Open-World Person Re-Identification by Multi-Label Assignment Inference

Brais Cancela[1]
brais.cancela@udc.es
Timothy M. Hospedales[2]
t.hospedales@qmul.ac.uk
Shaogang Gong[2]
s.gong@qmul.ac.uk

[1] VARPA Group,
Universidade da Coruña,
A Coruña, 15071, Spain

[2] School of EECS,
Queen Mary University of London,
London, E1 4NS, U.K.

The task of re-identification (ReID) is defined as the recognition of the same individual at different times and locations. State-of-the-art techniques share two very strong assumptions: *the total number of people in the scene is known* a priori, and there exists a total *overlap of identity* between a camera pair, that is, every person appears in both camera views. This is unrealistic for real-world re-identification scenarios, when there is no prior information about the same people reappearing in the scene at different views. We refer to this unconstrained setting as the 'open world' ReID problem. The open-world problem is more challenging for two reasons: (i) the total number of unique people within each camera and the scene as a whole (cross-cameras) are both unknown, and (ii) each subject may appear in some unknown subset of the cameras.

In this paper we consider for the first time the most general open-world re-identification problem. To address this, we introduce a new Conditional Random Field (CRF) model, making three important contributions: (1) No label information is needed a priori, allowing the system to detect when a new person enters the camera network; (2) An 'open world' solver, that is, the model does not assume that a person will (re)appear in every camera; and (3) Producing a person count as a byproduct. Our approach provides generality that is lacking in existing state of the art closed world ReID solutions.

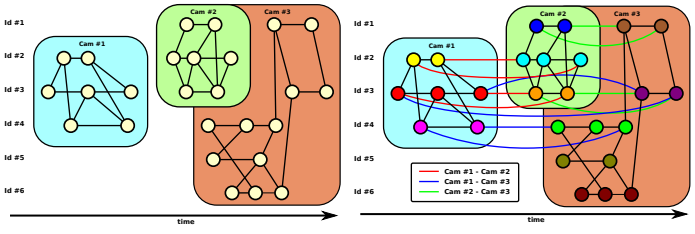
The objective of the CRF is to assign the most likely correct assignment of multiple id labels simultaneously to all the nodes in the CRF. We assume as input a set of N observations $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ across different camera views. Each observation $\mathbf{x}_i = \{c_i, t_i, \mathbf{p}_i, \mathbf{v}_i, \mathbf{a}_i\}$ consists of: A camera c_i making the detection; the time of detection t_i (we assume cameras are synchronized); the image position $\mathbf{p}_i$ and velocity $\mathbf{v}_i$ where the person was detected; and an appearance feature $\mathbf{a}_i$ from the detection bounding box. The re-identification task is to correctly assign identity labels $\mathcal{L} = \{l_i\}_{i=1}^{N}$, $l \in 1 \ldots L$ to all detections..

To address this task we propose a CRF $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where each node corresponds to a person detection (observation) $\mathcal{V} = \{v_i = x_i\}$. Each edge corresponds to a similarity between nodes/persons $\mathcal{E} = \{e_{ij} = (v_i, v_j)\}$, and the label of each node corresponds to the identity of that person/detection. Our aim is to find the set of labels $\mathcal{L}$ that best fits all the observations $\mathcal{X}$,

$$\mathcal{L}^* = \arg\min_{\mathcal{L}} \left(\sum_i U(l_i|\mathcal{X}) + \sum_{ij} B(l_i, l_j|\mathcal{X}) \right), \qquad (1)$$

where $U(l_i|\mathcal{X})$ and $B(l_i, l_j|\mathcal{X})$ denote unary and pairwise energy functions, respectively. Our algorithm proceeds in two steps, as explained in Algorithm 1. First, we solve the CRF allowing connections only between detections within the same camera. Second, we use that solution as an initial condition to build the connections between different cameras, creating the final CRF model. The structure and parameterisation of CRF at each stage is the same. We only increase the information included.

To evaluate our contribution, we focus on the challenging SAIVT-Softbio database [1], that includes 150 people recorded using 8 different cameras. Different people appear in different subsets of the cameras.



(a) Step 1 (b) Step 2

Figure 1: CRF illustration. In the first step, only detections within the same camera are connected. In the second step, a restricted connection between cameras is allowed.

Input: Detections $\mathcal{X}$
Output: Associations between detections $\mathcal{L}$
begin
  Compute within camera weights W and U,
  Solve the CRF Eq (1) with Alpha-expansion
  Solve Initial Hungarian to obtain H,
  Compute across camera weights W and U
  Solve the CRF Eq (1) with Alpha-expansion
end

Algorithm 1: Overview of CRF algorithm for open-world ReID.

Table 1: Re-identification among three cameras from SAIVT (3, 5 and 8).

	F_1-Score	Precision	Recall
Naive RankSVM	26.2%	22,0%	42,1%
Naive KISS	29.5%	19.7%	**66.1%**
RankSVM+CRF	42.0%	**53.7%**	39.4%
KISS+CRF	**48.3%**	50.3%	49.8%

Our contribution is agnostic to the appearance feature, and the base pairwise matching model used. To test our methodology, we consider the ELF [4] feature along with RankSVM [2, 4] and KISS [3] pairwise models. Furthermore, spatial and temporal information are included as information between cameras. As we address the open world problem with no prior information about the number of people or their camera overlap, no existing models directly apply. For baselines, we therefore define a more conventional 'engineering' generalisation to open world based on thresholding pairwise RankSVM and KISS scores.

To evaluate the performance of open-world problems the conventional CMC metric is insufficient. We therefore apply statistical analysis techniques. Given the final and ground truth labels, $\mathcal{L}^*$ and $\mathcal{L}_{gt}$, we evaluate all pairs. If two nodes have the same label in $\mathcal{L}_{gt}$ and in $\mathcal{L}^*$, it is a true positive; if they have different labels a true negative, and so on.

According to the obtained results (Table 1), our CRF model is more robust, as evidenced by its maintenance of high precision values. Moreover, it improves both of the base methods it is paired with. Because of the dichotomy between obtaining high precision and high recall, we conclude that the F-Score is the best overall metric to validate an open-world ReID algorithm.

A byproduct of open-world inference is a person count. Table 2 shows the estimated number of unique people among the approximately 600 detections across all three cameras. The estimated number of people along with the standard deviation of the estimate over multiple runs are given. In each case our framework improves on the baseline result, with KISS+CRF obtaining the best and most stable estimate.

Table 2: Inferring the number of distinct people in the dataset.

GT	Naive RankSVM	Naive KISS	RankSVM+CRF	KISS+CRF
48	61 ± 17.6	57.8 ± 11.2	65 ± 13.2	$\mathbf{54.1} \pm \mathbf{7.9}$

[1] Alina Bialkowski, Simon Denman, Patrick Lucey, Sridha Sridharan, and Clinton B Fookes. A database for person re-identification in multi-camera surveillance networks. In *DICTA*, 2012.

[2] Thorsten Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD*, pages 217–226, 2006.

[3] M Kostinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[4] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by support vector ranking. In *BMVC*, 2010.