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 Electronic Engineering and Computer Science,
Queen Mary University of London, London E1 4NS, UK

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. Moreover, 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. 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.

Associating groups of people, however, 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).

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, 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* which aims to explore more specific local spatial information between visual words that could be stable. Compared with existing popular representations of visual words, a single histogram of visual words [1] would lose all spatial distribution information, and dividing the image into grid blocks and concatenating their histograms still cannot cope with a common case in group images when people swap their positions (see examples in Fig. 1 (c)) and moreover it is not always valid since the corresponding image grid blocks between two group images are not always guaranteed to be foreground regions.

Our descriptors are built on visual words extracted in a group image. We first assign a label to each pixel of a given group image $\mathbf{I}$. The label is a visual word index here. The visual words are the clusters of SIFT+RGB features. The SIFT+RGB feature is the concatenation of SIFT vector and colour vector, where SIFT features [2] are extracted for each RGB channel at each pixel and the colour vector is the an average RGB colour vector of that pixel over a support region. 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.

For the center rectangular ring ratio-occurrence (CRRRO) descriptor, a holistic rectangular ring structure is expanding from the center of a group image. The ℓ rectangular rings divide a group image into ℓ non-overlapped regions, $P_1, \cdots, P_\ell$. Every rectangular ring is $0.5 \cdot N/\ell$ and $0.5 \cdot M/\ell$ thick along the vertical and horizontal directions respectively, 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.

For the block based ratio-occurrence (BRO) descriptor, we first divide a group image into $\omega_1 \times \omega_2$ grid blocks $B_1, B_2, \cdots, B_{\omega_1 \times \omega_2}$, and only the foreground blocks are considered. For each block B_i, we aim to extract rather simple spatial relationships between visual words in it by further dividing the block into small block regions using L-shaped partition [4] with a modification that the most inner four block regions are merged (see Fig. 2 (b)). These small block regions are denoted as $SB^i_0, \cdots, SB^i_{4\gamma}$ for some positive integer γ. As not all blocks B_i appear at the same positions in the group images and there may be other visually similar blocks in the same group image, we therefore further include a complementary image region $SB^i_{4\gamma+1}$, the image portion outside block B_i (Fig. 2 (b) with $\gamma = 1$), for representation of block B_i to reduce the ambiguity during association.

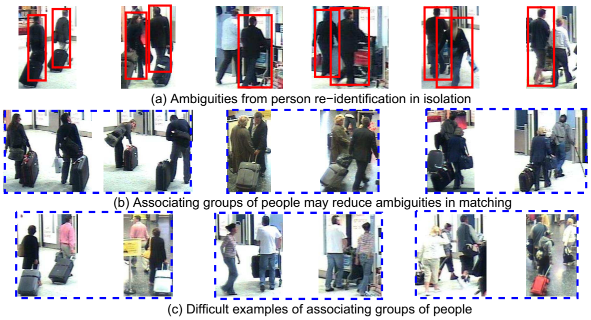
(a) Ambiguities from person re−identification in isolation
(b) Associating groups of people may reduce ambiguities in matching
(c) Difficult examples of associating groups of people

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

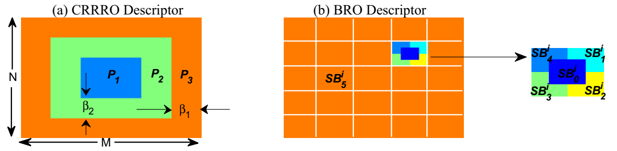

Figure 2: Partition of a group image by two descriptors. Left: the Center Rectangular Ring Ratio-Occurrence Descriptor ($\beta_1 = M/2\ell, \beta_2 = N/2\ell, \ell = 3$); Right: the Block based Ratio-Occurrence Descriptor ($\gamma = 1$), where white lines are to show the grids of the image.

The proposed descriptors CRRRO and BRO are finally constructed by exploring the corresponding *intra-* and *inter- ratio-occurrence maps*, which are notions introduced in this paper in order to explore the relationship between visual words within each zone and across different zones of a partition respectively. These two descriptors are then combined to represent a group image. Any two group images are matched by combining the distance metrics of the two proposed descriptors, where L_1 distance is used for CRRRO and a top-k match score metric is developed for BRO.

We conducted extensive experiments using the 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS) dataset [3], collected from a busy airport arrival hall, to evaluate the feasibility and performance of the proposed methods for associating groups of people in a crowded public space. Our results demonstrate compellingly the effectiveness of the proposed descriptor for group matching.

In addition, we also demonstrate a notable enhancement on individual person matching by utilising the group description as visual context. It is evidently shown that including group context improves notably the matching rate regardless the choice of person re-identification technique.

[1] 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.

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

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

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