Associating Groups of People

Wei-Shi Zheng
jason@dcs.qmul.ac.uk
Shaogang Gong
sgg@dcs.qmul.ac.uk
Tao Xiang
txiang@dcs.qmul.ac.uk

School of EECS,
Queen Mary University of London,
London E1 4NS, UK

Abstract

In a crowded public space, people often walk in groups, either with people they know or strangers. Associating a group of people over space and time can assist understanding individual's behaviours as it provides vital visual context for matching individuals within the group. Seemingly an 'easier' task compared with person matching given more and richer visual content, this problem is in fact very challenging because a group of people can be highly non-rigid with changing relative position of people within the group and severe self-occlusions. In this paper, for the first time, the problem of matching/associating groups of people over large space and time captured in multiple non-overlapping camera views is addressed. Specifically, a novel people group representation and a group matching algorithm are proposed. The former addresses changes in the relative positions of people in a group and the latter deals with variations in illumination and viewpoint across camera views. In addition, we demonstrate a notable enhancement on individual person matching by utilising the group description as visual context. Our methods are validated using the 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS) dataset on multiple camera views from a busy airport arrival hall.

1 Introduction

Object recognition has always been important for computer vision. In recent years, the focus of object recognition has shifted from recognising objects captured in isolation against clean background under well-controlled lighting conditions to a more challenging but also more useful problem of recognising objects subject to occlusion against cluttered background with drastic view angle and illumination changes. In particular, the problem of person re-identification or tracking (from disjoint views) has received increasing interest [6, 8, 9, 10, 13, 18], which aims to match a person observed in different non-overlapping locations over different camera views. In this paper, we consider a new problem, albeit closely related to the above, of associating groups of people over different camera views.

In a crowded public space, people often walk in groups, either with people they know or strangers. To be able to associate the same group of people over different camera views at different locations can bring about two significant benefits: (1) Matching a group of people over large space and time can be extremely useful in understanding and inferring longer-term association and more holistic behaviour of a group of people in public space. (2) It

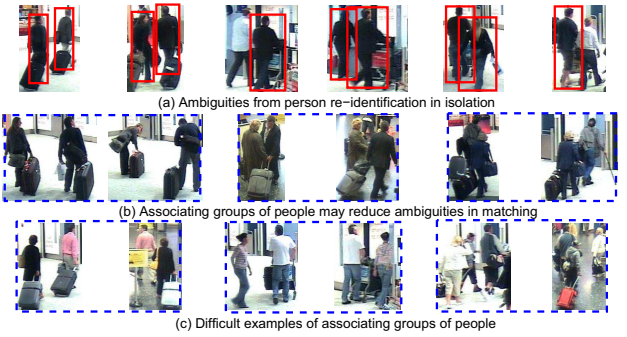

Figure 1: Advantages from and challenges in associating groups of people vs. person re-identification in isolation.

can provide vital visual context for assisting the match of individuals as the appearance of a person often undergoes drastic change across camera views caused by lighting and view angle variations. Most significantly, people appearing in public space are prone to occlusions by others near by. These viewing conditions make person re-identification an extremely hard problem. On the other hand, groups of people are less affected by occlusion which can provide a richer context and reduce ambiguity in discriminating an individual against others. This is illustrated by examples shown in Fig. 1 (a) where each of the six groups of people consists of one or two people in dark clothing. Based on appearance alone, it is difficult if not impossible to distinguish them in isolation. However, when they are considered in context by associating groups of people they appear together, it becomes much clearer that all candidates highlighted by red boxes are different people. Fig. 1 (b) shows examples of cases where matching groups of people together seems to be easier than matching individuals in isolation due to the changes in the appearance of people in different views caused by occlusion or change of body posture. We consider that the group context is more robust against these changes and more consistent over different views.

However, associating groups of people introduces new challenges: (1) Compared to an individual, the appearance of a group of people is highly non-rigid and the relative positions of the members can change significantly and often. (2) Although occlusions by other objects is less an issue, self-occlusion caused by people within the group remains a problem which can cause changes in group appearance. (3) Different from a relatively stable shape of every upright person which has similar aspect ratio, the aspect ratio of the shapes of different groups of people can be very different. Some difficult examples are shown in Fig. 1 (c).

Due to these challenges, existing representation descriptors and matching methods for person re-identification are not suitable for solving the group association problem. In this paper, a novel people group representation is proposed based on two new ratio-occurrence descriptors introduced here. Given this group representation, a group matching algorithm is formulated to achieve robustness against both changes in relative positions of people within a group and variations in illumination and view angle across different cameras. In addition, a new person re-identification method is introduced by utilising associated group of people as visual context to improve the matching of individuals across camera views.

To the best of our knowledge, there has been no previous attempt at addressing the problem of matching/associating groups of people over multiple camera views. There are related work reported in the literature on crowd detection and analysis [1, 2, 11, 15] and group activity recognition [7, 17]. However, these are not concerned with group association over space and time either within the same camera views or across different views. The proposed

model is validated using 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS) dataset captured by multiple camera views from a busy airport arrival hall [14].

2 Modelling Group Association

2.1 Group Representation

Given a gallery set and a probe set of images of different groups of people, we aim to find the best matched group template registered in the gallery for any probe group image.

Similar to [16, 18], we first assign a label to each pixel of a given group image $\mathbf{I}$. The label can be a simple colour or a visual word index of colour and gradient information together. Due to the change of camera view and varying positions and motions of a group of people, we consider that integration of local rotational invariant features and color density information is better for constructing visual words for indexing. In particular, we extract SIFT features [12] (a 128-dimensional vector) for each RGB channel at each pixel with a surrounding support region (12×12 in our experiment), and obtain an average RGB colour vector of that pixel over a support region (3×3 in our experiment), where the colour vector is normalized to $[0,1]^3$. The SIFT vector and colour vector are then concatenated for each pixel for representation, which we call the SIFT+RGB feature. The SIFT+RGB features are quantized into n clusters by K-means and a code book $\mathscr{A}$ of n visual words $\mathbf{w}_1$,...,$\mathbf{w}_n$ is built. Finally, an appearance label image is built by assigning a visual word index to the corresponding SIFT+RGB feature at each pixel of the group image. In order to remove background information, background subtraction is first performed. Then, only features extracted for foreground pixels are used to construct visual words for group image representation.

To represent the distribution of visual words of any image, a single histogram of visual words, which we call the holistic histogram, can be used [3]. However, this representation loses all spatial distribution information of the visual words. One way to alleviate this problem is to divide the image into grid blocks and concatenate the histograms of blocks one by one, for instance similar to [4]. However, this still cannot cope with a common case (examples in Fig. 1 (c)) in group images when people swap their positions. Moreover, corresponding image grid positions between two group images are not always guaranteed to represent foreground regions, therefore such a hard-wired grid block based representation is not always valid. In addition, it is noted that whilst global spatial relationships between people within a group can be highly unstable, local spatial relationships between small patches within a local region may be stable, e.g. within the bounding box of a person. In view of these characteristics of group images, we propose to represent a group using two descriptors: a *center rectangular ring ratio-occurrence descriptor* which aims to describe the ratio information of visual words within and between different rectangular ring regions, and a *block based ratio-occurrence descriptor* for exploring more specific local spatial information between visual words that could be stable. These two descriptors are finally combined for group representation.

Center Rectangular Ring Ratio-Occurrence Descriptor (CRRRO): Rectangular ring regions are considered to be approximately rotational invariant and efficient integral computation of visual words histogram is also available [16]. To that end, we define a holistic rectangular ring structure expanding from the center of a group image. The ℓ rectangular rings divide a group image into ℓ non-overlapped regions $P_1, \cdots, P_\ell$ from inside to outside. Every rectangular ring is $0.5 \cdot N/\ell$ and $0.5 \cdot M/\ell$ thick along the vertical and horizontal directions respectively (see Fig. 2 (a) with $\ell = 3$), where the group image is of size $M \times N$. Such

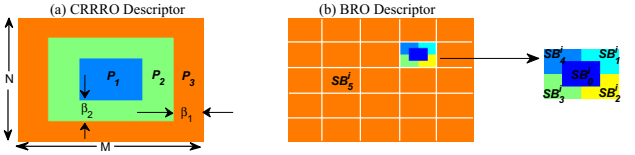

Figure 2: Partition of a group image by two descriptors. Left: the Center Rectangular Ring Ratio-Occurrence Descriptor ($\beta_1 = M/2\ell, \beta_2 = N/2\ell, \ell = 3$); Right: the Block based Ratio-Occurrence Descriptor ($\gamma = 1$), where white lines are to show the grids of the image.

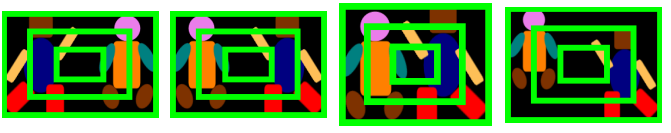

Figure 3: An illustration of a group of people against dark background.

a partitioning of a group image is especially useful for describing a pair of people because the distribution of constituent patches of each person in each ring is likely to be more stable against changes in relative positions between the two people over different viewpoints or scaling (see illustrations in Fig. 3).

After a partition of any image for representation, a common way for constructing a codebook is to concatenate the histogram of visual words from each ring. However, this ignores any spatial relationships existed between visual words from different ring-zones of a partition. We consider retaining such spatial relationships to be important and introduce a notion of *intra-* and *inter- ratio-occurrence maps* as follows. For each ring-region P_i, a histogram $\mathbf{h}_i$ is built, where $\mathbf{h}_i(a)$ indicates the frequency (occurrence) of visual word $\mathbf{w}_a$. Then for P_i, an *intra ratio-occurrence map* $\mathbf{H}_i$ is defined as

$$\mathbf{H}_i(a,b) = \frac{\mathbf{h}_i(a)}{\mathbf{h}_i(a) + \mathbf{h}_i(b) + \varepsilon}, \tag{1}$$

where ε is a very small positive value in order to avoid $0/0$. $\mathbf{H}_i(a,b)$ then represents the ratio-occurrence between words $\mathbf{w}_a$ and $\mathbf{w}_b$ within the region.

In order to capture any spatial relationship between visual words within and outside region P_i, we further define another two ratio occurrence maps for ring-region P_i. Define:

$$\mathbf{g}_i = \sum_{j=1}^{i-1} \mathbf{h}_j, \quad \mathbf{s}_i = \sum_{j=i+1}^{\ell} \mathbf{h}_j,$$

where $\mathbf{g}_i$ represents the distribution of visual words enclosed by the rectangular ring P_i and $\mathbf{s}_i$ represents the distribution of visual words outside P_i, where we define $\mathbf{g}_1 = \mathbf{0}$ and $\mathbf{s}_\ell = \mathbf{0}$. Then two *inter ratio-occurrence maps* $\mathbf{S}_i$ and $\mathbf{G}_i$ are formulated as follows:

$$\mathbf{G}_i(a,b) = \frac{\mathbf{g}_i(a)}{\mathbf{g}_i(a) + \mathbf{h}_i(b) + \varepsilon}, \quad \mathbf{S}_i(a,b) = \frac{\mathbf{s}_i(a)}{\mathbf{s}_i(a) + \mathbf{h}_i(b) + \varepsilon}. \tag{2}$$

Therefore, for each ring-region P_i, we construct a triplet representation $\mathbf{T}_r^i = \{\mathbf{H}_i, \mathbf{S}_i, \mathbf{G}_i\}$, and a group image is represented by a set $\{\mathbf{T}_r^i\}_{i=1}^{\ell}$. We shall demonstrate in our experiment that this group image representation using a set of triplet intra- and inter-ratio occurrence maps gives better performance for associating groups of people than that of using a conventional concatenation based representation.

Block based Ratio-Occurrence Descriptor (BRO): The descriptor designed above still cannot cope well with large non-center-rotational changes in people's positions within a group. It also does not utilise any local structure information that may be more stable or

consistent across different views of the same group, e.g. certain parts of a person can be visually more consistent than others. As we do not make any assumptions on people in a group being well segmented due to self-occlusion, we revisit a group image to explore patch (partial) information approximately by dividing it into $\omega_1 \times \omega_2$ grid blocks $B_1, B_2, \cdots, B_{\omega_1 \times \omega_2}$, and only the foreground blocks (defined as the block with more than 70 percent pixels are foreground) are considered. Due to the approximate partition of a group image and the low resolution of each patch or potential illumination change and occlusion, we extract rather simple (therefore probably more robust) spatial relationships between visual words in each foreground block by further dividing the block into small block regions using L-shaped partition [18] with a modification that the most inner four block regions are merged (see Fig. 2 (b)). This is because those block regions are always small and may not contain sufficient information. As a result, we obtain $4\gamma+1$ block regions within each block B_i denoted by $SB^i_0, \cdots, SB^i_{4\gamma}$ for some positive integer γ.

For associating groups of people over different views, we first note that not all blocks B_i appear at the same positions in the group images. For example, a pair of people may swap their positions resulting in the blocks corresponding to those foreground pixels change their positions in different images. Also, there may be other visually similar blocks in the same group image. Hence, describing local matches only based on features within block B_i could not be distinct enough. To reduce this ambiguity, for representing each block B_i, we further include a complementary image region $SB^i_{4\gamma+1}$, which is the image portion outside block B_i (see Fig. 2 (b) with $\gamma = 1$). Therefore, for each block B_i, we partition the group image into $SB^i_0, SB^i_1, \cdots, SB^i_{4\gamma}$ and $SB^i_{4\gamma+1}$. We demonstrate in our experiment that including such complementary region $SB^i_{4\gamma+1}$ would significantly enhance matching performance.

Like the Center Rectangular Ring Ratio-Occurrence Descriptor, for each block B_i, we learn an intra ratio-occurrence map $\mathbf{H}^i_j$ between visual words in each block region SB^i_j. Similarly, we explore an inter ratio-occurrence map $\mathbf{O}^i_j$ between different block regions SB^i_j. Since the size of each block region in block B_i would always be relatively much smaller than the complementary region $SB^i_{4\gamma+1}$, the ratio information between them will be sensitive to noise. Consequently we consider two simplified inter ratio-occurrence maps $\mathbf{O}^i_j$ between block B_i and its complementary region $SB^i_{4\gamma+1}$ formulated as follows:

$$\mathbf{O}^i_1(a,b) = \frac{\mathbf{t}_i(a)}{\mathbf{t}_i(a)+\mathbf{z}_i(b)+\varepsilon}, \quad \mathbf{O}^i_2(a,b) = \frac{\mathbf{z}_i(a)}{\mathbf{z}_i(a)+\mathbf{t}_i(b)+\varepsilon}, \tag{3}$$

where $\mathbf{z}_i$ and $\mathbf{t}_i$ are the histograms of visual words of block B_i and image region $SB^i_{4\gamma+1}$, respectively. Then, each block B_i is represented by $\mathbf{T}^i_b = \{\mathbf{H}^i_j\}_{j=0}^{4\gamma+1} \bigcup \{\mathbf{O}^i_j\}_{j=1}^{2}$, and a group image is represented by a set $\{\mathbf{T}^i_b\}_{i=1}^{m}$ where m is the amount of foreground blocks B_i.

These two proposed descriptors, CRRRO and BRO are specially designed for associating images of groups of people. Due to highly unstable positions of people within a group and likely partial occlusions among them, they explore the inter-person spatial relational information in a group and the likely local patch (partial) information for each person respectively.

2.2 Group Image Matching

We match two group images $\mathbf{I}_1$ and $\mathbf{I}_2$ by combining the distance metrics of the two proposed descriptors as follows:

$$d(\mathbf{I}_1,\mathbf{I}_2) = d_r\left(\{\mathbf{T}^i_r(\mathbf{I}_1)\}_{i=1}^{\ell}, \{\mathbf{T}^{i'}_r(\mathbf{I}_2)\}_{i'=1}^{\ell}\right) + \alpha \cdot d_b\left(\{\mathbf{T}^i_b(\mathbf{I}_1)\}_{i=1}^{m_1}, \{\mathbf{T}^{i'}_b(\mathbf{I}_2)\}_{i'=1}^{m_2}\right), \quad \alpha \geq 0, \tag{4}$$

where $\{\mathbf{T}_r^i(\mathbf{I}_1)\}_{i=1}^{\ell}$ indicates the center rectangular ring ratio-occurrence descriptor for group image $\mathbf{I}_1$ whilst $\{\mathbf{T}_b^i(\mathbf{I}_1)\}_{i=1}^{m_1}$ is for the block based descriptor.

For d_r, the L_1 norm metric is used to measure the distance between each corresponding ratio-occurrence map and d_r is obtained by averaging these distances. For d_b, since the spatial relationship between patches is not stable in different images of the same group and also not all the patches in one group image can be matched with those in another, it is inappropriate to directly measure the distance between the corresponding patches (blocks) of two group images. To address this problem, we assume that for each pair of group images, there exists at most k pairs of matched local patches between two images. We then define d_b as a *top k-match metric* where k is a positive integer as follows:

$$d_b\left(\{\mathbf{T}_b^i(\mathbf{I}_1)\}_{i=1}^{m_1},\{\mathbf{T}_b^{i'}(\mathbf{I}_2)\}_{i'=1}^{m_2}\right)=\min_{\mathbf{C},\mathbf{D}}\left\{k^{-1}\cdot||\mathbf{AC}-\mathbf{BD}||_1\right\},$$
$$\mathbf{A}\in\mathbb{R}^{q\times m_1},\mathbf{B}\in\mathbb{R}^{q\times m_2},\mathbf{C}\in\mathbb{R}^{m_1\times k},\mathbf{D}\in\mathbb{R}^{m_2\times k}, \quad (5)$$

where the i^{th} (i'^{th}) column of matrix $\mathbf{A}$ ($\mathbf{B}$) is the vector representation of $\mathbf{T}_b^i(\mathbf{I}_1)$ ($\mathbf{T}_b^{i'}(\mathbf{I}_2)$), each column $\mathbf{c}_j$ ($\mathbf{d}_j$) of $\mathbf{C}$ ($\mathbf{D}$) is an indicator vector in which only one entry is 1 and the others are zeros, and the columns of $\mathbf{C}$ ($\mathbf{D}$) are orthogonal. Note that m_1 and m_2, the amount of foreground blocks in two group images, may be unequal. Generally, directly solving Eq. (5) is hard. Noting that $\min_{\mathbf{C},\mathbf{D}}\left\{||\mathbf{AC}-\mathbf{BD}||_1\right\}\leq\sum_{j=1}^{k}\min_{\mathbf{c}_j,\mathbf{d}_j}\left\{||\mathbf{Ac}_j-\mathbf{Bd}_j||_1\right\}$ where $\{\mathbf{c}_j\}$ and $\{\mathbf{d}_j\}$ are sets of orthogonal indicator vectors, we therefore approximate the k-match metric value as follows: the most matched patches $\mathbf{a}_{i_1}$ and $\mathbf{b}_{i'_1}$ are first found by finding the smallest L_1 distance between columns of $\mathbf{A}$ and $\mathbf{B}$. We then remove $\mathbf{a}_{i_1}$ and $\mathbf{b}_{i'_1}$ from $\mathbf{A}$ and $\mathbf{B}$ respectively and find the next most matched pair. This procedure repeats until the top k matched patches are found.

3 Group as Contextual Cue for Person Re-identification

We wish to explore group information for reducing the ambiguity in person re-identification if a person would appear in the same group. Suppose a set of L paired samples $\{(\mathbf{I}_p^i,\mathbf{I}_g^i)\}_{i=1}^{L}$ are given, where $\mathbf{I}_g^i$ is the corresponding group image of the i^{th} person image $\mathbf{I}_p^i$. We introduce a group-contextual-descriptor similar in spirit to the center rectangular ring descriptor introduced above, with a minor modification that we expand the rectangular ring structure surrounding each person. This makes the group context person specific, i.e. two people in the same group would have different context. Note that, only context features at foreground pixels are extracted. As a result, the most inner rectangular region P_1 is the bounding box of a person, and for other outer rings, they are $\max\{M-x_1-0.5\cdot M_1,x_1-0.5\cdot M_1\}/(\ell-1)$ and $\max\{N-y_1-0.5\cdot N_1,y_1-0.5\cdot N_1\}/(\ell-1)$ thick along the horizontal and vertical directions, where (x_1,y_1) is the center of region P_1, M and N are width and height of the group image, and M_1 and N_1 are width and height of P_1. In particular, when $\ell=2$, the rectangular ring structure would divide a group image into two parts: a person-centred bounding box and a surrounding complementary image region.

There can be many ways to integrate group information for person re-identification. In this paper, we simply combine the distance metric d_p of some person descriptor such as the colour histogram and the distance metric d_r of the corresponding group context descriptor between two people. More specifically, denote the person descriptors of person image $\mathbf{I}_p^1$ and $\mathbf{I}_p^2$ as $\mathbf{P}_1$ and $\mathbf{P}_2$ respectively and denote their corresponding group context descriptors as $\mathbf{T}_1$ and $\mathbf{T}_2$ respectively. Then the distance between two people is computed as:

$$d(\mathbf{I}_p^1,\mathbf{I}_p^2)=d_p(\mathbf{P}_1,\mathbf{P}_2)+\beta\cdot d_r(\mathbf{T}_1,\mathbf{T}_2),\ \ \beta\geq 0. \quad (6)$$

4 Experiments

We conducted extensive experiments using the 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS) dataset to evaluate the feasibility and performance of the proposed methods for associating groups of people in a crowded public space.

Dataset & Parameter Settings: The i-LIDS MCTS dataset was captured at an airport arrival hall in the busy times under a multi-camera CCTV network. We extracted image frames captured from two non-overlapping camera views. In total, 64 groups were extracted and 274 group images were cropped. Most of the groups have 4 images, either from different camera views or from the same camera but captured at different locations at different times. These group images are of different sizes. Automatic detection of group image should be required in practice. In this paper, we take the first step on association of groups of people and focus on the evaluation of group descriptors. Sliding window based technique could be used in practice for group image detection based on the proposed group descriptors. From the group images, we extracted 476 person images for 119 pedestrians, most of which are with 4 images. All person images were normalized to 64×128 pixels. Different from other person datasets [6, 8, 18], these person images were captured by non-overlapping cameras, and many of them underwent large illumination change and were subject to occlusion.

For code book learning, additional 80 images (of size 640×480) were randomly selected with no overlap with the dataset described above. As described in Section 2, the SIFT+RGB features were extracted at each pixel of an image. In our experiments, a code book with 60 visual words (clusters) was built using K-means.

Unless otherwise stated, our descriptors are set as follows. For the center rectangular ring ratio-occurrence (CRRRO) descriptor, we set $\ell = 3$. For the block based ratio-occurrence (BRO) descriptor, each image was divided into 5×5 blocks, γ was set to 1, and the top 10-match score was computed. The default combination weight α in Eq. (4) was set to 0.8.

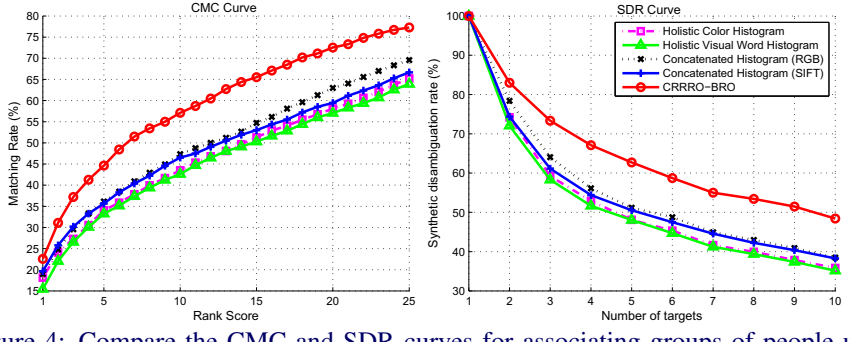

Figure 4: Compare the CMC and SDR curves for associating groups of people using the proposed CRRRO-BRO descriptor with those from other commonly used descriptors.

Evaluation of Group Association: We randomly selected one image from each group to build the gallery set and the other group images formed the probe set. For each group image in the probe set, we measured its similarity with each template image in the gallery. The s-nearest correct match for each group image was obtained. This procedure was repeated 10 times and the average cumulative match characteristic (CMC) curve [18] and the synthetic disambiguation rate (SDR) curve [8] were used to measure the performance, where the top 25 matching rates are shown for CMC curve and the SDR curve is able to give an overview of the whole CMC curve from the reacquisition point of view [8].

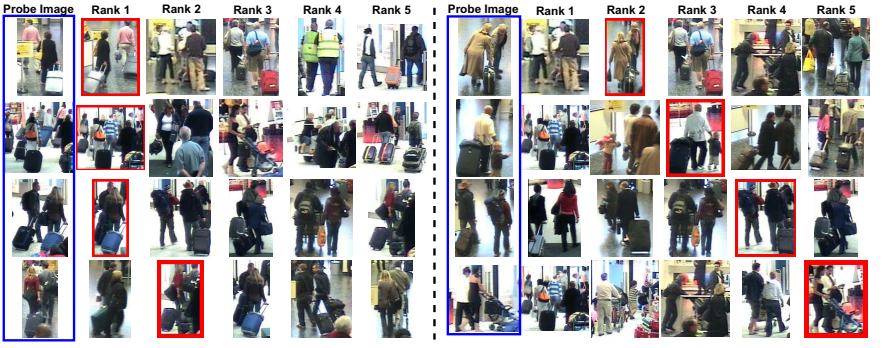

Figure 5: Examples of associating groups of people using our model. Correct matches are highlighted by red boxes.

The performance of the combined Center Rectangular Ring Ratio-Occurrence and Block based Ratio-Occurrence (CRRRO-BRO) descriptor approach (Eq. (4)) is shown in Fig. 4. We compare our model with two commonly used descriptors, colour histogram and visual word histogram of SIFT features (extracted at each colour channel) [3], which represent the distributions of colour or visual words of each group image holistically. We also apply these two descriptors to the designed center rectangular ring structure by concatenating the colour or visual word histogram of each rectangular ring. For the colour histogram, we selected the number of colour bins from $\{8, 16, 32, 64, 128\}$ and found 16 was the best one. In order to make the compared descriptors scale invariant, the histograms used in the compared methods were normalized [5]. For measurement, the *Chi-square* distance χ^2 [5] was used.

Results in Fig. 4 show the proposed CRRRO-BRO descriptor gives the best performance. It always keeps a notable margin to the CMC curve of the second best method, with 44.62% against 36.14% and 77.29% against 69.57% for rank 5 and 25 matching respectively. Compared to the existing holistic representations and the concatenation of local histograms representations, the proposed descriptor benefits from exploring the ratio information between visual words within and outside each local region. Moreover, Fig. 6 (b) shows that either using the proposed center based or block based descriptor can still achieve an overall improvement as compared to the concatenated histogram of visual words using SIFT+RGB features (described in Section 2) denoted by "Concatenated Histogram (Center, SIFT+RGB)" and "Concatenated Histogram (Block, SIFT+RGB, $k = 10$)" in the figure, respectively. This suggests the ratio maps can provide more information for matching. Finally, Fig. 5 shows some examples of associating groups of people using the proposed model (Eq. (4) with $\alpha = 0.8$). It demonstrates that our model is capable of establishing correct matching when there are large variations in people's appearances and their relative positions in a group caused by some very challenging viewing conditions including significantly different view angles and severe occlusions.

Evaluation of the Proposed Descriptors: To give more insight on how the proposed descriptors perform in different aspects, we show in Fig. 6 (a) comparative results between the combination CRRRO-BRO (Eq. (4)) and the individual CRRRO and BRO descriptors using the metrics d_r and d_b as described in Section 2.2. It shows that the combination of the center ring based and local block based descriptors utilises complementary information and improves the performance of each individual descriptor. Fig. 6 (b) evaluates the effects of using ratio map information as discussed in the last paragraph. Fig. 6 (c) shows that by exploring the inter ratio-occurrence between regions on the top of the intra one, an overall

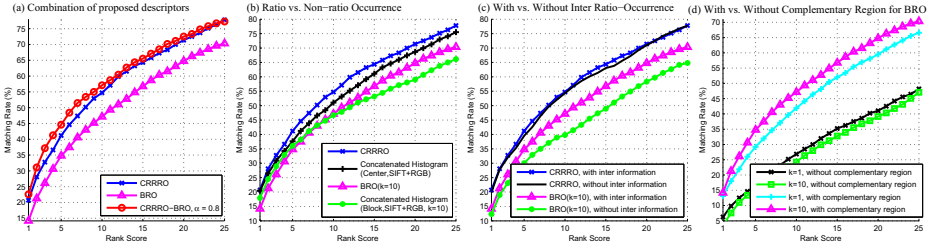

Figure 6: Evaluation of the proposed descriptors.

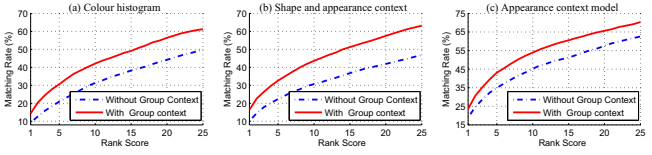

Figure 7: Improving person re-identification using group context.

better performance is obtained as compared with a model without utilising such information. For the block based ratio-occurrence descriptor, Fig. 6 (d) indicates that including the complementary region with respect to each block B_i can reduce the ambiguity during matching.

Person Re-Identification without/with Group Context: To show the effect of group context for improving person re-identification, we implemented three methods for person re-identification. One is to use colour histogram as representation of a person image and the other two are the appearance context model [18] and the shape and appearance context model [18]. For colour histogram, the number of colour bins was 16. For the appearance context model and the shape and appearance context model, we learned the corresponding code books from additional 40 person images extracted from the images used for code book learning for the proposed group descriptors. We employed the same code book sizes as suggested by [18], used 8 quantizing orientations for the HOG in [18] and 20 L-shaped regions for the plane partition in [18]. For group context, as described in Section 3, a two-rectangular-ring structure is expanded from the center of the bounding box of each person. For evaluation, one image for each person was randomly selected as the gallery template and the others were as probe images. This procedure was repeated 10 times and the average performances of these techniques without and with group context are shown in Fig. 7, where χ^2 distance is used for the colour histogram model and L_1 norm distance is for the other two person descriptors. It is evident that including group context notably improves the matching rate regardless of the choice of different person re-identification techniques. Over 10% improvement was always achieved for the colour histogram model and the shape and appearance context model, whilst about 8-9% improvement was obtained over the appearance context model. Note that the performance we obtained for the shape and appearance context model are not as high as that reported in [18]. This is because the person images from the i-LIDS MCTS dataset are much more challenging since they were captured from non-overlapping multiple camera views subject to significant occlusion, large variations in both view angle and illumination.

5 Conclusion

In this paper, for the first time, we have formulated the problem of associating groups of people over multiple non-overlapping camera views and proposed a center rectangular ring and

block based ratio-occurrence descriptors for effective representation of images of groups of people in crowded public spaces, and a top k-match model for matching possible local patches of two group images. We further demonstrated the advantages gained from utilising group context information in improving person re-identification under very challenging viewing conditions using the 2008 i-LIDS Multiple Camera Tracking Scenario dataset. Our ongoing work is to improve the descriptors and matching algorithms in order to cope with severe variations in the relative positions of people in groups.

Acknowledgement

This research was partially funded by the EU FP7 project SAMURAI, grant no. 217899.

References

[1] O. Arandjelović. Crowd detection from still images. In *BMVC*, 2008.

[2] G. J . Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *CVPR*, 2006.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nystrom method. *PAMI*, 26(2):214–225, 2004.

[6] N. Gheissari, T. B. Sebastian, P. H. Tu, J. Rittscher, and R. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR*, 2006.

[7] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In *ICCV*, 2003.

[8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[9] W. Hu, M. Hu, X. Zhou, J. Lou, T. Tan, and S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *PAMI*, 28(4):663–671, 2006.

[10] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. Tracking across multiple cameras with disjoint views. In *ICCV*, 2003.

[11] D. Kong, D. Gray, and H. Tao. Counting pedestrians in crowds using viewpoint invariant training. In *BMVC*, 2005.

[12] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2(60):91–110, 2004.

[13] C. Madden, E. Cheng, and M. Piccardi. Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Mach. Vision Appl.*, 18(3):233–247, 2007.

[14] UK Home Office. i-LIDS Multiple Camera Tracking Scenario Definition. 2008.

[15] V. Rabaud and S. Belongie. Counting crowded moving objects. In *CVPR*, 2006.

[16] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *CVPR*, 2006.

[17] S. Saxena, F. Brémond, M. Thonnat, and R. Ma. Crowd behavior recognition for video surveillance. In *10th International Conference on Advanced Concepts for Intelligent Vision Systems*, 2008.

[18] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007.