Chapter 9
Group Association: Assisting Re-identification by Visual Context

Wei-Shi Zheng, Shaogang Gong and Tao Xiang

Abstract In a crowded public space, people often walk in groups, either with people they know or with strangers. Associating a group of people over space and time can assist understanding an individual's behaviours as it provides vital visual context for matching individuals within the group. This seems to be an 'easier' task compared with person re-identification due to the availability of more and richer visual content in associating a group; however, solving this problem turns out to be rather 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 work, the problem of matching/associating groups of people over large space and time gaps 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 uses the proposed group descriptors for measuring the similarity between two candidate images. Based on group matching, we further formulate a method for matching individual person using the group description as visual context. These methods are validated using the 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS) dataset on multiple camera views from a busy airport arrival hall.

W.-S. Zheng (✉)
Sun Yat-sen University, Guangzhou, China
e-mail: wszheng@ieee.org

S. Gong
Queen Mary University of London, London, UK
e-mail: sgg@eecs.qmul.ac.uk

T. Xiang
Queen Mary University of London, London, UK
e-mail: txiang@eecs.qmul.ac.uk

S. Gong et al. (eds.), *Person Re-Identification*,
Advances in Computer Vision and Pattern Recognition,
DOI: 10.1007/978-1-4471-6296-4_9, © Springer-Verlag London 2014

9.1 Introduction

Object recognition has been a focus of computer vision research for the past five decades. 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 potentially more useful problem of recognising objects under occlusion against cluttered background with drastic view angle and illumination changes, known as 'recognition in the wild'. In particular, the problem of person re-identification or tracking between disjoint views has received increasing interest [1–6], which aims to match a person observed at different non-overlapping locations observed in different camera views. Typically, person re-identification is addressed by detecting and matching the visual appearance of isolated (segmented) individuals. In this work, we go beyond the conventional individual person re-identification by framing the re-identification problem in the context of associating groups of people in proximity across different camera views. We call this the *group association* problem. Moreover, we also consider how to explore a group of people as non-stationary visual context for assisting individual centred person re-identification within a group. This is often the condition under which re-identification needs be performed in a public space such as transport hubs.

In a crowded public space, people often walk in groups, either with people they know or with strangers. To be able to associate the same group of people over different camera views at different locations can bring about two 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 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 nearby. These viewing conditions make person re-identification in crowded spaces an extremely difficult 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. 9.1a where each of the six groups of people consists of one or two people in dark clothing. Based on visual 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. Figure 9.1b 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, and thus should be exploited for improving the matching of individual person.

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

Fig. 9.1 Advantages from and challenges in associating groups of people versus person re-identification in isolation

relative positions of the members can change significantly and frequently. (2) Although occlusion by other objects is less of 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. 9.1c.

Due to these challenges, conventional representations and matching methodsfor person re-identification are not suitable for solving the group association problem, because they are designed for person in isolation rather than in a group. In this work, a novel people group representation is presented based on two ratio-occurrence descriptors. This is in order to make the representation robust against within-group position changes. Given this group representation, a group matching algorithm is formulated to achieve group association robustness against both changes in relative positions of people within a group and variations in illumination and viewpoint across different camera views. 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. This group association model is validated using 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS) dataset captured by multiple camera views from a busy airport arrival hall [7].

The remaining sections are as follows. Section 9.2 overviews related work and assesses this work in context. Section 9.3 describes how the visual appearance of a group of people can be represented for robust group matching. Section 9.4 introduces the metric we used to measure the similarity between two group images and Sect. 9.5 formulates a method for utilising a group of people as contextual cues for individual person re-identification within the group. Section 9.6 presents experimental validation on these methods and Sect. 9.7 concludes the chapter.

9.2 Related Work

Contemporary work on person re-identification focuses on either finding distinctive visual appearance feature representations or learning discriminant models and matching distance metrics. Popular feature representations include colour histogram [4], principal axis histogram [2], rectangle region histogram [8], graph representation [9], spatial co-occurrence representation [3], and multiple feature based representation [4, 10]. For matching visual features of large variations due to either intra-class (same person) or inter-class (different people) appearance change [11], a number of methods have been reported in the literature including Adaboost [4], Primal RankSVM [12] and Relative Distance Comparison (RDC) [11]. These learning-based matching distance metric methods have shown to be effective for performing person re-identification regardless of a chosen feature representation. For assessing the usefulness of utilising group association as proximity context for person re-identification, the RankSVM-based matching method [12] is adopted in this work for matching individuals in a group.

The concept of exploiting contextual information for object recognition has been extensively studied in the literature. Most existing context modelling works require manual annotation/labelling of contextual information. Given both the annotated target objects and contextual information, one of the most widely used methods is to model the co-occurrence of context and object. Torralba et al. [13], Rabinovich et al. [14], Felzenszwalb et al. [15] and Zheng et al. [16] model how a target object category co-occurs frequently with other object categories (e.g. a person carrying a bag, a tennis ball with a tennis bracket) or where the target objects tend to appear (e.g. a TV in a living room). Besides co-occurrence information, spatial relationship between objects and context has also been explored for context modelling. Spatial relationships are typically modelled using Markov Random Field (MRF) or Conditionally Random Field (CRF) [17–19], or other graphical models [20]. These models incorporate the spatial support of target object against other objects either from the same category or from different categories and background, such as a boat on a river/sea or a car on a road. Based on a similar principle, Hoim et al. [21] and Bao et al. [22] proposed to infer the interdependence of object, 3D spatial geometry and the orientation and position of camera as context; and Galleguillos et al. [23] inferred the contextual interactions at pixel, region and object levels and combined them together using a multi-kernel learning algorithm [23, 24]. Comparing to those works, this work has two notable differences: (1) we focus on the problem of intra-category identification (individual re-identification whilst all people look alike) rather than inter-category classification (differentiating different object classes between for instance cars and bicycles); (2) we are specifically interested in exploring a group of people as non-stationary proximity context to assist in the matching of one of the individuals in the group.

There are other related works on crowd detection and analysis [25–28] and group activity recognition [29, 30]. However, those works are not concerned with

group association over space and time, either within the same camera view or across different views. A preliminary version of this work was reported in [31].

9.3 Group Image Representation

Given a gallery set and a probe set of images of different groups of people, we aim to design and construct suitable group image descriptors for matching gallery images with any probe image of a group of people.

9.3.1 From Pixel to Local Region-Based Feature Representation

Similar to [3, 32], 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 together with gradient information. Due to the change in camera view and varying positions and motions of a group of people, we consider that integration of local rotational invariant features and colour density information is better for constructing visual words for indexing. In particular, we extract SIFT features [33] (a 128-dimensional vector) for each RGB channel at each pixel with a surrounding support region (12×12 in our experiment). We also obtain an average RGB colour vector of pixel over a support region (3×3), where the colour vector is normalised 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 quantised into n clusters by K-means and a code book $\mathcal{A}$ 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.[1]

To represent the distribution of visual words of any image, a single histogram of visual words, which we call the holistic histogram, can be considered [34]. However autoedited1, this representation loses all spatial distribution information about 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 [35]. However, this representation will still be sensitive to the appearance changes in situations when people swap their positions in a group (Fig. 9.1c). Moreover, corresponding image grid positions between two group images are not always guaranteed to represent foreground regions, thus such a hard-wired grid block-based representation is not suitable.

Considering such characteristics of group images, we consider to represent a group image by constructing two local region-based descriptors: a *center rectangular ring*

[1] This step is omitted when continuous image sequences are not available.

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 combined to form a group image representation. These two descriptors are motivated by the observation that whilst global spatial relationships between people within a group can be highly unstable, local spatial relationships between small patches within a local region is more stable, e.g. within the bounding box of a person.

9.3.2 Center Rectangular Ring Ratio-Occurrence (CRRRO)

Rectangular ring regions are considered approximately rotationally invariant. An efficient integral computation of visual words histogram is also available [32]. Given both, we define a holistic rectangular ring structure expanding from the centre 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. 9.2a with $\ell = 3$), where the group image is of size $M \times N$. Such 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 scales (Fig. 9.3).

After a partition of any image for representation, a common approach to constructing a codebook is to concatenate the histogram of visual words from each ring. However, this ignores any spatial relationships existing between visual words from different ring-zones of a partition. We consider retaining such spatial relationships critical, thus we 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

$$\mathbf{H}_i(a, b) = \frac{\mathbf{h}_i(a)}{\mathbf{h}_i(a) + \mathbf{h}_i(b) + \varepsilon}, \tag{9.1}$$

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 relationships between visual words within and outside region P_i, we further define another two ratio occurrence maps for ring-region P_i as follows:

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

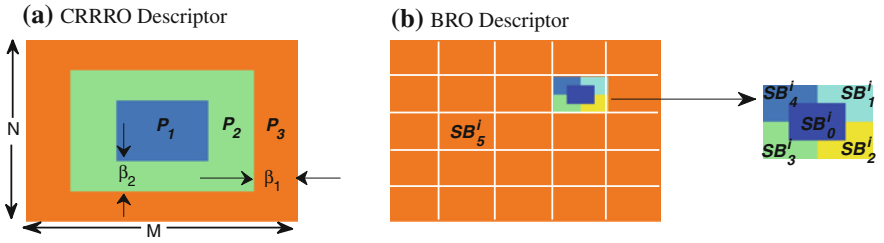

Fig. 9.2 Partition of a group image by two descriptors. *Left* the center rectangular ring ratio-occurrence descriptor ($\beta_1 = M/2\ell, \beta_2 = N/2\ell, \ell = 3$); *Right* the block based ratio-occurrence descriptor ($\gamma = 1$), where *white lines* show the grids of the image

Fig. 9.3 An illustration of a group of people against *dark* background

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:

$$\mathbf{G}_i(a,b) = \frac{\mathbf{g}_i(a)}{\mathbf{g}_i(a) + \mathbf{h}_i(b) + \varepsilon}, \quad \mathbf{S}_i(a,b) = \frac{\mathbf{s}_i(a)}{\mathbf{s}_i(a) + \mathbf{h}_i(b) + \varepsilon}. \tag{9.2}$$

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 show in the experiments 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.

9.3.3 Block-Based Ratio-Occurrence (BRO)

The CRRRO descriptor introduced 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[2] 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 potentially more robust) spatial relationships between visual words in each foreground block by further dividing the block into small block regions using L-shaped partition [3] with a modification that the most inner four block regions are merged (Fig. 9.2b). 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 $SB_0^i, \cdots, SB_{4\gamma}^i$ for some positive integer γ.

For associating groups of people over different views, we first note that not all blocks B_i appear in the same position in the group images. For example, a pair of people may swap their positions resulting in the blocks corresponding to those foreground pixels changing 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, region $SB_{4\gamma+1}^i$, which is the image portion outside block B_i (see Fig. 9.2b with $\gamma = 1$). Therefore, for each block B_i, we partition the group image into $SB_0^i, SB_1^i, \cdots, SB_{4\gamma}^i$ and $SB_{4\gamma+1}^i$. We show in the experiments that including such complementary region $SB_{4\gamma+1}^i$ would significantly enhance matching performance.

Similar to the CRRRO descriptor, for each block B_i, we learn an intra ratio-occurrence map $\mathbf{H}_j^i$ between visual words in each block region SB_j^i. Similarly, we explore an inter ratio-occurrence map $\mathbf{O}_j^i$ between different block regions SB_j^i. Since the size of each block region in block B_i would always be relatively much smaller than the complementary region $SB_{4\gamma+1}^i$, the ratio information between them will be sensitive to noise. Consequently, we consider two simplified inter ratio-occurrence maps $\mathbf{O}_j^i$ between block B_i and its complementary region $SB_{4\gamma+1}^i$ formulated as follows:

$$\mathbf{O}_1^i(a,b) = \frac{\mathbf{t}_i(a)}{\mathbf{t}_i(a) + \mathbf{z}_i(b) + \varepsilon}, \quad \mathbf{O}_2^i(a,b) = \frac{\mathbf{z}_i(a)}{\mathbf{z}_i(a) + \mathbf{t}_i(b) + \varepsilon}, \tag{9.3}$$

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

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

[2] A foreground block is defined as an image block with more than 70 % pixels being foreground.

9.4 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:

$$d(\mathbf{I}_1, \mathbf{I}_2) = d_r\left(\{\mathbf{T}_r^i(\mathbf{I}_1)\}_{i=1}^{\ell}, \{\mathbf{T}_r^{i'}(\mathbf{I}_2)\}_{i'=1}^{\ell}\right) + \alpha \cdot d_b\left(\{\mathbf{T}_b^i(\mathbf{I}_1)\}_{i=1}^{m_1}, \{\mathbf{T}_b^{i'}(\mathbf{I}_2)\}_{i'=1}^{m_2}\right), \quad \alpha \geq 0, \tag{9.4}$$

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. Note that L_1 norm metric is more robust and tolerant to noise as compared to Euclidean metric [36]. For d_b, since the spatial relationship between patches is not constant 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 exist 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:

$$d_b\left(\{\mathbf{T}_b^i(\mathbf{I}_1)\}_{i=1}^{m_1}, \{\mathbf{T}_b^{i'}(\mathbf{I}_2)\}_{i'=1}^{m_2}\right) = \min_{\mathbf{C},\mathbf{D}}\left\{k^{-1} \cdot ||\mathbf{AC} - \mathbf{BD}||_1\right\}, \quad \mathbf{A} \in \mathbb{R}^{q \times m_1}, \mathbf{B} \in \mathbb{R}^{q \times m_2}, \mathbf{C} \in \mathbb{R}^{m_1 \times k}, \mathbf{D} \in \mathbb{R}^{m_2 \times k}, \tag{9.5}$$

where the ith (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 number of foreground blocks in two group images, may be unequal. Generally, directly solving Eq. (9.5) is hard. Note that $\min_{\mathbf{C},\mathbf{D}}\left\{||\mathbf{AC} - \mathbf{BD}||_1\right\} \leq \sum_{j=1}^{k} \min_{\mathbf{c}_j,\mathbf{d}_j}\left\{||\mathbf{A}\mathbf{c}_j - \mathbf{B}\mathbf{d}_j||_1\right\}$ 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.

9.5 Exploring Group Context in Person Re-identification

9.5.1 Re-identification by Ranking

Person re-identification can be casted as a ranking problem [11, 12, 37], by which the problem is further addressed either in terms of feature selection or matching distance metric learning. This approach aims to learn a set of most discriminant and robust features, based on which a weighted L1 norm distance is used to measure the similarity between a pair of person images.

More specifically, person re-identification by ranking the relevance of their image features can be formulated as follows: There exist a set of relevance scores $\lambda = \{r_1, r_2, \cdots, r_\rho\}$ such that $r_\rho \succ r_{\rho-1} \succ \cdots \succ r_1$ where ρ is the number of scores and $\succ$ indicates the order. Most commonly, this problem only has two relevance considerations: relevant and related irrelevant observation feature vectors, that is, the correct and incorrect (but may still be visually similar) matches. Given a dataset $X = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ where $\mathbf{x}_i$ is a multi-dimensional feature vector representing the appearance of a person captured in one view, y_i is its label and m is the number of training samples. Each vector $\mathbf{x}_i (\in R^d)$ has an associated set of relevant observation feature vectors $\mathbf{d}_i^+ = \{\mathbf{x}_{i,1}^+, \mathbf{x}_{i,2}^+, \cdots, \mathbf{x}_{i,m^+(\mathbf{x}_i)}^+\}$ and irrelevant observation feature vectors $\mathbf{d}_i^- = \{\mathbf{x}_{i,1}^-, \mathbf{x}_{i,2}^-, \cdots, \mathbf{x}_{i,m^-(\mathbf{x}_i)}^-\}$ corresponding to correct and incorrect matches from another camera view, respectively. Here $m^+(\mathbf{x}_i)$ and $m^-(\mathbf{x}_i)$ are the respective numbers of relevant and irrelevant observations for query $\mathbf{x}_i$. In general, $m^+(\mathbf{x}_i) << m^-(\mathbf{x}_i)$ because there are likely only a few instances of correct match and many incorrect matches. The goal of ranking any paired image relevance is to learn function δ for all pairs of $(\mathbf{x}_i, \mathbf{x}_{i,j}^+)$ and $(\mathbf{x}_i, \mathbf{x}_{i,j'}^-)$ such that the relevant score $\delta(\mathbf{x}_i, \mathbf{x}_{i,j}^+)$ is larger than $\delta(\mathbf{x}_i, \mathbf{x}_{i,j'}^-)$.

Here, we seek to compute the score δ in terms of the pairwise sample $(\mathbf{x}_i, \mathbf{x}_{i,j})$ by a linear function $\mathbf{w}$ as follows:

$$\delta(\mathbf{x}_i, \mathbf{x}_{i,j'}) = \mathbf{w}^\top |\mathbf{x}_i - \mathbf{x}_{i,j}|, \tag{9.6}$$

where $|\mathbf{x}_i - \mathbf{x}_{i,j}| = \left[|\mathbf{x}_i(1) - \mathbf{x}_{i,j}(1)|, \cdots, |\mathbf{x}_i(d) - \mathbf{x}_{i,j}(d)|\right]^\top$. We call $|\mathbf{x}_i - \mathbf{x}_{i,j}|$ the absolute difference vector.

Note that for a query feature vector $\mathbf{x}_i$, we wish to have the following rank relationship for a relevant feature vector $\mathbf{x}_{i,j}^+$ and a related irrelevant feature vector $\mathbf{x}_{i,j'}^-$

$$\mathbf{w}^\top (|\mathbf{x}_i - \mathbf{x}_{i,j}^+| - |\mathbf{x}_i - \mathbf{x}_{i,j'}^-|) > 0. \tag{9.7}$$

Let $\hat{\mathbf{x}}_s^+ = |\mathbf{x}_i - \mathbf{x}_{i,j}^+|$ and $\hat{\mathbf{x}}_s^- = |\mathbf{x}_i - \mathbf{x}_{i,j'}^-|$. Then, by going through all samples $\mathbf{x}_i$ in the dataset X, we obtain a corresponding set of these pairwise relevant difference vectors denoted by $P = \{(\hat{\mathbf{x}}_s^+, \hat{\mathbf{x}}_s^-)\}$ where $\mathbf{w}^\top (\hat{\mathbf{x}}_s^+ - \hat{\mathbf{x}}_s^-) > 0$ is expected. A RankSVM model then is defined as the minimisation of the following objective function:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{s=1}^{|P|}\xi_s$$
$$\text{s.t. } \mathbf{w}^\top(\hat{\mathbf{x}}_s^+ - \hat{\mathbf{x}}_s^-) \geq 1 - \xi_s,\ s = 1, \cdots, |P|,\ \xi_s \geq 0,\ \ s = 1, \cdots, |P|, \quad (9.8)$$

where C is a parameter that trades margin size against training error.

A computational difficulty in using a SVM to solve the ranking problem is the potentially large size of P. In problems with lots of queries and/or queries represented as feature vectors of high dimensionality, the size of P means that forming the $\hat{\mathbf{x}}_s^+ - \hat{\mathbf{x}}_s^-$ vectors becomes computationally challenging. In the case of person re-identification, the ratio of positive to negative observation samples is $m : m \cdot (m-1)$ and as m increases the size of P can become very large rapidly. Hence, the RankSVM in Eq. (9.8) can be computationally intractable for large-scale constraint problems due to memory usage.

Chapelle and Keerthi [38] proposed primal RankSVM that relaxes the constrained RankSVM and formulated a non-constraint model as follows:

$$\mathbf{w} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{s=1}^{|P|}\ell\left(0, 1 - \mathbf{w}^\top\left(\hat{\mathbf{x}}_s^+ - \hat{\mathbf{x}}_s^-\right)\right)^2, \quad (9.9)$$

where C is a positive importance weight on the ranking performance and ℓ is the hinge loss function. Moreover, a Newton optimisation method is introduced to reduce the training time of the SVM. Additionally, it removes the need for an explicit computation of the $\hat{\mathbf{x}}_s^+ - \hat{\mathbf{x}}_s^-$ pairs through the use of a sparse matrix. However, in the case of person re-identification the size of the training set can also be a limiting factor. The effort required to construct all the $\hat{\mathbf{x}}_s^+$ and $\hat{\mathbf{x}}_s^-$ for model learning is determined by the ratio of positive to negative samples as well as the feature dimension d. As the number of related observation feature vectors is increased, i.e. more people are observed, the space complexity (memory cost) of creating all the training samples is

$$O\left(\sum_{i=1}^{m} d \cdot m^+(\mathbf{x}_i) \cdot m^-(\mathbf{x}_i)\right), \quad (9.10)$$

where $m^-(\mathbf{x}_i) = m - m^+(\mathbf{x}_i) - 1$ for the problems addressed here.

9.5.2 Re-identification with Group Context

We wish to explore group information for reducing the ambiguity in person re-identification if a person stays in the same group. Suppose a set of L paired samples $\{(\mathbf{I}_p^i, \mathbf{I}_g^i)\}_{i=1}^L$ is 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 in Sect. 9.3.1, with a 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 $\max\{M - a_1 - 0.5 \cdot M_1, a_1 - 0.5 \cdot M_1\}/(\ell - 1)$ and $\max\{N - b_1 - 0.5 \cdot N_1, b_1 - 0.5 \cdot N_1\}/(\ell - 1)$ thick along the horizontal and vertical directions, where (a_1,b_1) is the centre 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.

To integrate group information for person re-identification, in this work, we adopt to combine the distance metric d_p of a pair of person descriptors and the distance metric d_r of the corresponding group context descriptors computed from a probe and gallery image pair to be matched. 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:

$$d(\mathbf{I}_p^1, \mathbf{I}_p^2) = d_p(\mathbf{P}_1, \mathbf{P}_2) + \beta \cdot d_r(\mathbf{T}_1, \mathbf{T}_2), \quad \beta \geq 0, \tag{9.11}$$

where d_r is defined in Sect. 9.4 and d_p is formulated as

$$d_p(\mathbf{P}_1, \mathbf{P}_2) = -\delta(\mathbf{P}_1, \mathbf{P}_2) = -\mathbf{w}^\top |\mathbf{P}_1 - \mathbf{P}_2|, \tag{9.12}$$

where $\mathbf{w}$ is learned by RankSVM as described in Sect. 9.5.1.

For making use of group context in assisting person re-identification, we consider the following processing steps:

1. Detect a target person;
2. Extract features for each person and measure its distance from the gallery person images using the ranking distance in Eq. (9.12);
3. Segment the group of people around a detected person;
4. Represent each group of people using the group descriptor described in Sect. 9.3;
5. Measure the distance of each group descriptor from the group images to their corresponding gallery person images using the matching distance given in Sect. 9.4;
6. Combine the two distances using Eq. (9.11).

In computing the group descriptors, we focus on demonstrating the effectiveness of the proposed group descriptors and the group assisted matching model for improving person re-identification. We consider that person detection and the segmentation of groups of people in steps (1) and (3) above are performed using standard techniques readily available.

9.6 Experiments

We conducted extensive experiments using the 2008 i-LIDS MCTS dataset to evaluate the feasibility and performance of the proposed methods for associating groups of people and for person re-identification assisted by group context in a crowded public space.

9.6.1 Dataset and Settings

The i-LIDS MCTS dataset was captured at an airport arrival hall by 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 four images, either from different camera views or from the same camera but captured at different locations at different time. These group images are of different sizes. From the group images, we extracted 476 person images for 119 pedestrians, most of which are with four images. All person images were normalised to 64×128 pixels. Different from other person re-identification datasets [1, 3, 4], these images were captured by non-overlapping camera views, 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 Sect. 9.3, 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 CRRRO descriptor, we set $\ell = 3$. For the 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. (9.4)) was set to 0.8. For the colour histogram, the number of colour bins was set to 16.

9.6.2 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 [3] and the synthetic disambiguation rate (SDR) curve [4] 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 [4].

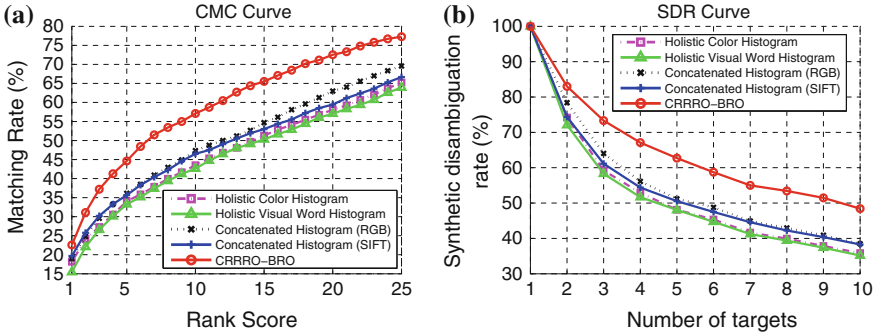

Fig. 9.4 Compare the CMC and SDR *curves* for associating groups of people using the proposed CRRRO–BRO descriptor with those from other commonly used descriptors

Fig. 9.5 Examples of associating groups of people: the 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. (9.4)) is shown in Fig. 9.4. We compare our model with two commonly used descriptors, colour histogram and visual word histogram of SIFT features (extracted at each colour channel) [34], which represent the distributions of colour or visual words of each group image holistically. We also applied these two descriptors to the designed center rectangular ring structure by concatenating the colour or visual word histogram of each rectangular ring. In order to make the compared descriptors scale invariant, the histograms used in the compared methods were normalised [39]. For measurement, the *Chi-square* distance χ^2 [39] was used.

Results in Fig. 9.4 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

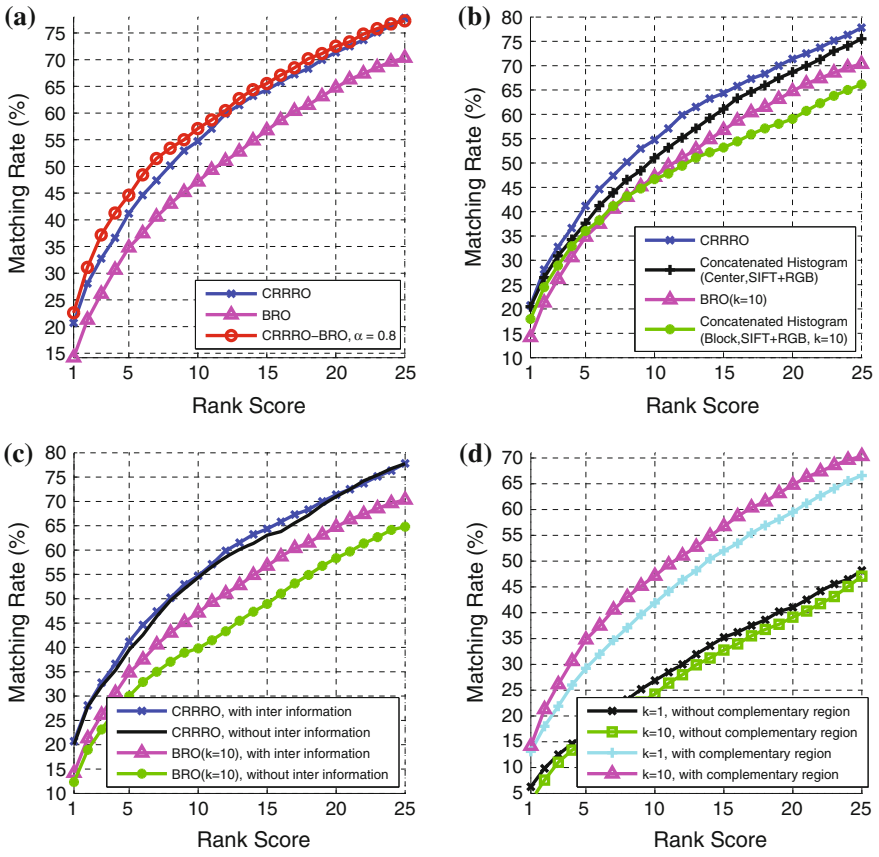

Fig. 9.6 Evaluation of the proposed group image descriptors

local region. Moreover, Fig. 9.6b shows that either using the proposed centre-based or block-based descriptor can still achieve an overall improvement as compared to the concatenated histogram of visual words using SIFT+RGB features (Sect. 9.3) 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. 9.5 shows some examples of associating groups of people using the proposed model (Eq. (9.4)) with $\alpha = 0.8$). It demonstrates that this 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 challenging viewing conditions, including significantly different view angles and severe occlusions.

9.6.3 Evaluation of Local Region-Based Descriptors

To give more insight into how the proposed local region-based group image descriptors perform individually and in combination, we show in Fig. 9.6a comparative results between the combination CRRRO–BRO (Eq. (9.4)) and the individual CRRRO and BRO descriptors using the metrics d_r and d_b as described in Sect. 9.4. It shows that the combination of the centre ring-based and local block-based descriptors utilises complementary information and improves the performance of each individual descriptor. Figure 9.6b evaluates the effects of using ratio map information as discussed above. Figure 9.6c shows that by exploring the inter ratio-occurrence between regions on the top of the intra one, an overall better performance is obtained as compared with a model without utilising such information. For the block-based ratio-occurrence descriptor, Fig. 9.6d indicates that including the complementary region with respect to each block B_i can reduce the ambiguity during matching.

9.6.4 Improving Person Re-identification by Group Context

RankSVM was adopted for matching individual person without using group context. To represent a person image, a mixture of colour and texture histogram features was used, similar to those employed by [4, 12]. Specifically, we divided a person image into six horizontal stripes. For each stripe, the RGB, YCbCr, HSV colour features and two types of texture features extracted by Schmid and Gabor filters were computed across different radiuses and scales, and totally 13 Schmid filters and 8 Gabor filters were obtained. In total, 29 feature channels were constructed for each stripe and each feature channel was represented by a 16-dimensional histogram vector. The details are given in [4, 12]. Each person image was thus represented by a feature vector in a 2,784-dimensional feature space $\mathcal{Z}$. Since the features computed for this representation include low-level features widely used by existing person re-identification techniques, this representation is considered as generic and representative. With group context, as described in Sect. 9.5.2, a two-rectangular-ring structure is expanded from the centre of the bounding box of each person, and the group matching score is fused with the RankSVM score, where we set $C = 0.005$ in Eq. (9.9).

For evaluating whether there is any benefit to re-identification when using group context information, we randomly selected all images of p people (classes) to set up the test set, and the images of the rest of the people (classes) were used for training. Different values of p were used to evaluate the matching performance of models learned with different amounts of training data. Each test set was composed of a gallery set and a probe set. The gallery set consisted of one image for each person, and the remaining images were used as the probe set. This procedure was repeated 10 times and the average performances of these techniques without and with group context are shown in Fig. 9.7. It is evident that including group context

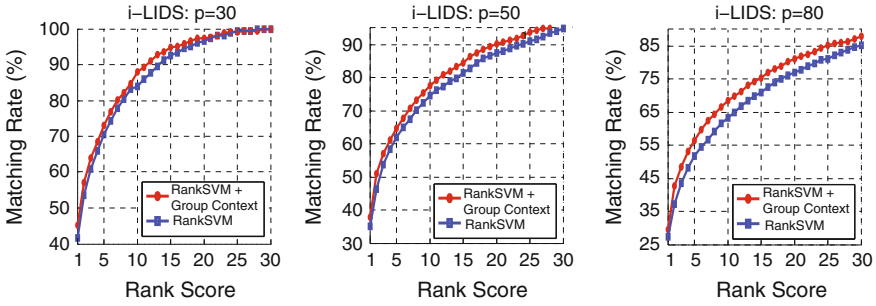

Fig. 9.7 Improving person re-identification using group context

notably improves the matching rate regardless of the choice of different person re-identification techniques. Although RankSVM has been shown in the literature as a very effective method for person re-identification, a clear margin of improvement is consistently achieved over the baseline RankSVM model when group context information is utilised. This suggests that group context helps alleviate the appearance variations due to occlusion, large variations in both view angle and illumination caused by non-overlapping multiple camera views.

9.7 Conclusions

In this work, we considered and addressed the problem of associating groups of people over multiple non-overlapping camera views and formulated local region-based group image descriptors in the form of both a centre rectangular ring and block based ratio-occurrence descriptors. They are designed specifically for the representation of images of groups of people in crowded public spaces. We evaluated their effectiveness using a top k-match distance model. Moreover, we demonstrated the advantages gained from utilising group context information in improving person re-identification under challenging viewing conditions using the 2008 i-LIDS MCTS dataset.

A number of future research directions are identified. First, both the grouping matching method and how the scores of group matching and person matching can benefit from further investigation, e.g. by exploiting more effective distance metric learning methods. Second, the problem of automatically identifying groups, especially groups of people who move together over a sustained period of time, needs to be solved more systematically in order to fully apply the presented method, e.g. by exploiting crowd analysis and modelling crowd flow patterns. Finally, dynamical contextual information inferred from groups can be further used to complement the method presented in this work, which is not utilised in the current model.

References

1. Gheissari, N., Sebastian, T.B., Tu, P.H., Rittscher, J., Hartley, R.: Person reidentification using spatiotemporal appearance. In: Procedings of the IEEE Conference on Computer Vision and Pattern Recognition (2006)
2. Hu, W., Hu, M., Zhou, X., Lou, J., Tan, T., Maybank, S.: Principal axis-based correspondence between multiple cameras for people tracking. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 663–671 (2006)
3. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.: Shape and appearance context modeling. In: Proceedings of the International Conference on Computer Vision (2007)
4. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of the European Conference on Computer Vision (2008)
5. Javed, O., Rasheed, Z., Shafique, K., Shah, M.: Tracking across multiple cameras with disjoint views. In: Proceedings of the International Conference on Computer Vision (2003)
6. Madden, C., Cheng, E., Piccardi, M.: Tracking people across disjoint camera views by an illumination-tolerant appearance representation. Mach. Vision Appl. **18**(3), 233–247 (2007)
7. HOSDB: Imagery library for intelligent detection systems (i-lids). In: Proceedings of the IEEE Conference on Crime and Security (2006)
8. Dollar, P., Tu, Z., Tao, H., Belongie, S.: Feature mining for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2007)
9. Gheissari, N., Sebastian, T., Hartley, R.: Person reidentification using spatiotemporal appearance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2006)
10. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. Comput. Vis. Image Underst. **117**(2), 130–144 (2013)
11. Zheng, W., Gong, S., Xiang, T.: Re-identification by relative distance comparison. IEEE Trans. Pattern Anal. Mach. Intell. **35**(3), 653–668 (2013)
12. Prosser, B., Zheng, W., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: Proceedings of the British Machine Vision Conference (2010)
13. Torralba, A., Murphy, K., Freeman, W., Rubin, M.: Context-based vision system for place and object recognition. In: Proceedings of the International Conference on Computer Vision (2003)
14. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: Proceedings of the International Conference on Computer Vision (2007)
15. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
16. Zheng, W., Gong, S., Xiang, T.: Quantifying and transferring contextual information in object detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**(4), 762–777 (2012)
17. Kumar, S., Hebert, M.: A hierarchical field framework for unified context-based classification. In: Proceedings of the International Conference on Computer Vision (2005)
18. Carbonetto, P., de Freitas, N., Barnard, K.: A statistical model for general contextual object recognition. In: Proceedings of the European Conference on Computer Vision (2004)
19. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. In : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2008)
20. Gupta, A., Davis, L.S.: Beyond nouns: exploiting prepositions and comparative adjectives for learning visual classifier. In: Proceedings of the European Conference on Computer Vision (2008)
21. Hoiem, D., Efros, A., Hebert, M.: Putting objects in perspective. Int. J. Comput. Vision **80**(1), 3–15 (2008)
22. Bao, S.Y.Z., Sun, M., Savarese, S.: Toward coherent object detection and scene layout understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 65–72 (2010)

23. Galleguillos, C., McFee, B., Belongie, S., Lanckriet, G.: Multi-class object localization by combining local contextual interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2010)
24. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: Proceedings of the IEEE International Conference on Computer Vision (2009)
25. Brostow, G.J., Cipolla, R.: Unsupervised bayesian detection of independent motion in crowds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2006)
26. Arandjelović, O.: Crowd detection from still images. In: Proceedings of the British Machine Vision Conference (2008)
27. Kong, D., Gray, D., Tao, H.: Counting pedestrians in crowds using viewpoint invariant training. In: Proceedings of the British Machine Vision Conference (2005)
28. Rabaud, V., Belongie, S.: Counting crowded moving objects. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2006)
29. Gong, S., Xiang, T.: Recognition of group activities using dynamic probabilistic networks. In: Proceedings of the International Conference on Computer Vision (2003)
30. Saxena, S., Brémond, F., Thonnat, M., Ma, R.: Crowd behavior recognition for video surveillance. In: Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems (2008)
31. Zheng, W., Gong, S., Xiang, T.: Associating groups of people. In: Proceedings of the British Machine Vision Conference (2009)
32. Savarese, S., Winn, J., Criminisi, A.: Discriminative object class models of appearance and shape by correlatons. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2006)
33. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **2**(60), 91–110 (2004)
34. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Proceedings of the European Conference on Computer Vision, International Workshop on Statistical Learning in Computer Vision (2004)
35. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2005)
36. He, R., Zheng, W.S., Hu, B.G.: Maximum correntropy criterion for robust face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **33**(8), 1561–1576 (2011)
37. Zheng, W., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 649–656. Colorado Springs (2011)
38. Chapelle, O., Keerthi, S.S.: Efficient algorithms for ranking with svms. Inf. Retrieval **13**(3), 201–215 (2010)
39. Fowlkes, C., Belongie, S., Chung, F., Malik, J.: Spectral grouping using the nystrom method. IEEE Trans. Pattern Anal. Mach. Intell. **26**(2), 214–225 (2004)