**. In this setting, the clothes of gallery members #c, #h, and #k are changed at the time of the input sequence to make the person identification problem setting more difficult, which is intentional so that biometric cues alone cannot perfectly identify the subjects. Three absentees (#x, #y, and #z) and one unregistered person (probe #18) are arranged to demonstrate that the proposed method can handle such situations.

In this experiment, the label for each gallery is denoted by a corresponding gallery ID for convenience as $\mathbf{L} = \{\#a, \#b, \cdots, \#un\}$, where #un is the label for an unregistered person.

Parameters: The standard deviation of the feature vector element was set at $\sigma = 394.5$, which is determined from the other

[*1] We calculated the foot's position on the ground plane from the bottom center coordinate of the bounding box.

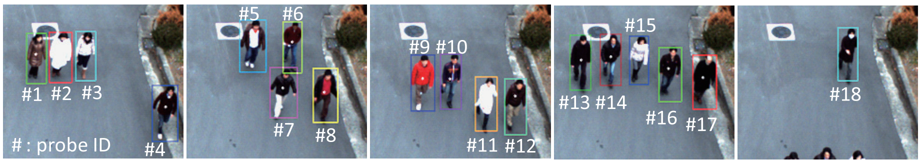


Fig. 8 Snapshots of the input sequence for preliminary experiment.

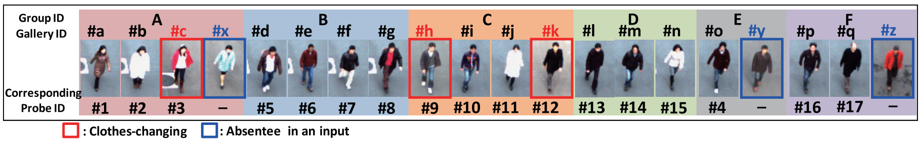
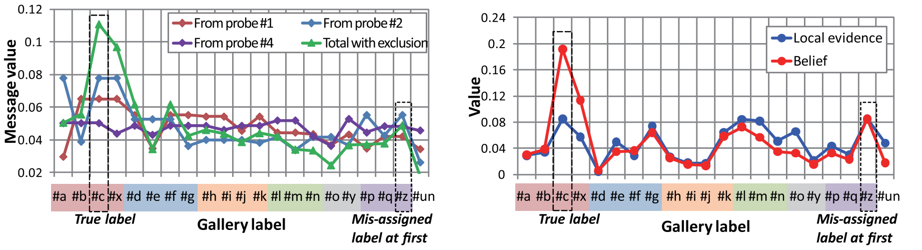


Fig. 9 Gallery set for the input shown in Fig. 8.

Table 1 Initial label correspondence for an input shown in Fig. 8. The numerical value in the table represents the belief.

Probe ID	Gallery ID																				
	#a	#b	#c	#x	#d	#e	#f	#g	#h	#i	#j	#k	#l	#m	#n	#o	#y	#p	#q	#z	#un
#1	0.177	0.023	0.016	0.011	0.002	0.120	0.025	0.031	0.027	0.057	0.115	0.077	0.054	0.018	0.058	0.051	0.023	0.044	0.012	0.011	0.048
#2	0.014	0.171	0.060	0.032	0.032	0.027	0.041	0.087	0.041	0.004	0.018	0.046	0.029	0.051	0.013	0.037	0.052	0.060	0.096	0.045	0.048
#3	0.029	0.034	0.085	0.058	0.005	0.050	0.029	0.074	0.027	0.017	0.017	0.065	0.084	0.082	0.051	0.066	0.021	0.043	0.030	0.085	0.048
#5	0.001	0.077	0.028	0.003	0.564	0.005	0.038	0.081	0.017	0.000	0.001	0.007	0.006	0.015	0.001	0.005	0.024	0.010	0.050	0.021	0.048
#6	0.099	0.001	0.007	0.057	0.000	0.065	0.003	0.009	0.004	0.124	0.040	0.024	0.052	0.017	0.354	0.075	0.003	0.011	0.001	0.009	0.048
#7	0.028	0.047	0.028	0.027	0.005	0.045	0.091	0.053	0.077	0.017	0.011	0.049	0.062	0.098	0.010	0.049	0.131	0.053	0.049	0.023	0.048
#8	0.008	0.089	0.055	0.032	0.028	0.020	0.044	0.136	0.029	0.003	0.005	0.045	0.038	0.079	0.008	0.029	0.079	0.031	0.126	0.066	0.048
#9	0.043	0.048	0.057	0.018	0.006	0.097	0.037	0.044	0.082	0.027	0.033	0.084	0.054	0.054	0.031	0.058	0.051	0.075	0.032	0.021	0.048
#10	0.095	0.005	0.007	0.023	0.000	0.117	0.016	0.009	0.038	0.230	0.049	0.068	0.040	0.026	0.071	0.079	0.022	0.048	0.005	0.004	0.048
#11	0.112	0.017	0.010	0.038	0.001	0.050	0.019	0.014	0.024	0.068	0.238	0.078	0.048	0.015	0.066	0.053	0.025	0.054	0.016	0.007	0.048
#12	0.048	0.015	0.036	0.110	0.001	0.053	0.024	0.045	0.021	0.043	0.018	0.067	0.116	0.079	0.079	0.076	0.028	0.028	0.019	0.045	0.048
#13	0.048	0.031	0.053	0.020	0.007	0.106	0.044	0.065	0.057	0.028	0.039	0.075	0.087	0.045	0.030	0.054	0.036	0.053	0.033	0.038	0.048
#14	0.018	0.046	0.117	0.035	0.007	0.041	0.026	0.054	0.067	0.013	0.012	0.042	0.082	0.139	0.024	0.046	0.044	0.061	0.034	0.045	0.048
#15	0.114	0.007	0.014	0.032	0.000	0.059	0.007	0.010	0.014	0.091	0.089	0.063	0.051	0.020	0.247	0.069	0.010	0.039	0.005	0.010	0.048
#4	0.045	0.012	0.023	0.099	0.000	0.065	0.018	0.018	0.027	0.063	0.042	0.039	0.054	0.046	0.136	0.157	0.017	0.058	0.008	0.025	0.048
#16	0.031	0.035	0.031	0.045	0.002	0.077	0.044	0.027	0.074	0.036	0.031	0.062	0.037	0.060	0.034	0.114	0.052	0.112	0.028	0.019	0.048
#17	0.014	0.092	0.033	0.051	0.006	0.028	0.048	0.050	0.032	0.006	0.017	0.063	0.038	0.057	0.010	0.039	0.081	0.056	0.183	0.049	0.048
#18	0.095	0.026	0.009	0.038	0.001	0.071	0.036	0.014	0.039	0.077	0.084	0.091	0.033	0.032	0.061	0.087	0.044	0.085	0.021	0.007	0.048

□ : True correspondence 🟨 : Assigned correspondence



(a) Received messages (b) Local evidence and belief

Fig. 11 Received messages and belief of probe #3 at the first message passing.

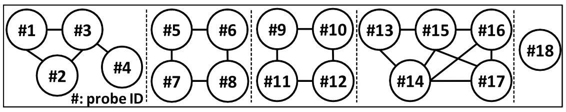


Fig. 10 Node connection between the probe subjects in Fig. 8.

Table 2 Compatibility for a pair of labels in the same group between probe subjects in Fig. 8.

Probe pair (#i, #j)	$\psi_{ij}(x_i = l_k, x_j = l_l)$ $l_k \neq l_l$ and $g(l_k) = g(l_l)$
(#1, #3)	0.42
(#2, #3)	0.73
(#4, #3)	0.16

preliminary experiment. Local evidence for the label of the unregistered person was set at $C_{un} = \frac{1}{N_l}$, where N_l is the number of gallery labels.

Results: **Table 1** shows the initial label correspondence via straightforward labeling. In this table, seven probe subjects (#3, #6, #7, #12, #13, #15, and #18) are initially mislabeled because of the within-class variation of the gait features caused by walking manner variations, clothes changes, and silhouette noise.

We illustrate the message effect on improving the belief from the initial state by taking probe #3 as an example. As shown in **Fig. 10**, probe #3 is connected to probes #1 and #2, which truly belong to group A (the same group as probe #3), and probe #4, which truly belongs to group E. Initially, probe #3 is mislabeled as #z and probes #1, #2, and #4 are correctly labeled, as shown in Table 1. The received message and the belief of probe #3 after the first message passing is then shown in **Fig. 11**. In this figure, we see that the messages from probe #1 and probe #2 contribute much to boost the belief for the label #c. This is because probes #1 and #2 have high initial beliefs (local evidence) for their true labels, and high compatibilities for a pair of labels which belong to the same group as probe #3, as shown in **Table 2**.

On another note, in the message shown in Fig. 11, the message about the label #x (absentee) is relatively high because #x is also a member of group A. The belief of probe #3 for the label #x, however, does not exceed that for the true label #c because local evidence for the true label #c is essentially higher than that for the label #x, even though the message magnitude for label #x is nearly equal to that for the label #c.

In this way, the initial mislabel assignments gradually improve with iteration of the message passing. Note that probe #18, which is an unregistered person and is initially mislabeled as #k, is not connected to any probe subject, as shown in Fig. 10, but is only connected to the factor node in this case. The assigned label to probe #18 is therefore changed only by exclusive force with an increase in the beliefs of the other labels.

The labeling accuracy of the proposed method under no seed node and of the straightforward method are shown in **Table 3** (in this experiment, the result of proposed method is unchanged with or without seed nodes). In this table, we can see that the proposed method significantly improves the labeling accuracy.

5.1.3 Evaluation for the Dataset from the Real Surveillance Camera

Gallery and probe data set: We arranged the real image sequences (320 pixels × 240 pixels/9 fps/jpeg format) which are obtained from the surveillance cameras installed in a Japanese elementary school. In this experiment, a scenario of person re-identification across two non-overlapping cameras is assumed and we collected gallery and probe subsequences from the two different cameras. The numbers of gallery and probe subjects are shown in **Table 4**, and the examples are shown in **Fig. 12**.

In this dataset, the observation angle of each subject is different to some extent between gallery and probe sequences and the trajectory and walking manner of each subject are more fluctuated than those in the dataset used in previous section.

Parameters: The standard deviation of the feature vector element was set at $\sigma = 1071.1$ and the seed decision threshold was set at $T_s = 0.7$. Both of these values were determined based on the training dataset composed of 40 subjects which are also extracted from the same cameras. Local evidence for the label of the unregistered person was set in the same way as the preliminary experiment.

Result: **Table 5** shows the labeling accuracy. In this table, we can see that the proposed method improves the labeling accuracy even for the real situation.

5.2 Experiment with Simulation Data

5.2.1 Settings

Observed space and trajectory: We assumed an input video sequence in which each walking person is captured by a surveillance camera in a virtually constructed space. We set the whole space to be 10 [m] × 2,000 [m] and the observed space to be 10 [m] × 20 [m] as shown in **Fig. 13**. In such a space, we arranged the initial position for each person, gave them velocities, and then moved them. For simplicity, we assumed that each person walked with constant velocity and that the walking direction was only the Y-direction, as shown in Fig. 13.

Gallery and probe data set: In all the simulation experiments, the number of gallery subjects (registered persons) is set to 1,000, and gait features for all of the gallery and probe subjects are randomly chosen from the gait database proposed in Ref. [26]. Note that the gait database [26] has expanded and includes 1,580 subjects at time of writing. We used two side-view sequences as the probe and gallery sequences.

We then considered the following three scenarios, and we defined the gallery and probe settings for each scenario as shown in **Table 6**.

Set *A*: *Person identification when going to elementary school in a group*: All of the gallery subjects are grouped. The registered

Table 3 Labeling accuracy for the input shown in Fig. 8.

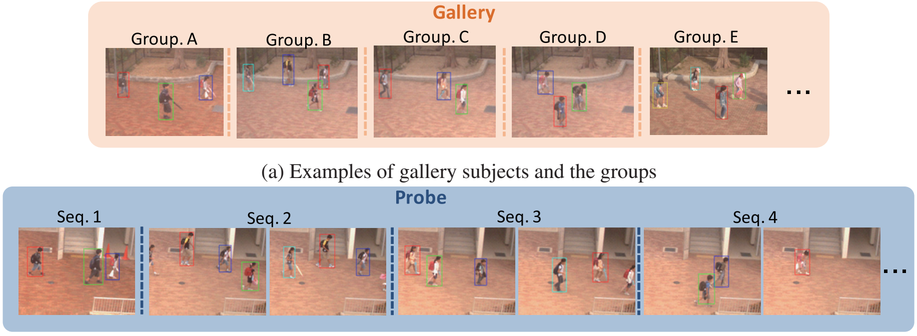
Method	Labeling accuracy
Straightforward	0.61
Proposed	1.0

Table 5 Labeling accuracy for the dataset from the real surveillance camera.

Method	Labeling accuracy
Straightforward	0.70
Proposed	0.87

Table 4 Gallery and probe settings in the dataset from the real surveillance camera.

Gallery setting				Probe setting						
Number of group	Number of member	Subject number		Absentee	Number of group	Subject number				
						Registered			Unregistered	Total
		Group belonging	Stand alone			Group belonging		Stand alone		
						In a group	In isolation			
14	2 to 5	39	1	0	16	37	2	1	7	47



(a) Examples of gallery subjects and the groups

(b) Examples of probe subjects

Fig. 12 Examples of gallery and probe subjects in the dataset from the real surveillance camera.

Table 6 Gallery and probe settings in simulation experiments.

Data set	Gallery setting					Probe setting						
	Number of group	Number of member	Subject number			Absentee	Number of group	Subject number				
								Registered			Unregistered	Total
								Group belonging		Stand alone		
			Group belonging	Stand alone	Total			In a group	In isolation			
I	125	6 to 10	1000	0	1000	0	125	1000	0	0	0	1000
A						10		980	10		10	
B	100	2 to 10	500	500		500	50	200	50	250		510
C											500	1000

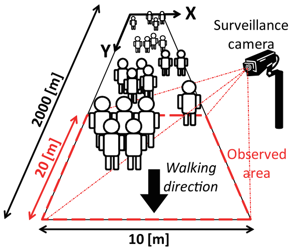


Fig. 13 Assumed environment in simulation experiments.

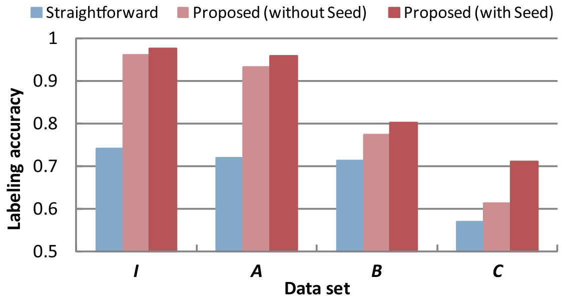


Fig. 14 Results for simulation data set.

person and the unregistered person correspond to a school student and an intruder, respectively. Absentees and isolated persons correspond to absent students and early or late arrival students. There can be a small number of unregistered persons, absentees, and isolated persons.

Set *B*: *Person identification in amusement theme parks*: Substantial numbers of the gallery subjects are assumed to be standalone (persons who belong to groups of only one member). The registered person and the unregistered person correspond to a fee-paying fair visitor and an unfair visitor who enters the park without the due entrance procedure. The absentee corresponds to a registered person who is in the park but is not captured by surveillance camera. The isolated person corresponds to a registered person who is lost or separated from their group with another objective. There can be a small number of unregistered persons and isolated persons in addition to some absentees.

Set *C*: *Person re-identification in network cameras*: We assume that there are two cameras which have different fields of view, and regard one side camera as the gallery-side camera and the other as the probe-side camera. Some of the gallery subjects are assumed to be standalone. A registered person corresponds to a person who is captured by the gallery-side camera, and an unregistered person corresponds to a person who is captured only by the probe-side camera. An absentee corresponds to a registered person who is not captured by the probe-side camera. An isolated person corresponds to a registered person who is separated from their group with another objective. There can be some unregistered persons and absentees, and a small number of isolated persons.

We also arranged the ideal scenario, where all gallery subjects are grouped and there are no absentees, isolated persons, or unregistered persons (denoted as set *I* in Table 6). We arranged 10 different sets randomly for each scenario. The performance for each data set is evaluated by averaging their results.

Parameters: The standard deviation of the feature vector element was set at σ = 366.2 and the seed decision threshold was set at T_s = 0.8. Both of these values were determined based on the gait database used. The local evidence for the label of an unregistered person C_{un} significantly influences the performance of the many-to-many labeling scheme in the presence of an unregistered person, particularly in the presence of a relatively large number of unregistered persons in an input sequence such as set *C*. Thus, we set the parameter at $C_{un} = 0$ for set *I*, $C_{un} = 0.002$ for sets *A* and *B*, and $C_{un} = 0.005$ for set *C*, so that the performance of the straightforward method for each data set becomes the best. Note that we also conducted the same experiments under no seed node ($T_s = 0.0$) to verify the effectiveness of seed node.

5.2.2 Results

Figure 14 shows the labeling accuracy. In this figure, we see that the proposed method discernibly improves the labeling accuracy for each data set and the introduction of seed node contributes the performance improvement. In particular, when the ratio of the number of persons in a group is high, the effectiveness of the proposed method is greatest, as shown in the results for sets *I* and *A*, while the performance improvements for sets *B* and *C* are relatively low.

Basically, the belief values for isolated persons, standalone persons, and unregistered persons for their own true labels are not expected to be directly boosted by the messages. Thus, in the case where such a person has the highest belief for a wrong label about another person at the first stage, it is difficult to recover the true label, except in the case where the wrong label is a label about a person in a group in an input sequence; that is, the exclusive force for the wrong label is expected (as the label change of probe #18 shows in the experiment in Section 5.1.2). This is one of the major reasons why the performances of the proposed method for sets *B* and *C* are lower than those of sets *I* and *A*.

6. Discussion

6.1 Limitation

While the proposed method significantly improves the labeling performance, there are still some subjects who are mislabeled, and subjects whose labels are negatively changed via message passing, even for the ideal set *I* in the simulation experiments. We list the typical cases of failure for the proposed method as follows.

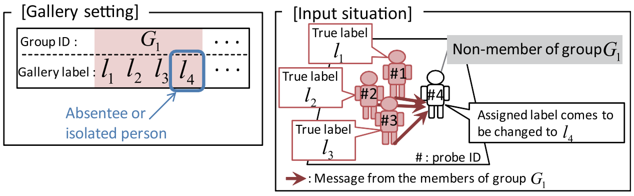


Fig. 15 Example situation of negative label change.

6.1.1 Mislabel within the Same Group Members

When a person in a group is mislabeled as another person in the same group at first, it is difficult to recover the true label because the belief for the true label and the wrongly assigned label are boosted to the same degree. Mislabeling within the same group members is, however, relatively rare compared with mislabeling between different groups. The rate of this kind of mislabel is relatively low.

6.1.2 Negative Label Change in the Presence of an Absentee or an Isolated Person in a Group

As shown in **Fig. 15**, when the following three incidents occur simultaneously, where i) an absentee or an isolated person exists in a group (the gallery subject with the label l_4 in group G_1), ii) *another person* [*2] (probe #4) comes close to the group members (probe #1, #2, and #3) with similar velocity vector in an input sequence, and iii) *another person* is not set as a seed. Then, *another person* may possibly be mislabeled as an absentee or an isolated person by messages from the group members. At the same time, if *another person* is mislabeled as an isolated person in such a case, the initial correct label assignment for the identical isolated person is excluded by *another person* and changed to the other incorrect label. Note that this often occurs in the presence of a number of standalone persons, unregistered persons, and isolated persons, such as our simulation sets *B* and *C*, because the event probability of the above incident increases.

Though the initial mislabel assignment and negative label changes as listed above possibly cause other negative label changes through the propagation of an undesirable message using the proposed method, the impact of such a negative effect is basically smaller than that of the positive effects in total, as shown in the results of the proposed method (Fig. 14).

6.2 Effect of the Seed Node on Performance

The contribution of seed node to the performance improvement of the proposed method is demonstrated in the simulation results (Fig. 14). The advantages of introducing seed node in graph are considered as followings.

- The avoidance of negative label change: As discussed in Section 6.1.2, the negative label change is not occurred if *another person* (which is described in Section 6.1.2) is set as a seed.
- The enhancement of message effect: According to the Eq. (18), a seed node can send more discriminative messages for the labels which belong to the same group of the assigned seed label as following example. We consider again the situation shown in Section 4.4 (Fig. 6 and Fig. 7 (b)), and let assume that the probe #2 is set as a seed with seed label l_2, that is, the local evidence for the label l_2 is set to 1 and that for each of all the other label is set to 0 in Fig. 7 (b). In this case, the messages from probe #2 to probe #1 for the labels l_1 and l_3 become as, $m_{21}(x_1 = l_1) = 1.0$ and $m_{21}(x_1 = l_3) = 0.0$, respectively [*3] as shown in **Fig. 16**. Therefore, the messages from a seed node promote the belief updates of its neighbor nodes, and positive label changes of them are also expected to be promoted as a result. Though, the negative label changes are possibly promoted, in particular, in the case that a seed node is assigned false label as seed label, such negative case is assumed to be occurred less often than positive case.

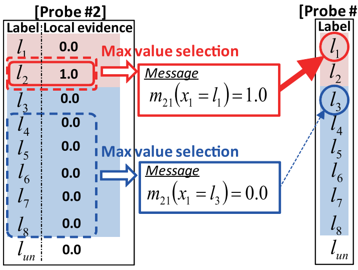


Fig. 16 Messages from a seed node (probe #2) to the other node (probe #1) under the setting shown in Fig. 6 (Section 4.4).

6.3 Effect of the Absence of Homography Calibration on Performance

We assume the homography calibration for the calculation of the position of each subject as described in Section 4.2. The cost of calibration is, however, expensive in some practical systems. One of the alternative ways is a direct use of the image pixel coordinate system instead of the world coordinate system to represent the trajectory of each individual. In many of practical surveillance systems, the camera captures the scene from near the top view or oblique view just like the scene used in our experiments. In such views, it is assumed that the direct use of image pixel coordinate does not have a serious impact on the performance of the proposed method.

To examine this, we conducted an additional experiment for the dataset used in Section 5.1.2 and we used the image pixel coordinate directly for the calculation of the positions of individuals. The parameters are set in pixel units, and we decided the parameters $d_{\max} = 160$ [pixel] and $d_{\min} = 30$ [pixel] based on the road width (approx. 320 pixels) and human width (approx. 30 pixels), and the $v_{\max} = 15$ [pixel/sec] and $v_{\min} = 0$ [pixel/sec] based on the average velocity 60 [pixel/sec] which roughly estimated from the dataset. As a result, we get the same result with that shown in Table 3, though the neighbor relationships among probe subjects are slightly changed.

6.4 Issues Toward the Practical System

The proposed method is based on some assumptions as described in Section 3. In terms of the total system (practical surveillance system), however, the following challenging issues

[*2] Not only a standalone person or an unregistered person, but also a person from another group.

[*3] This is an extreme case and the degree of magnitude relation between these messages are biased by constant value *C* in actual.

are required to be addressed in future work.

6.4.1 Obtaining the Group Affiliation

In practice, we need some kinds of registration procedures to associate the group affiliations with the individuals in advance. This is not such a serious problem in surveillance systems at factories and schools, where the potential observed persons are well-known in advance, i.e., the school children and the factory workers. Also, the registration can be achieved relatively easily with a system constructed at a place where the entrance and exit are controlled, i.e., where the group affiliation of each person can be easily checked and registered at the entrance gate, as in amusement or theme parks, stadiums, theaters, and airports. Alternatively, group affiliations can be derived by manual annotation (by user interaction) of the video sequence, and also inferred automatically by means of grouping techniques, such as data mining and clustering. In particular, social behavior-based group finding techniques have been developed in recent years [11], [31]. In these methods, the group is estimated based on trajectory, distance, and velocity of pedestrians. Thus, these methods bear affinity with the proposed method in terms of focusing such kinds of social behaviors, and the integration with these techniques is future work for the practical use of the proposed method.

6.4.2 Obtaining the Trajectory and Biometric Cue

Segmentation and tracking of each person are essential for the acquisitions of the trajectory and biometric cue, and these are not easy tasks when the scene is crowed, in particular, in the presence of occlusion among persons. To evaluate the proposed method for more practical scenes including such occlusion relationships, state-of-the-art techniques of segmentation and tracking, such as Refs. [3], [38], [44] are required to be applied for this problem. Moreover, cross-view matching of biometric cue is also essential and in the case of gait-based identification, the view transformation model [21] can be applied for this issue. The integration of these techniques with the proposed method also remains in future work.

7. Conclusions

In this paper, we proposed the behavior-based group context for person identification in video sequences and integrated it in the framework of CRF. In the proposed method, by means of message passing, the belief of individual identity is propagated to neighborhoods based on their group affiliation information and their behavioral differences, such as the spatial distance and the velocity vector difference in an input sequence, so that the same group members enhance one member's belief as those group members enhance each others' beliefs. In our experiments, we showed that the proposed method significantly improves the performance compared with the straightforward method based on biometric cues alone.

Our future work includes construction of the model for optimal selection of local evidence for the label of an unregistered person C_{un}. This is a rather general issue for many-to-many matching problems when considering an unregistered person.

Acknowledgments This work was partially supported by Grant-in-Aid for Scientific Research (S) 21220003 and "R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society," Strategic Funds for the Promotion of Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government.

References

[1] Oliva, A. and Torralba, A.: The role of context in object recognition, *Trends in Cognitive Sciences*, Vol.11, No.12, pp.520–527 (2007).

[2] Boykov, Y. and Jolly, M.: Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images, *Proc. Int. Conf. Computer Vision*, pp.105–112 (2001).

[3] Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E. and Gool, L.V.: Online multiperson tracking-by-detection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.33, No.9, pp.1820–1833 (2011).

[4] Cai, Y., Takala, V. and Pietikainen, M.: Matching Groups of People By Covariance Descriptor, *Proc. 20th International Conference on Pattern Recognition*, pp.2744–2747 (2010).

[5] Cho, T.S., Avidan, S. and Freeman, W.T.: The patch transform, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.32, No.8, pp.1489–1501 (2010).

[6] Cho, T.S., Avidan, S. and Freeman, W.T.: A probabilistic image jigsaw puzzle solver, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.183–190 (2010).

[7] Divvala, S.K., Hoiem, D., Hays, J.H., Efros, A.A. and Hebert, M.: An empirical study of context in object detection, *IEEE Conference on Computer Vision and Pattern Recognition* (2009).

[8] Gallagher, A.C. and Chen, T.: Using group prior to identify people in consumer images, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1–8 (2007).

[9] Galleguillos, C. and Belongie, S.: Context Based Object Categorization: A critical survey, *Computer Vision and Image Understanding*, Vol.114, pp.712–722 (2010).

[10] Galleguillos, C., Rabinovich, A. and Belongie, S.: Object categorization using co-occurrence, location and appearance, *IEEE Conference on Computer Vision and Pattern Recognition*, Vol.1, No.2, pp.1–8 (2008).

[11] Ge, W., Collins, R.T. and Ruback, R.B.: Automatically Detecting the Small Group Structure of a Crowd, *IEEE Workshop on Applications of Computer Vision*, pp.1–8 (2009).

[12] Gupta, A., Kembhavi, A. and Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.31, No.10, pp.1775–1789 (2009).

[13] Han, D., Bo, L. and Sminchisescu, C.: Selection and context for action recognition, *IEEE International Conference on Computer Vision*, pp.1933–1940 (2009).

[14] Han, J. and Bhanu, B.: Individual Recognition Using Gait Energy Image, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.28, No.2, pp.316–322 (2006).

[15] Heitz, G. and Koller, D.: Learning spatial context: Using stuff to find things, *European Conference on Computer Vision*, No.30–43 (2008).

[16] Hoiem, D., Efros, A.A. and Hebert, M.: Putting Objects in Perspective, *International Journal of Computer Vision*, No.1, pp.3–15 (2008).

[17] Hu, W., Tan, T., Wang, L. and Maybank, S.: A survey on visual surveillance of object motion and behaviors, *IEEE Trans. Systems, Man and Cybernetics*, Vol.34, pp.334–352 (2004).

[18] Ikizler-Cinbis, N. and Sclaroff, S.: Object, scene and actions: Combining multiple features for human action recognition, *European Conference on Computer Vision*, Vol.1, pp.494–507 (2010).

[19] Lan, T., Wang, Y., Mori, G. and Robinovitch, S.: Retrieving Actions in Group Contexts, *ECCV Workshop on Sign, Gesture and Activity* (2010).

[20] Lin, D., Kapoor, A., Hua, G. and Baker, S.: Joint people, event, and location recognition in personal photo collections using cross-domain context, *Proc. 11th European Conference on Computer Vision*, pp.243–256 (2010).

[21] Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T. and Yagi, Y.: Gait Recognition Using a view transformation model in the frequency domain, *European Conference on Computer Vision*, Vol.3, pp.151–163 (2006).

[22] Marques, O., Barenholtz, E. and Charvillat, V.: Context modeling in computer vision: Techniques, implications, and applications, *Multimedia Tools and Applications*, Vol.51, No.1, pp.303–339 (2011).

[23] Marszalek, M., Laptev, I. and Schmid, C.: Actions in context, *IEEE Conference on Computer Vision and Pattern Recognition* (2009).

[24] Naaman, M., Yeh, R.B., Garcia-Molina, H. and Paepcke, A.: Leveraging context to resolve identity in photo albums, *Proc. 5th ACM/IEEE-*

CS Joint Conference on Digital Libraries, pp.178–187 (2005).

[25] Nixon, M.S. and Carter, J.N.: Automatic recognition by gait, *Proc. IEEE*, Vol.94, No.11, pp.2013–2024 (2006).

[26] Okumura, M., Iwama, H., Makihara, Y. and Yagi, Y.: Performance evaluation of vision-based gait recognition using a very large-scale gait database, *IEEE 4th International Conference on Biometrics: Theory, Applications and Systems* (2010).

[27] Parikh, D., Zitnick, C.L. and Chen, T.: From appearance to context-based recognition: Dense labeling in small images, *IEEE Conference on Computer Vision and Pattern Recognition* (2008).

[28] Pellegrini, S., Ess, A., Schindler, K. and VanGool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking, *IEEE 12th International Conference on Computer Vision*, pp.261–268 (2009).

[29] Perko, R. and Leonardis, A.: A framework for visual-context-aware object detection in still images, *Computer Vision and Image Understanding*, Vol.114, No.6, pp.700–711 (2010).

[30] Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E. and Belongie, S.: Objects in context, *IEEE International Conference on Computer Vision*, pp.1–8 (2007).

[31] Sochman, J. and Hogg, D.C.: Who Knows Who - Inverting the Social Force Model for Finding Groups, *IEEE International Workshop on Socially Intelligent Surveillance and Monitoring*, pp.1–8 (2011).

[32] Song, Y. and Leung, T.: Context-aided human recognition- clustering, *European Conference on Computer Vision* (2006).

[33] Stone, Z., Zickler, T. and Darrell, T.: Autotagging Facebook: Social network context improves photo annotation, *Proc. CVPR Workshop on Internet Vision*, pp.1–8 (2008).

[34] Torralba, A., Murphy, K.P. and Freeman, W.T.: Using the forest to see the trees: Exploiting context for visual object detection and localization, *Comm. ACM*, Vol.53, No.3, pp.107–114 (2010).

[35] Choi, W., Shahid K. and Savarese, S.: Learning context for collective activity recognition, *IEEE Conference on Computer Vision and Pattern Recognition* (2011).

[36] Zheng, W.-S., Gong, S. and Xiang T.: Associating Groups of People, *British Machine Vision Conference*, Vol.5 (2009).

[37] Wu, J., Osuntogun, A., Choudhury, T., Philipose, M. and Rehg, J.M.: A scalable approach to activity recognition based on object use, *IEEE International Conference on Computer Vision*, pp.1–8 (2007).

[38] Wu, M., Peng, X., Zhang, Q. and Zhao, R.: Segmenting and tracking multiple objects under occlusion using multi-label graph cut, *Computers and Electrical Engineering*, Vol.36, No.5, pp.927–934 (2010).

[39] Yamaguchi, K., Berg, A.C., Ortiz, L.E. and Berg, T.L.: Who are you with and where are you going?, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1345–1352 (2011).

[40] Yao, B. and Fei-Fei, L.: Grouplet: A structured image representation for recognizing human and object interactions, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.9–16 (2010).

[41] Yao, B. and Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.17–24 (2010).

[42] Yedidia, J.S., Freeman, W.T. and Weiss, Y.: Understanding belief propagation and its generalizations, *Exploring artificial intelligence in the new millennium*, chapter8, pp.239–269, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA (2003).

[43] Zhang, L., Chen, L., Li, M. and Zhang, H.: Automated annotation of human faces in family albums, *Proc. 11th ACM International Conference on Multimedia*, pp.355–358 (2003).

[44] Zhao, T., Nevatia, R. and Wu, B.: Segmentation and tracking of multiple humans in crowded environments, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.30, pp.1198–1211 (2008).

[45] Zhao, W., Chellappa, R., Phillips, P.J. and Rosenfeld, A.: Face recognition: A literature survey, *ACM Computing Surveys*, Vol.35, No.4, pp.399–458 (2003).

[46] Zheng, W.-S., Gong, S. and Xiang, T.: Quantifying contextual information for object detection, *IEEE International Conference on Computer Vision*, pp.932–939 (2009).

Haruyuki Iwama received his B.S. and M.S. degrees in Engineering Science from Osaka University in 2003 and 2005. In 2005, he joined the Production Technology Development Group, Sharp Corporation, where he worked on vision-based inspection systems. He is currently a Ph.D. candidate and a Specially Appointed Researcher of the Institute of Scientific and Industrial Research, Osaka University. His research interests are segmentation, group context-based person identification, and gait recognition.

Yasushi Makihara received his B.S., M.S., and Ph.D. degrees in Engineering from Osaka University in 2001, 2002, and 2005, respectively. He is currently an Assistant Professor of the Institute of Scientific and Industrial Research, Osaka University. His research interests are gait recognition, morphing, and temporal super resolution. He is a member of IPSJ, RJS, and JSME.

Yasushi Yagi is the Director of the Institute of Scientific and Industrial Research, Osaka university, Ibaraki, Japan. He received his Ph.D. degrees from Osaka University in 1991. In 1985, he joined the Product Development Laboratory, Mitsubishi Electric Corporation, where he worked on robotics and inspections. He became a Research Associate in 1990, a Lecturer in 1993, an Associate Professor in 1996, and a Professor in 2003 at Osaka University. International conferences for which he has served as Chair include: FG1998 (Financial Chair), OMINVIS2003 (Organizing Chair), ROBIO2006 (Program co-chair), ACCV2007 (Program chair), PSVIT2009 (Financial Chair), ICRA2009 (Technical Visit Chair), ACCV2009 (General Chair), ACPR2011 (Program Co-chair) and ACPR2013 (General Chair). He has also served as the Editor of IEEE ICRA Conference Editorial Board (2007–2011). He is the Editorial member of IJCV and the Editor-in-Chief of IPSJ Transactions on Computer Vision & Applications. He was awarded ACM VRST2003 Honorable Mention Award, IEEE ROBIO2006 Finalist of T.J. Tan Best Paper in Robotics, IEEE ICRA2008 Finalist for Best Vision Paper, MIRU2008 Nagao Award, and PSIVT2010 Best Paper Award. His research interests are computer vision, medical engineering and robotics. He is a fellow of IPSJ and a member of IEICE, RSJ, and IEEE.

(Communicated by *Nassir Navab*)