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.

Abstract

Person re-identification methods have recently made tremendous progress on maximizing re-identification accuracy between camera pairs. However, this line of work mostly shares an critical limitation - it assumes re-identification in a 'closed world'. That is, between a known set of people who all appear in both views of a single pair of cameras. This is clearly far from a realistic application scenario. In this study, we take a significant step toward a more realistic 'open world' scenario. We consider associating persons observed in more than two cameras where: multiple within-camera detections are possible; different people can transit between different cameras – so that there is only partial and unknown *overlap of identity* between people observed by each camera; and the total number of unique people among all cameras is itself unknown. To address this significantly more challenging open world scenario, we propose a novel framework based on online Conditional Random Field (CRF) inference. Experiments demonstrate the robustness of our approach in contrast to the limitations of conventional approaches in the open world context.

1 Introduction

The task of re-identification (ReID) is often defined as the recognition of the same individual at different times and locations, which may involve different cameras, views, poses and lighting. This challenge is now widely studied by the computer vision community, due to its fundamentally challenging nature, and important practical role underpinning many visual surveillance functionalities including person search and tracking across disjoint cameras.

Re-identification studies generally frame the task as a closed set matching problem. Given a predefined 'gallery' set of known individuals, systems try to label each new 'probe' detection with the identity of the matching gallery individual. Studies have investigated good feature representations [4] and discriminative models [9] to maximise the chance of correct matching. They considered the contexts of single-shot [4, 9] (one image per person per camera) as well as multi-shot [12] (a series of images per person per camera, obtained from tracking) scenarios. However, most studies 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. Although this constrained framing of the ReID problem is a good starting point, it 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.

The closed-world problem is significantly simpler, because it can be divided into a series of independent tasks: "*For each probe person, find the top most similar in the gallery*". In the unconstrained variant, if there are two cameras people may only be seen by one; or if there are more than two, then people may appear in any subset of the cameras. This means there are more possible outcomes (no match), and every unknown identity problem is no longer independent, they become strongly inter-related. For example, consider intuitively the task of trying to match a person with a red-shirt against a gallery in the conventional closed world context. The match is simply the one whose shirt most clearly red. In the open world scenario, these could be completely separate people if two distinct red-shirt people were observed independently in each camera and not in the other. Moreover, if there are two red-shirt probes: in a closed world context, these would be given as distinct. In an open-world context there is additional ambiguity: Are they distinct people, or due to a broken track? The classical approaches clearly make too strong assumptions for this type of scenario.

In this paper we consider for the first time the most general open-world re-identification problem, where there is no prior information about the number of people or their overlap of identity across cameras. To address this, we introduce a new Conditional Random Field (CRF) model, overcoming the entailed challenges of effective graph construction, local optima and efficient inference. Our framework can answer qualitatively more general queries than existing re-identification systems such as: "*How many people are in the scene?*", "*If a person leaves a camera, which other cameras did he appear in, or did he simply disappear?*".

1.1 Related Work

Closed World Re-Identification: There has now been extensive work on closed-world re-identification (ReID). Studies have generally addressed good feature representations [4, 15, 16, 20] and/or learning matching models discriminatively [7, 9, 15, 16]. Most works have considered the 'single-shot' scenario of exactly one image per person using datasets like VIPeR [7]; while others considered 'multi-shot' – how to constructively aggregate information from multiple detections/shots of each person that might be obtained from tracking – using datasets like ETHZ [3]. Further review is beyond scope of this work, so we point the reader to a recent book [6] and survey [18] that summarise the main issues [5].

Towards Open World ReID: Going beyond closed world ReID discussed above, a few recent studies have begun to consider some open-world aspects of ReID. For instance, [12] introduced a CRF model to address multi-shot re-identification when the shots are not assumed to be correctly pre-grouped within each camera: Corresponding to realistic input with track association errors and split detections. Temporal information from each shot is used to restrict the connections between the nodes of the CRF. The system is only tested with the ETHZ dataset, which is recorded using a moving camera. However, pose and illumination variation is not high, and the more constrained assumption of full overlapping person sets is made. Recently, [11] introduces a probabilistic graphical model to associate within-camera trajectories across disjoint cameras. This model reasons generatively about

the appearance of each person, lighting change between cameras and the association between trajectories. Efficient Gibbs sampling is used to find the best solution. However, it still requires prior-knowledge of the number of people in the scene, and unlike [12], it assumes that within camera association is already performed perfectly. No existing work has considered the fully unconstrained open-world problem addressed in this work where within-camera re-identification is not assumed a-priori, person identities only partially overlap across two or more cameras, i.e. no guarantee of all people reappearing in every camera view, and the total number of persons is unknown. In [21], a transfer learning framework is defined to verify a probe person against a set of targets against a large amount of unlabeled data. However, this assumes the target and background people are split a-priori, it reasons about a single probe person at a time instead of jointly about all probes, and it only applies within two cameras.

Set Association: Although the open world scenario has not been addressed before, some existing algorithms are related to this challenge. The Hungarian Method (HM) [14] performs a set-match and can find the best pairwise correspondence between two sets of detections. It is a good solution for the closed world single-shot problem. However it will find an association even if the two sets are partially overlapped or totally disjoint (i.e. none reappearing). Thus even if every person in camera A does not appear in camera B, HM obtains a complete set of matches. Moreover, it cannot deal with multi-shot as it only makes 1:1 connections. We will exploit the HM to define a subset of credible matches for a global CRF to reason about. A classical CRF model [2] could also be used, with pairwise similarity measures to weight links between detections. However, several problems arise: (i) how to define the graph structure and label space, and (ii) CRFs tend to minimise the number of distinct labels used, thus tending to assign every detection to one identity. In this work, we develop a novel CRF model that incrementally constructs an appropriate graph to address these issues.

Our Framework: Contrary to classical ReID, the challenge in an 'open world' scenario also includes within-camera association, i.e. encompassing within-camera ambiguity due to tracking errors. An open world model not only has to distinguish when two detections belong to the same person, as in classical ReID, but also has to recognise when a new person enters in the scene, as in a classical tracking system. We build on CRFs, as they are state-of-the-art solvers for closely-related topics of re-identification [12] and tracking [19]. However, we relax the conventional constraint on requiring a priori set of known labels, and address issues in efficiency and convergence. Specifically, we introduce a novel two-step CRF model, that exploits spatio-temporal information where available. The first step matches within camera detections that belong to the same person. The second step considers both within and across-camera matching, using inter-camera information to revise initial within-camera estimates.

The proposed model makes 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 enables the flexibility lacking in existing state of the art closed world ReID solutions. Finally we also discuss some different evaluation criteria, as the classic Cumulative Matching Criteria (CMC) that assumes known number of people in a ReID scenario is no longer suitable.

2 A Framework for Open World Re-Identification

In this section we first formalise the task and our model representation. In this model, different candidates of people with unknown id labels are represented as nodes in a CRF. The

objective of the CRF is to infer the most likely correct assignment of multiple id labels simultaneously to all the nodes in the CRF (see Figure 1). 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 $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, with the following structure. 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 $\mathcal{X}$,

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

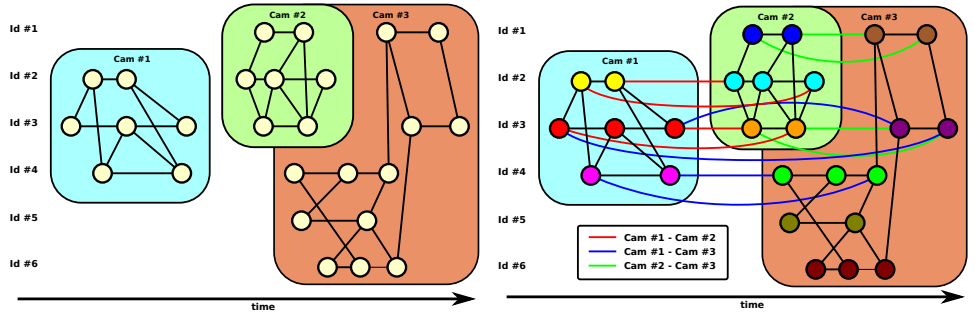
Here $U(l_i|\mathcal{X})$ and $B(l_i, l_j|\mathcal{X})$ denote unary and pairwise energy functions, respectively. U is an $L \times N$ matrix defining the cost of assigning any label l_i to any observation $\mathbf{x}_i$. Importantly, in the open world context, we do not know the total number of people in the network, so $L = N$ to account for the limiting case where every single detection is a unique person. B is also an $N \times N$ matrix, defining the cost of assigning l_i and l_j to a particular pair of observations. We decompose B into two matrices, $B(i, j) = W(i, j)C(\mathcal{L}(i), \mathcal{L}(j))$, where $W(i, j)$ is the weight of the similarity between the two nodes x_i and x_j and $C(\mathcal{L}(i), \mathcal{L}(j))$ is the cost of assigning the labels $\mathcal{L}(i)$ and $\mathcal{L}(j)$ to their respective nodes. We shall define W later, whilst C is a $N \times N$ matrix defined as

$$C(l_i, l_j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

As mentioned before, U is an $N \times N$ matrix. Thus, assuming that we have as many labels as person detections (observations), the cost of assigning any label l_j to any observation $\mathbf{x}_i$ is also a pairwise similarity measure between the observations $\mathbf{x}_i$ and $\mathbf{x}_j$. $B(i, j) = 0$ means there is no direct connection between the two detections. Non-zero values will depend on the appearance features, and spatio-temporal information (if available). The accuracy of pairwise correspondences is higher within the same camera than between cameras, due to less appearance change and stronger continuity. For this reason, our algorithm proceeds in two steps, as illustrated in Fig. 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, but additional information is included.

2.1 Label Assignment as Within-Camera Tracking

A characteristic of CRF models is that they try to reduce the number of labels in the output. For that reason, while creating a fully-connected CRF for solving the open-world re-identification problem is elegant, it is very hard to tune. Small variations in the pairwise potential causes every connected detection to be grouped with the same label, even if the similarity is low. Thus, we restrict the number of direct links between detections.



(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.

First, all the detections included in the observation set are sorted according to the time they were detected. Then, we establish the similarity between detections by creating the unary potential $\tilde{U}$, defined as

$$\tilde{U}(i,j) = \begin{cases} 0 & \text{if } i = j \\ 1 - \delta^c_{i,j} & \text{if } |t_i - t_j| < \tau_c \text{ and } c_i = c_j \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

where $\delta^c_{i,j} \in [0,1]$ is the probability of assigning label l_i to the detection x_j in camera c. Similarly, the pairwise weight $\tilde{W}$ is defined as

$$\tilde{W}(i,j) = \begin{cases} \left(1 - \dfrac{|t_i - t_j|}{\tau_c}\right) \alpha^c_{i,j} & \text{if } |t_i - t_j| < \tau_c \text{ and } c_i = c_j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where t_i and t_j are the times that detections x_i and x_j were recorded, respectively; $\alpha^c_{i,j} \in [0,1]$ is the appearance similarity between detections i and j in camera c; and τ_c a time threshold. Note that the strength between two detections decreases with the time gap similarly to [12].

As explained before, the number of connections between the nodes, using these matrices, is too high. A fully-connected CRF tends to use fewer labels, which is an undesirable property for our model. Thus, we reduce the number of direct connections to two at most for each detection based on higher $\tilde{W}(i,j)$ values. First, we define $\tilde{W}_w$ and $\tilde{U}_w$ as

$$\tilde{W}_w(i,j) = \begin{cases} \tilde{W}(i,j) & \text{if } |P| < 2, \text{ where } P = \{x \in N - \{i\}, \tilde{W}(i,x) > \tilde{W}(i,j)\} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$\tilde{U}_w(i,j) = \begin{cases} \tilde{U}(i,j) & \text{if } |P| < 2, \text{ where } P = \{x \in N - \{i\}, \tilde{W}(i,x) > \tilde{W}(i,j)\} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Two links per node is a good balance. This value can be modified, but we found this is a good connection density (higher values highly increase the false positive rate, whilst a lower value increase the false negative rate). To enforce symmetry, we define W_w and U_w as

$$W_w = \tilde{W}_w + \tilde{W}_w^T \qquad U_w = \tilde{U}_w + \tilde{U}_w^T \tag{7}$$

U_w, W_v and C define our CRF, which can be solved efficiently using the alpha-expansion algorithm [2]. At this point, we have connections between nodes (associated person detections) in the same camera, which are denoted by G. Next, we establish links across cameras.

Input: $U_w, W_w, \{H\}$
Output: U, W
begin
  $U = U_w, W = W_w, T = \emptyset$.
  foreach $c_1, c_2 \in |c|, c_1 \neq c_2$ **do**
    $[p,q] = \text{Hungarian}(H^{c_1,c_2})$.
    for $i = 1..|p|$ **do**
      if $H^{c_1,c_2}(p_i,q_i) > \alpha_t^{c_1,c_2}$ **then**
        $W(p_i,q_i) = W(q_i,p_i) = \frac{f^{c_1,c_2}}{\max(\{f^{c_i,c_j}\})}$.
        $T \cup (p_i, q_i, \frac{f^{c_1,c_2}}{\max(\{f^{c_i,c_j}\})})$.
      end
    end
  end
  for $i = 1..|T|$ **do**
    Without taking into account (p_i,q_i) connection:
    Select $S_i(p) | \forall j \in S_i(p)$, if exists a path between p_i and j using W
    Select $S_i(q) | \forall j \in S_i(q)$, if exists a path between q_i and j using W
    Update the states $U(S_i(q), S_i(p))$ and $U(S_i(p), S_i(q))$.
  end
end

Algorithm 1: Constructing unary and binary CRF potentials.

2.2 Cross-Camera Association

To simplify association across cameras, we only take into account direct connections between the first and the last appearance of a person in each camera. Let $\mathcal{L}_v$ be the labels associated with each node after using the local CRF model. Given the sorted detections, we create two sets B and E enclosing the first and the last label appearances, as follows:

$$\forall p \in [1..N] \quad p \in B \quad \text{if } \forall q \in [1..(p-1)], G(q) \neq G(p) \tag{8}$$

$$\forall p \in [1..N] \quad p \in E \quad \text{if } \forall q \in [(p+1)..N], G(q) \neq G(p) \tag{9}$$

Once we have these sets, we need to select which are the correct matches between the detections. With the same reasoning as before, we want to reduce the number of connections between detections. Assuming the labels obtained in the first step are correct, we can conclude that the final detection of each person in each camera occurs when the subject leaves the camera field of view. The same also happens with the initial detections. Based on this reasoning, we can conclude that every final detection in one camera is related with, at most, one detection in another camera. Thus, for each pair of cameras c_1 and c_2, we create the matrix H^{c_1,c_2}, which stores the affinity between detections i and j, as

$$H_{i,j}^{c_1,c_2} = \begin{cases} \beta_{i,j}^{c_1,c_2} & \text{if } c_i = c_1 \wedge c_j = c_2 \wedge ((i \in B \wedge j \in E) \vee (i \in E \wedge j \in B)) \\ \infty & \text{otherwise} \end{cases} \tag{10}$$

where β^{c_1,c_2} is a cross-camera pairwise person-affinity measure based on appearance and spatio-temporal cues. The lower the β value, the stronger connection. Using the Hungarian Method [14], we search for the most plausible assignment of correct labels. In other words, the Hungarian method is used to find a small subset of plausible links between detections in different cameras. The detected links are included in the CRF as explained in Algorithm 1. In the first loop, we compute the Hungarian Method to obtain the connections between nodes in different cameras. Then, for each pair of connected nodes, we remove that connection and we look for all the connections each node has. So, we obtain all the different states each node can have, without taking into account the new connection. Finally, we enable the connection and we update the unary potential of all the connected nodes, updating the new

Input: Detections $\mathcal{X}$
Output: Associations between detections $\mathcal{L}$
begin
  Compute within camera weights W and U (Eq. 7),
  Solve the CRF Eq (1) with Alpha-expansion [8]
  Solve Initial Hungarian to obtain H (Eq. 10),
  Compute across camera weights W and U (Alg. 1)
  Solve the CRF Eq (1) with Alpha-expansion [8])
end

Algorithm 2: Overview of CRF algorithm for open-world ReID.

states the nodes can reach with this new connection. The weight of this connection is further adapted by the expected quality/reliability of affinities computed between these two cameras according to the estimated F_1 score f^{c_1,c_2} across cameras:

$$W(i,j) = \begin{cases} \tilde{W}(i,j) & \text{if } c_i = c_j \\ \frac{f^{c_1,c_2}}{\max(\{f^{c_i,c_j}\})} & \text{if } c_i \neq c_j \text{ and } x_i \text{ and } x_j \text{ are linked} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

This is done because we want to rely more on connections between camera pairs that can match reliably, and less on unreliable pairs. Finally, we create the CRF using the matrices U, W and C. To solve this CRF, we use the alpha-expansion algorithm again. An overview of our two-step CRF algorithm is described in Algorithm 2.

2.3 Pairwise Affinity Measures

The model depends on pairwise within and across-camera similarity measures δ^c (Eq. 3), α^c (Eq. 4) and β^{c_1,c_2} (Eq. 10). These are all learned in the training step.

Within-camera: Various techniques can be used to compute similarities: $\delta^c, \alpha^c \in [0,1]$. For simplicity, we assume $\delta^c = \alpha^c$. To obtain these, we train a pairwise appearance-based person-similarity model $d^c(\cdot,\cdot)$ per camera. Let λ^+ be the set containing all the matching pairs, whereas λ^- the opposite; and $d^c(\mathbf{a}_i,\mathbf{a}_j)$ some pairwise distance metric (KISS or distance to the hyperplane in the RankSVM model). To normalise the distances for comparability across cameras, δ^c and α^c are then defined as

$$\alpha^c_{i,j} = \delta^c_{i,j} = \frac{|(\mathbf{a}_l,\mathbf{a}_m) \in \lambda^+, d^c(\mathbf{a}_l,\mathbf{a}_m) \leq d^c(\mathbf{a}_i,\mathbf{a}_j)|}{|(\mathbf{a}_l,\mathbf{a}_m) \in \lambda^+, d^c(\mathbf{a}_l,\mathbf{a}_m) \leq d^c(\mathbf{a}_i,\mathbf{a}_j)| + |(\mathbf{a}_n,\mathbf{a}_p) \in \lambda^-, d^c(\mathbf{a}_l,\mathbf{a}_m) \leq d^c(\mathbf{a}_i,\mathbf{a}_j)|} \tag{12}$$

Across-camera: To obtain the across-camera measure for cameras c_1 and c_2 with respective detections x_i and x_j, we compute KISS or RankSVM similarity measures: one for appearance ($d^{c_1,c_2}(\mathbf{a}_i,\mathbf{a}_j)$), and another one for the combination of both position and velocity ($\rho^{c_1,c_2}(\mathbf{p}_i;\mathbf{v}_i,\mathbf{p}_j;\mathbf{v}_j)$). We combine the two distances to obtain the similarity measure

$$\beta^{c_1,c_2}_{i,j} = \gamma^{c_1,c_2} d^{c_1,c_2}(\mathbf{a}_i,\mathbf{a}_j) + (1-\gamma^{c_1,c_2})\rho^{c_1,c_2}(\mathbf{p}_i;\mathbf{v}_i,\mathbf{p}_j;\mathbf{v}_j) \quad \gamma^{c_1,c_2} \in [0,1] \tag{13}$$

3 Experiments

Dataset: To evaluate our contribution, we need a dataset that reflects the open-world challenge. Many classic ReID datasets, such as VIPER or ETHZ assume total overlap of persons across cameras. PrID dataset [8] has a multiple-shot version with partial overlap, but it contains only two different cameras. Thus, we decide to focus on the challenging SAIVT-Softbio

Figure 2: SAIVT-SoftBio dataset. The dataset contains 150 people recorded over an eight camera network. It includes several angle orientations and sudden illumination changes.

database [1] (see Fig. 2), that includes 150 people recorded using 8 different cameras. To our knowledge, it is the only dataset that simultaneously meets all the requirements for a full open-world task: Multi-shot data and multiple cameras with camera-transition uncertainty.

Experimental Settings: Our contribution is agnostic to the appearance feature, and the base pairwise matching model used. To evaluate the system, we divide the dataset into 3 disjoint subsets. The first third (train set), is used to train all pairwise within and across-camera matching models, giving d and ρ in Eqs. (13)-(16). We consider the ELF [17] feature with RankSVM [10, 17] and KISS[1] [13] pairwise models. The second portion (calibration set) is used to calibrate all thresholds, α^c, δ^c, and β^{c_1,c_2} measures; and γ^{c_1,c_2} values. The best combination of these parameters is obtained by looking for the best F-score, denoted by f^c or f^{c_1,c_2}, depending if its within or between cameras. The final third is used to evaluate the performance. We average performance over 10 random splits.

Baselines: 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 of RankSVM [10, 17] and KISS [13]. We train both on the training set, and then use the calibration set to optimise the threshold for the pairwise affinity. Pairs with affinity over threshold are declared as sharing the same label. We denote these NaiveRankSVM and NaiveKISS.

Evaluation Metrics: To evaluate the performance of open-world problems the conventional CMC metric is insufficient, due to partial overlap of a variable number of labels and > 2 cameras. We therefore apply statistical analysis: Given the final and ground truth labels, $\mathcal{L}^*$ and $\mathcal{L}_{gt}$, we analyse all pairs. If two nodes have the same label in $\mathcal{L}_{gt}$ and in $\mathcal{L}^*$, it is a true positive. The same label in $\mathcal{L}^*$ and different in $\mathcal{L}_{gt}$, a false positive, and so on. As the number of negative pairs is very large, accuracy and specificity have high values (≈ 1). Precision (percentage of pair matches that are correct), recall (percentage of correct pair matches that are detected) and their combination, the F-score, are better measures to use.

3.1 Results

Open World Re-identification: We evaluate on SAIVT, using five images per person per camera. We consider three cameras (Cam 3, 5 and 8, that are challenging according to [1]), where a person that appears in one camera may or may not appear in the others. Table 1 shows results obtained from analyzing every possible pair (within and between cameras). We present variants of our framework using both RankSVM and KISS as the base pairwise models. The CRF results are based on global inference across all three cameras, however columns break down association performance as evaluated within individual cameras (first three), across each pair of cameras (middle three), and across all three cameras ("whole model"). The baseline methods obtain somewhat better recall, due to their non-conservative nature. However on the other hand, the low number of false negatives causes a huge incre-

[1] For KISS, we reduce the dimension of ELF to 100 with PCA, as it is not robust to high dimensional data

Table 1: Re-identification among three cameras from SAIVT. The last column shows the global performance. Other columns show local performance. E.g., C3-C8 shows the quality of the connections between camera 3 and camera 8 when the whole CRF model is computed.

F_1-Score	C3	C5	C8	C3 - C5	C3 - C8	C5 - C8	Whole model
Naive RankSVM	31.7%	34.1%	27.1%	15.9%	20.1%	24.6%	26.2%
Naive KISS	32.6%	29.4%	34.7%	23.4%	31.0%	29.6%	29.5%
RankSVM+CRF	50.1%	41.1%	**73.2%**	18.2%	43.4%	32.4%	42.0%
KISS+CRF	**57.3%**	**52.0%**	70.0%	**30.3%**	**47.6%**	**43.7%**	**48.3%**
Precision	**C3**	**C5**	**C8**	**C3 - C5**	**C3 - C8**	**C5 - C8**	**Whole model**
Naive RankSVM	30,2%	22,2%	36,7%	14,9%	27,7%	25,7%	22,0%
Naive KISS	22.0%	20.0%	22.0%	15.9%	20.7%	19.9%	19.7%
RankSVM+CRF	**63.8%**	**61.4%**	**62.3%**	37.2%	**55.4%**	45.2%	**53.7%**
KISS+CRF	56.4%	59.2%	58.5%	**38.0%**	48.4%	**47.1%**	50.3%
Recall	**C3**	**C5**	**C8**	**C3 - C5**	**C3 - C8**	**C5 - C8**	**Whole model**
Naive RankSVM	50,6%	**87,6%**	44,2%	24,7%	29,4%	43,4%	42,1%
Naive KISS	**70.1%**	63.3%	91.7%	**50.3%**	**70.1%**	**65.4%**	**66.1%**
RankSVM+CRF	47.4%	38.8%	**94.0%**	15.5%	43.1%	30.8%	39.4%
KISS+CRF	62.8%	50.1%	91.1%	28.5%	51.1%	44.7%	49.8%

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

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

ment in the number of false positives, resulting in significantly worse precision. 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 recall and precision, we conclude that the F-Score is the best overall metric to validate an open-world ReID algorithm.

Estimating the number of people: An important general question of interest to camera network operators is how many unique people are observed by the camera network in a given time period? This is implicit in the open world ReID task. Inference in our CRF model computes this as a byproduct[2], so we can answer this question directly. 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.

4 Conclusion

We have proposed the first method to address the most practical 'open world' variant of the re-identification problem. That is, when no information is provided a priori about the number or distribution of people. We develop a two-step CRF model using both appearance, temporal and spatial information that can be solved by fast energy minimization techniques using graph cuts. Evaluation on a challenging public dataset with three cameras demonstrates that the model improves on engineered baselines built on either of two classic pairwise ReID techniques. Moreover, important metadata such as person counts can be generated as a by-product of inference in our model. In our future work, we would like to test our algorithm with more cameras and build explicit person and camera lighting models.

[2]One assumption is made in this point: since we assume we are going to have more than 1 detection per person and per camera, labels with only one associated detection are treated as noise, and removed.

References

[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] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.

[3] Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and appearance for mobile scene analysis. In *ICCV*, 2007.

[4] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[5] Shaogang Gong, Marco Cristani, Chen Change Loy, and Timothy M. Hospedales. The re-identification challenge. In Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors, *Person Re-identification*, pages 1–20. Springer, 2014.

[6] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy, editors. *Person Re-Identification*. Springer, 2014.

[7] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*. 2008.

[8] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102. Springer, 2011.

[9] Martin Hirzer, Peter M Roth, Martin Köstinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*. 2012.

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

[11] Vijay John, Gwenn Englebienne, and Ben Krose. Solving person re-identification in non-overlapping camera using efficient gibbs sampling. In *BMVC*, 2013.

[12] Svebor Karaman and Andrew D Bagdanov. Identity inference: generalizing person re-identification scenarios. In *ECCV 2012. Workshops and Demonstrations*, 2012.

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

[14] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[15] Ryan Layne, Timothy M. Hospedales, and Shaogang Gong. Person re-identification by attributes. In *BMVC*, 2012.

[16] Ryan Layne, Timothy M. Hospedales, and Shaogang Gong. Re-id: Hunting attributes in the wild. In *BMVC*, 2014.

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

[18] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People re-identification in surveillance and forensics: a survey. *ACM Computing Surveys*, December 2013.

[19] Bo Yang and Ram Nevatia. An online learned crf model for multi-target tracking. In *CVPR*, 2012.

[20] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

[21] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Transfer re-identification: From person to set-based verification. In *CVPR*, 2012.