From re-identification to identity inference: labelling consistency by local similarity constraints

Svebor Karaman, Giuseppe Lisanti, Andrew D. Bagdanov, Alberto Del Bimbo

Abstract In this chapter we introduce the problem of identity inference as a generalization of person re-identification. It is most appropriate to distinguish identity inference from re-identification in situations where a large number of observations must be identified without knowing *a priori* that groups of test images represent the same individual. The standard single- and multi-shot person re-identification common in the literature are special cases of our formulation. We present an approach to solving identity inference by modeling it as a labeling problem in a Conditional Random Field (CRF). The CRF model ensures that the final labeling gives similar labels to detections that are similar in feature space. Experimental results are given on the ETHZ, i-LIDS and CAVIAR datasets. Our approach yields state-of-the-art performance for multi-shot re-identification, and our results on the more general identity inference problem demonstrate that we are able to infer the identity of very many examples even with very few labelled images in the gallery.

Key words: Re-identification, identity inference, conditional random fields, video surveillance.

1 Introduction

Person re-identification is traditionally defined as the recognition of an individual at different times, possibly imaged from different camera views and/or locations, and considering a large number of candidate individuals in a known gallery. It is a standard component of multi-camera surveillance systems as it is a way to associate multiple observations of the same individual over time. Particularly in scenarios in which the long-term behaviour of persons must be characterized, accurate re-identification is es-

Media Integration and Communication Center
University of Florence, Viale Morgagni 65, Florence, Italy
svebor.karaman@unifi.it, {bagdanov, delbimbo, lisanti}@dsi.unifi.it

sential. In realistic, wide-area surveillance scenarios such as airports, metro and train stations, re-identification systems should be capable of robustly associating a unique identity with hundreds, if not thousands, of individual observations collected from a distributed network of many sensors.

Re-identification performance has traditionally been evaluated as a retrieval problem. Given a gallery consisting of a number images of known individuals, for each test image or group of test images of an unknown person the goal of re-identification is to return a ranked list of individuals from the gallery. Configurations of the re-identification problem are generally categorized according to how much group structure is available in the gallery and test image sets. In a *single-shot image set* there is no grouping information available. Though there might be multiple images of an individual, there is no knowledge of which images correspond to that person. In a *multi-shot image set*, on the other hand, there is explicit grouping information available. That is, it is known which images correspond to the same individual, though of course the identities corresponding to each group are not known and the re-identification problem is to determine them.

The categorization of re-identification scenarios into multi- and single-shot configurations is useful for establishing benchmarks and standardized datasets for experimentation on the discriminative power of descriptors for person re-identification. However, these scenarios are not particularly realistic with respect to many real-world application scenarios. In video surveillance scenarios, for example, it is more common to have a few individuals of interest and to desire that all occurrences of them be labelled. In this case the number of unlabelled test images to re-identify is typically much larger than the number of gallery images available. Another unrealistic aspect of traditional person re-identification is its formulation as a retrieval problem. In most video surveillance applications, the accuracy of re-identification at rank-1 is the most critical metric and higher ranks are of much less interest.

Based on these observations, in this chapter we describe a generalization of person re-identification which we call *identity inference*. The identity inference formulation is expressive enough to represent existing single- and multi-shot scenarios, while at the same time also modelling a larger class of problems not considered in the literature. In particular, we demonstrate how identity inference models problems where only a few labelled examples are available, but where identities must be inferred for a large number of probe images. In addition to describing identity inference problems, our formalism is also useful for precisely specifying the various multi- and single-shot re-identification modalities in the literature. We show how a Conditional Random Field (CRF) can then be used to efficiently and accurately solve a broad range of identity inference problems, including existing person re-identification scenarios as well as more difficult tasks involving a lot of test images.

In the next section we review the literature on person re-identification. In section 3 we introduce our formulation of the identity inference problem and in section 4 propose a solution based on label inference in a Conditional Random Field. Section 5 contains a number of experiments illustrating the effectiveness of our approach for both the re-identification and identity inference problems. We conclude in section 6 with a discussion of our results.

2 Related work

Person re-identification has applications in tracking, target re-acquisition, verification and long-term activity modelling. The most popular approaches to person re-identification are *appearance-based* techniques which must overcome problems such as varying illumination conditions, poses changes and target occlusion. Within the broad class of appearance-based approaches to person re-identification, we distinguish *learning-based* methods, which generally require a training stage in which statistics of multiple images of each person is used to build a discriminative models of persons to be re-identified, from *direct* methods which require no initial training phase.

The majority of existing research on the person re-identification problem has concentrated on the development of sophisticated features for describing the visual appearance of targets. In [19] were introduced discriminative appearance-based models using Partial Least Squares (PLS) over texture, gradients and color features. The authors of [13] use an ensemble of local features learned using a boosting procedure, while in [1] the authors use a covariance matrix of features computed in a grid of overlapping cells. The SDALF descriptor introduced in [11] exploits axis symmetry and asymmetry and represents each part of a person by a weighted color histogram, maximally stable color regions and texture information from recurrent highly-structured patches. In [8] the authors fit a Custom Pictorial Structure (CPS) model consisting of head, chest, thighs and legs part descriptors using color histograms and Maximally Stable Color Region (MSCR). The Global Color Context (GCC) of [7] uses a quantization of color measurements into color words and then builds a color context modeling the self-similarity for each word using a polar grid. The Asymmetry-based Histogram Plus Epitome (AHPE) approach in [3] represents a person by a global mean color histogram and recurrent local patterns through epitomic analysis. A common feature of most appearance-based approaches is that they compute an aggregate or mean appearance model over multiple observations of the same individual (for multi-shot modalities).

The approaches mentioned above concentrate on feature representation and not specifically on the classification or ranking technique. An approach which does concentrate specifically on the ranking approach is the Ensemble RankSVM technique of [17], which learns a ranking SVM model to solve single-shot re-identification problems. The Probabilistic Distance Comparison (PRDC) approach [24] introduced a comparison model which maximizes the probability of a pair of correctly matched images having a smaller distance than that of an incorrectly matched pair. The same authors in [22] then model person re-identification as a transfer ranking problem where the goal is to transfer similarity observations from a small gallery to a larger, unlabelled probe set. Metric learning approaches in which the metric space is adapted to the gallery data have also been successfully applied recently to the re-identification problem [10, 16].

We believe that in realistic scenarios many unlabelled images will be available while only few detections with known identities will be given, which is a scenario not covered by the standard classification of single- and multi-shot cases. We propose a CRF model that is able to encode a "soft grouping" property of unlabelled images. Our application of CRFs to identity inference is similar in spirit to semi-supervised techniques based on the graph Laplacian like manifold ranking [26, 6]. These techniques, however, do not

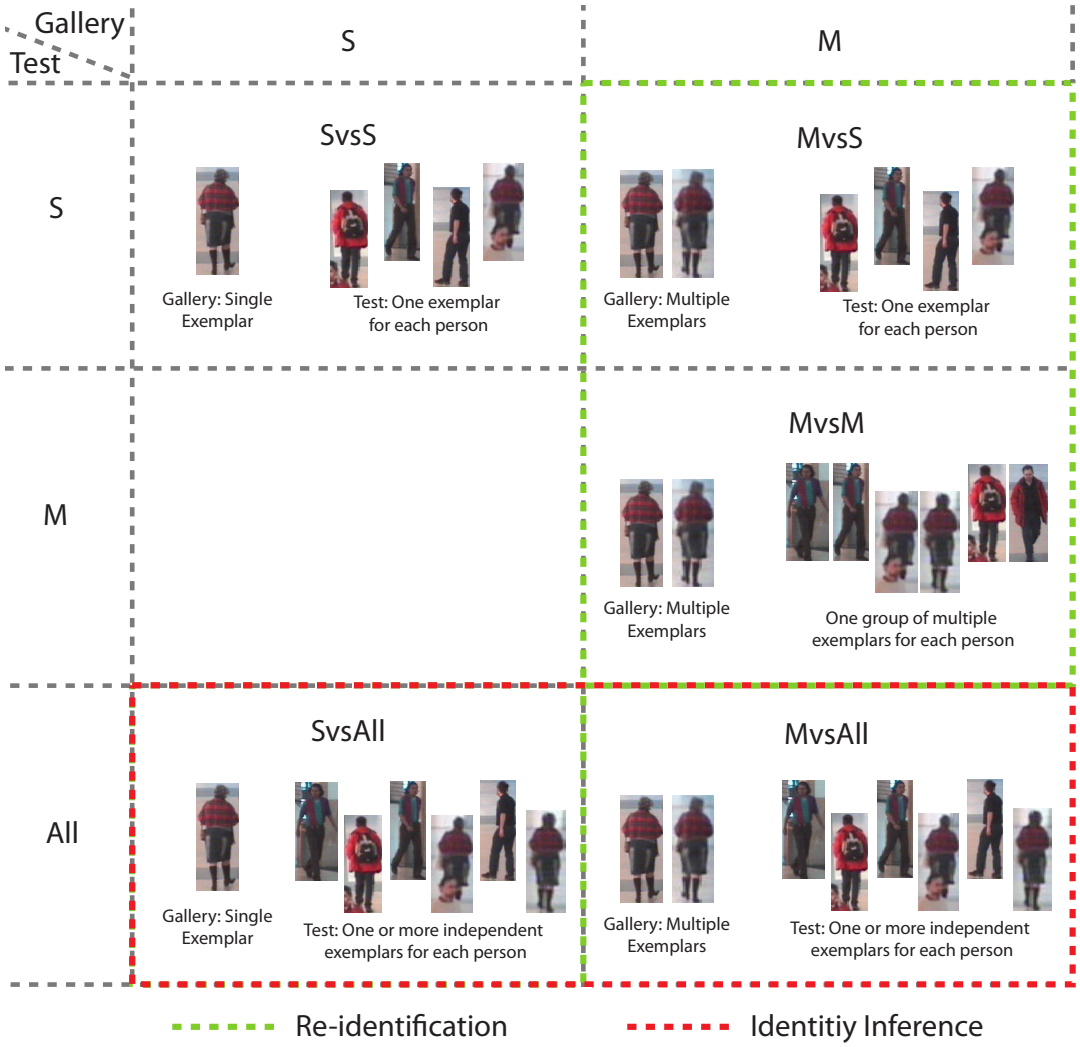


Fig. 1: Re-identification and identity-inference protocols.

immediately generalize to multi-shot modalities and it is unclear how to use them for batch re-identification of more than one probe image at a time.

3 Identity inference as generalization of re-identification

In this section we give a formal definition of the re-identification and identity inference problems. The literature on person re-identification considers several different configurations of gallery and test images. The modality of a specific re-identification problem depends on whether the gallery and/or test subsets contain single or multiple instances of each individual. Here we consider each modality in turn and show how each can be represented as an instance of our definition of re-identification. A summary of the different protocols is given in figure 1. Despite the importance of the problem, there is much confusion about how each of the classical re-identification modalities are defined. One of our goals in this chapter is to formally define how, given a set of images of people extracted from video sequences, each type of re-identification problem is determined.

Let $\mathscr{L} = \{1, \ldots N\}$ be a label set for a re-identification scenario, where each element represents a unique individual appearing in a video sequence or collection of sequences. Given a number of instances (images) of individuals from $\mathscr{L}$ detected in a video collection:

$$\mathscr{I} = \{x_i \mid i = 1 \ldots D\},$$

we assume that each image x_i of an individual is represented by a feature vector $\mathbf{x}_i \equiv \mathbf{x}(x_i)$ and that the label corresponding to instance x_i is given by $y_i \equiv y(x_i)$. Note that we interchangeably use the implicit notation y_i and $\mathbf{x}_i$ for the label and feature vector corresponding to image x_i, or the explicit functional notation $y(x_i)$ and $\mathbf{x}(x_i)$, as appropriate.

An instance of a re-identification problem, represented as a tuple $\mathscr{R} = (\mathscr{G}, \mathscr{T})$, is completely characterized by its gallery and test image sets ($\mathscr{G}$ and $\mathscr{T}$, respectively). Formally, the gallery images are defined as:

$$\mathscr{G} = \left\{\mathscr{G}_j \mid j = 1 \ldots N\right\}, \text{where } \mathscr{G}_j \subset \{x \mid y(x) = j\}.$$

That is, for each individual $i \in \mathscr{L}$, a subset of all available images is chosen to form his gallery $\mathscr{G}_i$. The set of test images is defined as:

$$\mathscr{T} = \left\{\mathscr{T}_j \mid j = 1 \ldots M\right\} \subset \mathscr{P}(\mathscr{I}),$$

where $\mathscr{P}$ is the powerset operator (i.e. $\mathscr{P}(I)$ is the set of all subsets of $\mathscr{I}$). We further require for all $\mathscr{T}_j \in \mathscr{T}$ that $x, x' \in \mathscr{T}_j \Rightarrow y(x) = y(x')$ (sets in $\mathscr{T}$ have homogeneous labels), and $\mathscr{T}_j \in \mathscr{T} \Rightarrow \mathscr{T}_j \cap \mathscr{G}_i = \emptyset, \forall i \in \{1 \ldots N\}$ (the test and gallery sets are disjoint). A solution to an instance of a re-identification problem is a mapping from the test images $\mathscr{T}$ to the set of all permutations of $\mathscr{L}$.

3.1 Re-identification scenarios

In this section we formally define each of the standard re-identification modalities commonly discussed in the literature. Though we define single test scenarios for each modality, in practice each scenario is repeated over a number of random trials to evaluate performance.

Single-versus-all re-identification (SvsAll) is often referred to as simply *single-shot re-identification* or *single-versus-single* (SvsS) but could better be described as *single-versus-all* (SvsAll)[1] re-identification (see figure 1). In the SvsAll re-identification scenario a single gallery images is given for each individual, and *all remaining instances* of each individual are used for testing: $M = D - N$. Formally, a single-versus-all re-identification problem is a tuple $\mathscr{R}_{\text{SvsAll}} = (\mathscr{G}, \mathscr{T})$, where:

[1] We prefer the SvsAll terminology as the SvsS terminology has been misinterpreted at least once in the literature.

$$\mathscr{G}_j = \{x\} \text{ for some } x \in \{x \mid y(x) = j\}, \text{ and}$$
$$\mathscr{T}_j = \left\{\{x\} \mid x \in \mathscr{I} \setminus \mathscr{G}_j \text{ and } y(x) = j\right\}.$$

In a single-versus-all instance of re-identification the gallery sets are all singletons, containing only a single example of each individual.

This re-identification modality was first described by Farenzena et al. [11] and Schwartz et al. [19]. Note that despite its simplicity, this configuration is susceptible to misinterpretation. At least one author has interpreted the SvsS modality to be one in which a single gallery image per subject is used, and a *single* randomly chosen probe image is also chosen for each subject [7]. SvsAll re-identification is a realistic model of scenarios where no reliable data association can be performed between observations before re-identification is performed. This could be the case, for example, when very low bitrate video is processed or in cases where imaging conditions do not allow reliable detection rates.

Multi-versus-single shot re-identification (MvsS) is defined using G gallery images of each person, while each of the test sets $\mathscr{T}_j$ contains only a single image. In this case $M = N$, as there are exactly as many singleton test sets $\mathscr{T}_j$ as persons depicted in the gallery. Formally, a MvsS re-identification problem is a tuple $\mathscr{R}_{\text{MvsS}} = (\mathscr{G}, \mathscr{T})$, where:

$$\mathscr{G}_j \subset \{x \mid y(x) = j\} \text{ and } |\mathscr{G}_j| = G \; \forall j \text{ and}$$
$$\mathscr{T}_j = \{x\} \text{ for some } x \notin \mathscr{G}_j \text{ s.t. } y(x) = j.$$

The MvsS configuration is not precisely a generalization of the SvsAll person re-identification problem in that, after selecting G gallery images for each individual, only a *single* test image is selected to form the test sets $\mathscr{T}_j$.

The MvsS re-identification scenario has been used in only a few works in the literature [11, 7]. We do not consider it to be an important modality, though it might be an appropriate model of verification scenarios where a fixed set of gallery individuals are enrolled and then must be unambiguously re-identified on the basis of a single image.

Multi-versus-multi shot re-identification (MvsM) is the case in which the gallery and test sets of each person both have G images. In this case $M = N$, there is again as many gallery sets as test sets. After selecting the G gallery images for each of the N individuals, only a fraction of the remaining images of each person are used to form the test set. Formally, a MvsM re-identification problem is a tuple $\mathscr{R}_{\text{MvsM}} = (\mathscr{G}, \mathscr{T})$, where:

$$\mathscr{G}_j \subset \{x \mid y(x) = j\} \text{ and } |\mathscr{G}_j| = G \; \forall j \text{ and}$$
$$\mathscr{T}_j \subset \left\{x \mid y(x) = j \text{ and } x \notin \mathscr{G}_j\right\} \text{ and } |\mathscr{T}_j| = G \; \forall j.$$

Note that the MvsM configuration is *not* a generalization of the SvsAll case in which all of the available imagery for each target is used as test imagery. The goal in MvsM re-identification is to re-identify each *group* of test images, leveraging the knowledge that images in each group are all of the same individual.

The MvsM re-identification modality is the most commonly reported one in the literature [11, 3, 8, 1]. It is representative of scenarios in which some amount of reliable data association can be performed before re-identification. However, it is not a completely realistic formulation since data association is never completely correct and there will always be uncertainty about group structure in probe observations.

3.2 Identity inference

Identity inference addresses the problem of having *few* labeled images while desiring to label *many* unknown images without explicit knowledge that groups of images represent the same individual. The formulation of the *single-versus-all* re-identification falls within the scope of identity inference, but neither the multi-versus-single nor the multi-versus-multi formulations are a generalization of this case to multiple gallery images. In the MvsS and MvsM cases the test set is either a singleton for each person (MvsS) or a group of images (MvsM) of the same size as the gallery image set for each person. Identity inference could be described as a *multi-versus-all* configuration. Formally, it is a tuple $\mathscr{R}_{\text{MvsAll}} = (\mathscr{G}, \mathscr{T})$, where:

$$\begin{aligned} &\mathscr{G}_j \subset \{x \mid y(x) = j\} \text{ and } |\mathscr{G}_j| = G \text{ and} \\ &\mathscr{T}_j = \left\{\{x\} \mid x \in \mathscr{I} \setminus \mathscr{G}_j \text{ and } y(x) = j\right\}. \end{aligned}$$

In instances of identity inference a set of G gallery images is chosen for each individual. All remaining images of each individual is then used as an element of the test set without any identity grouping information. As in the SvsAll case, the test images sets are all singletons.

Identity inference as a generalization of person re-identification was first introduced in [14]. It encompasses both SvsAll and MvsAll person re-identification and represents, in our opinion, one of the most realistic scenarios in practice. The re-identification modalities accurately model situations where an operator is interested in inferring the identity of past, unlabeled observations on the basis of very few labeled examples of each person. In practice, the number of labeled images available is significantly less than the number of images for which labels are desired.

4 A CRF model for identity inference

Conditional Random Fields (CRFs) have been used to model the statistical structure of problems such as semantic image segmentation [4] and stereo matching [18]. In this section we show how we model the identity inference problem as a minimum energy labelling problem in a CRF.

A CRF is defined in general case by a graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, a set of random variables $\mathscr{Y} = \left\{Y_j \mid j = 1 \ldots |V|\right\}$ which represent the statistical structure of the problem being modelled, and a set of possible labels $\mathscr{L}$. The vertices $\mathscr{V}$ index the random variables in

$\mathscr{V}$ and the edges $\mathscr{E}$ encode the statistical dependence relations between the random variables. The labelling problem is then to find an assignment $\tilde{\mathbf{y}}$ of labels to nodes that minimizes an energy function E over possible labellings $\mathbf{y}^* = (y_i^*)_{i=1}^{|V|}$: $\tilde{\mathbf{y}} = \arg\min_{\mathbf{y}^*} E(\mathbf{y}^*)$.

The energy function $E(\mathbf{y}^*)$ is defined as:

$$E(\mathbf{y}^*) = \sum_{i \in \mathscr{V}} \phi_i(y_i^*) + \lambda \sum_{(i,j) \in \mathscr{E}} \psi_{ij}(y_i^*, y_j^*), \tag{1}$$

where $\phi_i(y_i^*)$ is a unary data cost encoding the penalty of assigning label y_i^* to vertex i and $\psi_{ij}(y_i^*, y_j^*)$ is a binary smoothness cost representing the conditional penalty of assigning labels y_i^* and y_j^* respectively to vertices i and j. The parameter λ in equation (1) controls the trade-off between data and smoothness costs.

To minimize properly defined energy functions [15] and find an optimal labelling $\tilde{\mathbf{y}}$ in a CRF, the graph cut approach has been shown to be competitive [20] against other methods proposed in the literature such as Max-Product Loopy Belief Propagation [12] and Tree-Reweighted Message Passing [23]. The multi-label problem is solved by iterating the α-expansion move [5] where the binary labelling is expressed as each node either keeping its current label or taking the label α selected for the iteration. In all experiments we use the graph cut approach to minimizing the energy in equation (1). If higher ranks are desired, an inference algorithm like loopy belief propagation that returns the marginal distributions at each node can be used [12]. We feel that rank-1 precision is the most significant performance measure, and found loopy belief propagation to be many times slower than graph cuts on our inference problem.

CRF topology: We can map an identity inference problem $\mathscr{R} = (\mathscr{G}, \mathscr{T})$ onto a CRF by defining the vertex and edge sets $\mathscr{V}$ and $\mathscr{E}$ in terms of the gallery and test image sets defined by $\mathscr{G}$ and $\mathscr{T}$. We have found two configurations of vertices and edges to be useful for solving identity inference problems. The first uses vertices to represent groups of images in the test set $\mathscr{T}$ and is particularly useful for modelling MvsM re-identification problems:

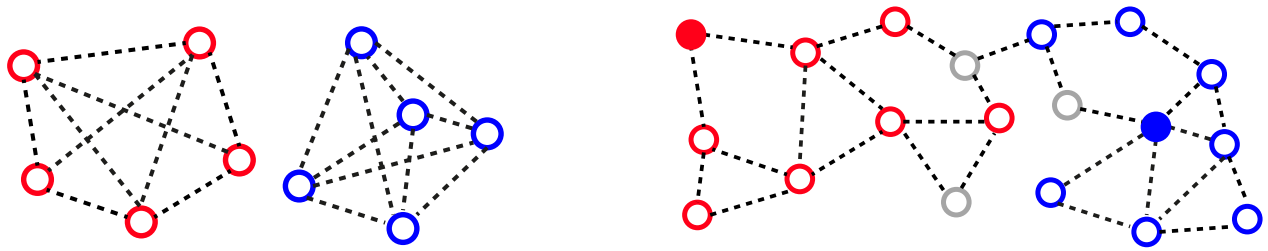
$$\mathscr{V} = \bigcup_{i=1}^{N} \mathscr{T}_i \text{ and } \mathscr{E} = \left\{ (x_i, x_j) \mid x_i, x_j \in \mathscr{T}_l \text{ for some } l \right\}.$$

The edge topology in this CRF is completely determined by the group structure as expressed by the $\mathscr{T}_j$.

When no identity grouping information is available for the test set, as in the general identity inference case as well as in SvsAll re-identification, we instead use the following formulation of the CRF:

$$\mathscr{V} = I \text{ and } \mathscr{E} = \bigcup_{x_i \in \mathscr{V}} \left\{ (x_i, x_j) \mid x_j \in \text{kNN}(x_i) \right\},$$

where the $\text{kNN}(x_i)$ maps an image to its k most similar images in feature space. Note that we hence treat equally training and test images when building $\mathscr{E}$. The topology of this CRF formulation, in the absence of explicit group information, uses feature simi-



(a) MvsM (probes only) (b) SvsAll

Fig. 2: Illustrations of the CRF topology for the MvsM (a) and SvsAll (b) modalities. Filled circles represent gallery images, unfilled circles probes. Color indicates the ground truth label.

larity to form connections between nodes. Illustrations of the topology for the MvsM and SvsAll scenarios are given in figure 2.

Data and smoothness costs: The unary data cost determines the penalty of assigning label y_i^* to vertex i given $\mathbf{x}(x_i)$, the observed feature representation of image x_i. We define it as:

$$\phi_i(y_i^*) = \min_{x \in \mathscr{G}_{y_i^*}} ||\mathbf{x}(x) - \mathbf{x}(x_i)||_2. \tag{2}$$

That is, the cost of assigning label y_i^* is proportional to the minimum L2-distance between the feature representation of image x_i and any gallery image of individual y_i^*. The data cost is L1-normalized for each vertex i, and hence is a cost distribution over the labels. The data cost can be seen as the individual assignment cost.

We use the smoothness cost $\psi_{ij}(y_i^*, y_j^*)$ to ensure local consistency between labels in neighbouring nodes, it is composed of a label cost $\psi(y_i^*, y_j^*)$ and a weighting factor w_{ij}:

$$\psi_{ij}(y_i^*, y_j^*) = w_{ij}\psi(y_i^*, y_j^*), \tag{3}$$

$$\psi(y_i^*, y_j^*) = \begin{cases} 0 & \text{if } y_i^* = y_j^* \\ \frac{1}{|\mathscr{G}_{y_i^*}||\mathscr{G}_{y_j^*}|} \sum\limits_{\substack{x \in \mathscr{G}_{y_i^*} \\ x' \in \mathscr{G}_{y_j^*}}} ||\mathbf{x}(x) - \mathbf{x}(x')||_2 & \text{otherwise.} \end{cases} \tag{4}$$

The label cost $\psi(y_i^*, y_j^*)$ depends only on the labels. The more similar two labels are in terms of the available gallery images for them, the lower the cost for them to coexist in a neighbourhood of the CRF. The label cost are L1-normalized, and thus is a cost distribution over all labels. Note that the label cost is fixed to 0 if $y_i^* = y_j^*$ (see equation (4)). The weighting factors w_{ij} allow the smoothness cost between nodes i and j to be flexibly controlled according to the problem at hand. In the experiments presented in this chapter, we define the weights w_{ij} from equation (3) between vertices i and j in the CRF in terms of feature similarity:

$$w_{ij} = exp(-||\mathbf{x}(x_i) - \mathbf{x}(x_j)||_2). \quad (5)$$

This definition gives a higher cost to a potential labelling $\mathbf{y}^*$ that labels similar images differently. But as the similarity between nodes decreases, so does the cost of keeping these two different labels. Hence, our method would still allow connected nodes to share different labels but will tend to discourage this situation especially for very similar images and/or very different identities.

5 Experimental results

In this section we describe a series of experiments we performed to evaluate the effectiveness of the approach described in section 4 for solving identity inference problems. In the next section we describe the three datasets we use in all experiments and the feature descriptor we use to represent appearance. In section 5.2 we report on the re-identification experiments performed on these three datasets, and in section 5.3 we report results for identity inference.

Note that in all experiments we only evaluate performance at rank-1 and not across all ranks (as is done in some re-identification works). We believe that rank-1 performance, that is *classification* performance, is the most important metric for person re-identification since it is indicative of how well fully-automatic re-identification performs. Consequently, all plots in this section are not CMC curves, but rather plots of rank-1 re-identification accuracies for various parameter settings.

5.1 Datasets and feature representation

We evaluate the performance of our CRF approach on a variety of commonly used datasets for re-identification. To describe the visual appearance of images in gallery and probe sets we use a simple descriptor that captures color and shape information.

5.1.1 Datasets

For evaluating identity inference performance we are particularly interested in test scenarios where there are many images of each test subject. However, most publicly available datasets for re-identification possess exactly the opposite property in that they contain very few images per person. The VIPER [13] dataset only provides a pair of images for each identity and thus no meaningful structure can be defined for our approach. Another popular dataset for re-identification is i-LIDS [25] which has on average 4 images per person. Although this is a rather small number of images per person we want to demonstrate the robustness of our approach also on this dataset. The most interesting publicly available datasets for our approach are CAVIAR [8], which contains between 10 and 20 images for each person extracted from two different view of an indoor envi-

Table 1: Re-identification dataset characteristics.

	ETHZ1	**ETHZ2**	**ETHZ3**	**CAVIAR**	**i-LIDS**
Number of cameras	1	1	1	2	3
Environment	Outdoor	Outdoor	Outdoor	Indoor	Indoor
Number of identities	83	35	28	72	119
Minimun number of images/person	7	6	5	10	2
Average number of images/person	58	56	62	17	4
Maximun number of images/person	226	206	356	20	8
Average detection size	60×132	63×135	66×148	34×81	68×175

ronment, and the ETHZ [19] dataset, which consists of three video sequences, where on average each person appears in more than 50 images. The characteristics of the selected datasets are summarized in table 1 and some details are given below:

- **ETHZ.** The ETHZ Zurich dataset [19] consists of detections of persons extracted from three sequences acquired outdoors. This dataset is divided in distinct datasets, corresponding to three different sequences in which different persons appear.

  1. The ETHZ1 sequence contains images of 83 persons. The number of detections per person ranges from 7 to 226, the average being 58. Detections have an average resolution of 60×132.
  2. ETHZ2 contains images of 35 persons. The number of detections per person ranges from 6 to 206, with an average of 56. This sequence seems to have been recorded on a bright, sunny day and the strong illumination tends to partially wash out differences in appearance, making this sequence one of the most difficult.
  3. ETHZ3 contains images of 28 persons, with the number of detections per person ranging between 5 and 356 (62 on average). The resolution of each detection is quite high (66×148), facilitating better description of each one. The small number of persons, the high resolution and the large number of images per person make this sequence the easiest of the three ETHZ datasets.

- **CAVIAR.** The CAVIAR dataset consists of several sequences recorded in a shopping center. It contains 72 persons imaged from two different views and was designed to maximize variability with respect to resolution changes, illumination conditions and pose changes. As detailed in table 1, the number of images per person is either 10 or 20 with an average of 17. While the number of persons, cameras and detections make this dataset interesting, the very small average resolution of 34×81 makes it difficult to extract discriminative features.
- **i-LIDS.** The i-LIDS dataset consists of multiple camera views from a busy airport arrival hall. It contains 119 people imaged in different lighting conditions and most of the time with baggage that in part occlude the person. The number of images per person is low, with a minimum of 2, a maximum of 8 and an average of 4. The average resolution of the detections (68×175) is rather high, especially with respect to the other datasets.

5.1.2 A descriptor for re-identification

In our experiments we use a descriptor based on both color and shape information that requires no foreground/background segmentation and does not rely on body-part localization. Given an input image of a target, it is resized to a canonical size of 64×128 pixels with coordinates between $[-1,1]$ and origin $(0,0)$ at the center of the image. Then we divide it into overlapping horizontal stripes of 16 pixels in height and from each stripe we extract an RGB histogram. The use of horizontal stripes allows us to capture the vertical color distribution in the image, while overlapping stripes allow us to maintain color correlation information between adjacent stripes in the final descriptor. We equalize all RGB color channels before extracting the histogram. Histograms are quantized to $4 \times 4 \times 4$ bins.

Descriptors of visual appearance for person recognition can be highly susceptible to background clutter, and many approaches to person re-identification use sophisticated background modeling techniques to separate foreground from background signals [11, 2, 3]. We use a more straightforward approach that weights the contribution of each pixel to its corresponding histogram bin according to an Epanechnikov kernel centered on the target image:

$$K(x,y) = \begin{cases} \frac{3}{4}(1-(\frac{x}{W})^2-(\frac{y}{H})^2) & \text{if} \quad |(\frac{x}{W})^2+(\frac{y}{H})^2| \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where W and H are, respectively, the width and height of the target image. This discards (or diminishes the influence of) background information and avoids the need to learn a background model for each scenario. To the weighted RGB histograms we concatenate a set of Histogram of Oriented Gradients (HOG) descriptors computed on a grid over the image as described in [9]. The HOG descriptor captures local structure and texture in the image that are not captured by the color histograms.

The use of the Hellinger kernel, that is a simple application of the square root to all descriptor bins, is well known in the image classification community [21] and helps control the influence of dimensions in the descriptor that tend to have disproportionately high values with respect to the others. In preliminary experiments we found this to improve robustness of Euclidean distances between descriptors and we therefore take the square root of all histogram bins (both RGB and HOG) to form our final descriptor.

5.2 Multi-shot re-identification results

To evaluate our approach in comparison with other state-of-the-art methods [11, 8, 3], we performed experiments on each dataset described above for the MvsM re-identification scenario. We evaluate performance for galleries varying in size: 2, 5 and 10 images per person for ETHZ; 2, 3 and 5 images per person for CAVIAR; and 2 and 3 images per person for i-LIDS.

Note that grouping information in the test set is explicitly encoded in the CRF. Edges only link test images that correspond to the same individual, and one test images is connected to all other test images of that individual. In these experiments we fix $\lambda = 1$ in the energy function of equation (1), and the weight on the edges is defined according to the features similarity as detailed in equation (5).

Results for MvsM person re-identification are presented in figure 3a, 3c and 3e for ETHZ, figure 4a for CAVIAR, and figure 5a for i-LIDS. The NN curve in these figures corresponds to labelling each test image with the nearest gallery image label without exploiting group knowledge, while the GroupNN approach exploits group knowledge by assigning each group of test images the label for which the average distance between test images of that group and gallery individuals of that label is minimal. We refer to our approach as "CRF" in all plots, and for each configuration we randomly select the gallery and test images and average performance over ten trials.

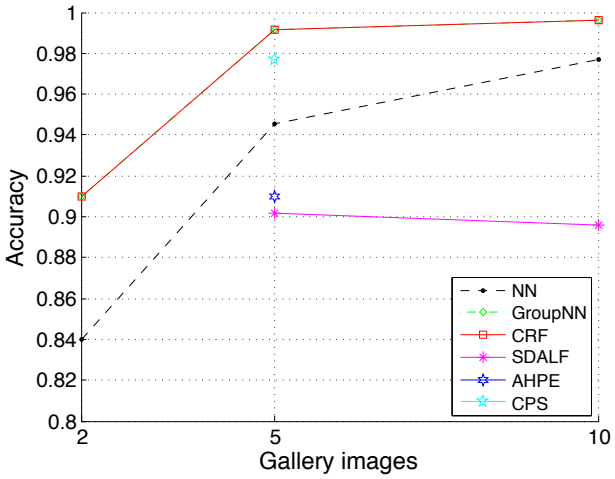
5.2.1 Multi-shot re-identification performance on ETHZ

For the MvsM scenarios on ETHZ we tested $M \in \{2,5,10\}$. We now detail results on each sequence and compare with the state-of-the-art when available.

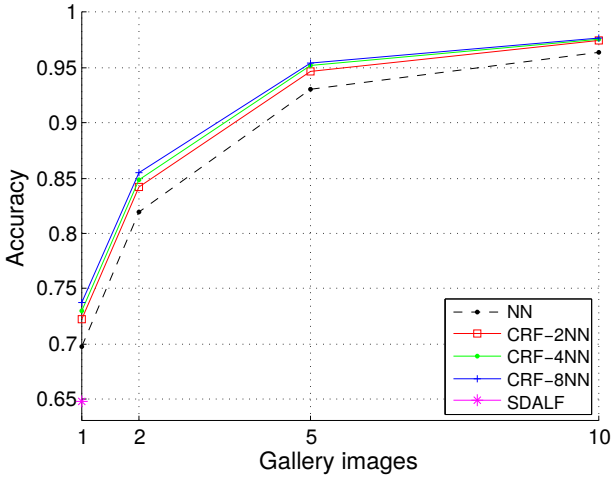
ETHZ1: Performance on ETHZ1 (figure 3a) starts at 84% accuracy at rank-1 for the simple NN classification approach and at 91% for both the GroupNN and our CRF approach for $M = 2$. Using 5 images, GroupNN and the CRF reach an accuracy of about 99.2%. The state-of-the-art on ETHZ1 for $M = 5$ is CPS at 97.7%. With 10 gallery and test images per subject, the CRF approach reachs 99.6% accuracy while the NN classification peaks at 97.7%. The SDALF approach obtains 89.6% on this scenario.

ETHZ2: On ETHZ2 (figure 3c), which is the most difficult of the ETHZ datasets, performance at $M = 2$ starts at 81.7% for the simple NN baseline and 90% for both the GroupNN and our CRF approach. Using 5 images, methods exploiting group knowledge reach 99.1%. The state-of-the-art on this dataset is SDALF at 91.6%, AHPE at 90.6%, and CPS which reaches 97.3%. Finally, when using 10 images for the gallery and test sets, methods using grouping knowledge stays at 99.1%. Note that, as with ETHZ1, SDALF performance at $M = 10$ is less than at $M = 5$ with 89.6% rank-1 accuracy.

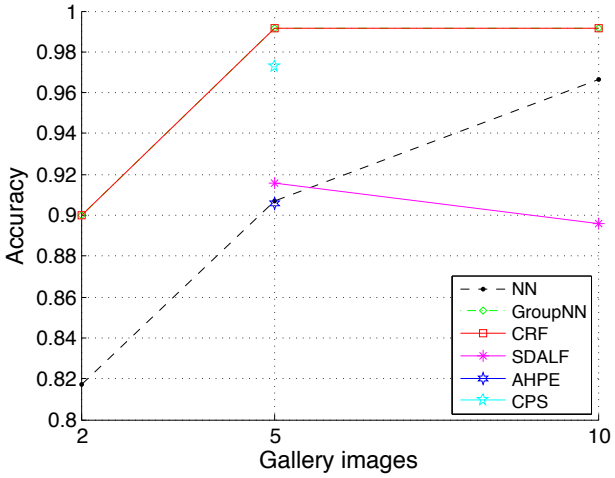
ETHZ3: On ETHZ3 (figure 3e), which is the "easiest" of the ETHZ datasets, performance start at 91.4% rank-1 accuracy for the simple NN baseline and at 96.8% for both the groupNN and our CRF approach with $M = 2$. The NN classification reaches 97% using 5 images and 99.3% using 10 images. Methods using group knowledge saturate the performance on this dataset for both 5 and 10 images. The SDALF approach obtains 93.7% and 89.6% accuracy using 5 and 10 images, respectively. For $M = 5$, the AHPE approach obtains 94% while the CPS method arrives at 98% rank-1 accuracy.



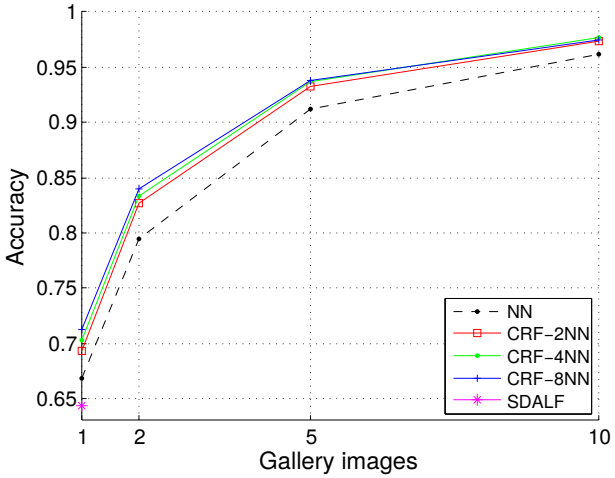
(a) ETHZ1 MvsM



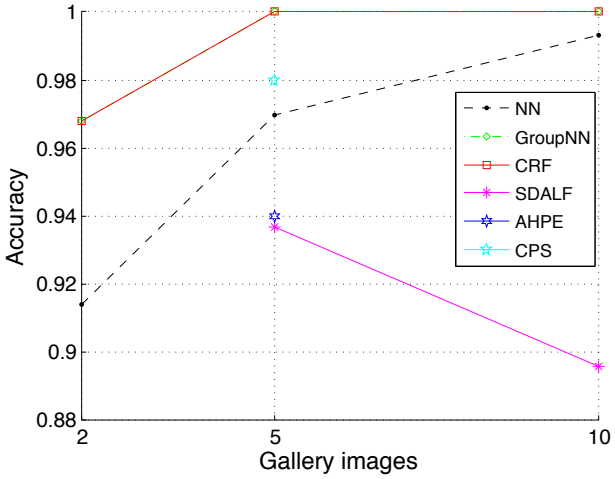
(b) ETHZ1 {M,S}vsAll



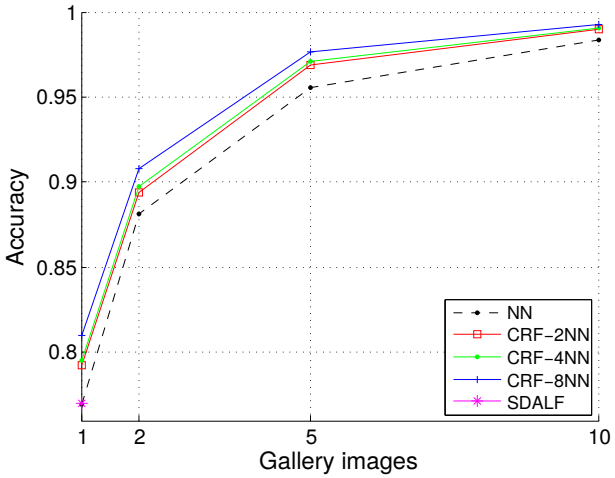
(c) ETHZ2 MvsM



(d) ETHZ2 {M,S}vsAll

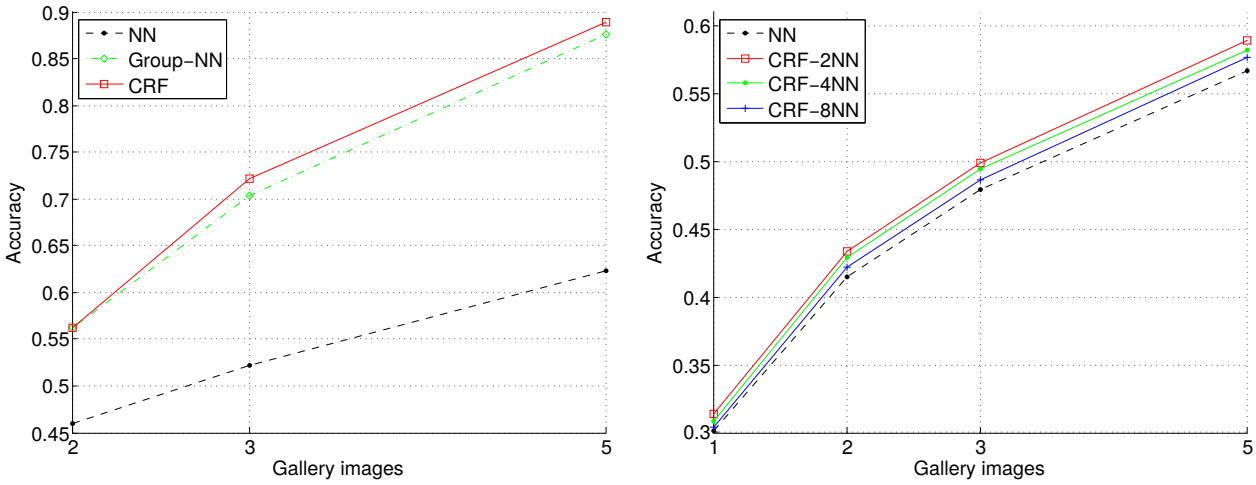


(e) ETHZ3 MvsM



(f) ETHZ3 {M,S}vsAll

Fig. 3: MvsM (left column) and SvsAll and MvsAll (right column) re-identification accuracy on ETHZ. Note that these are *not* CMC curves, but are rank-1 *classification* accuracies over varying gallery and test set sizes.



(a) CAVIAR MvsM (b) CAVIAR {M,S}vsAll

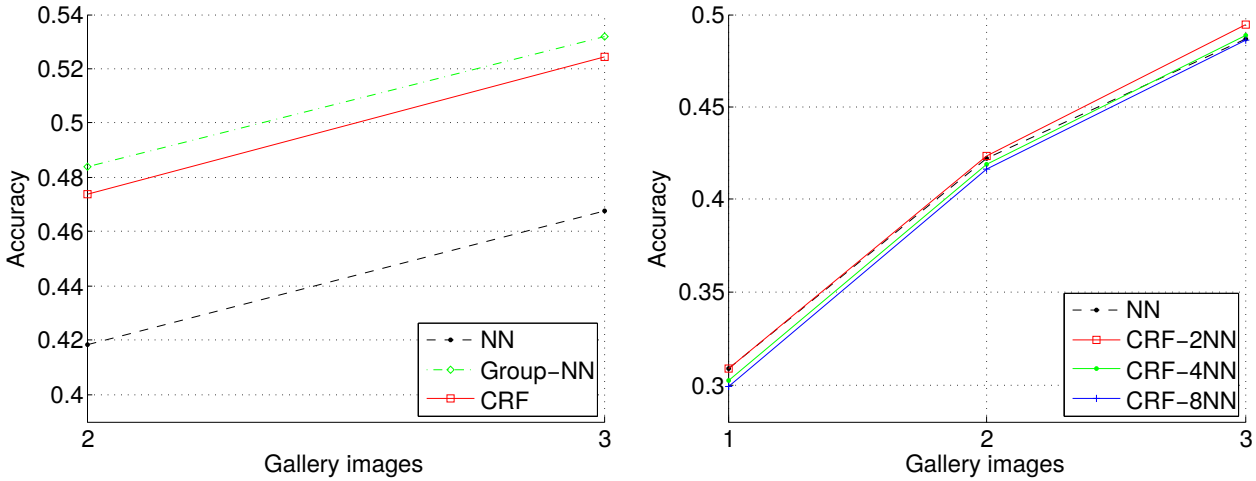
Fig. 4: MvsM (left) and SvsAll and MvsAll (right) re-identification accuracy on CAVIAR.

5.2.2 Multi-shot re-identification performance on CAVIAR

For MvsM re-identification on CAVIAR we performed experiments with $M \in \{2,3,5\}$ (see figure 4a). Performance begins at 46% accuracy for the NN classification and 55% for approaches that use group structure in the probe image set. Using 3 images, the NN baseline reaches an accuracy of 52%, the GroupNN approach reaches 70.6%, while the CRF reaches 72%. These results can be compared with SDALF performance at 8.5%, HPE at 7.5% and the CPS performance at 13% for $M = 3$. Finally, with $M = 5$ the difference between methods exploiting group structure and those that do not becomes even more prominent. Nearest neighbor achieves 62.7%, while the GroupNN and CRF approaches reach 86.9% and 88.4%, respectively. The best state-of-the-art result on CAVIAR for $M = 5$ is CPS with 17.5%.

5.2.3 Multi-shot re-identification performance on i-LIDS

For the MvsM modality on i-LIDS we only tested $M \in \{2,3\}$ due to the limited number of images per person. Using $M = 2$ images, the NN classification yields an accuracy of 41.8%, while GroupNN yields a performance of 48.4% and our CRF approach 47.5%. The state-of-the-art on this configuration is: 39% for SDALF, 32% for HPE and 44% for CPS. Using $M = 3$ images yields only a small improvement as for many identities in i-LIDS there are fewer than 6 images. In these cases the gallery and probe image sets are limited to 2 images. GroupNN outperforms our CRF approach on this dataset due to the limited number of images per person.



(a) i-LIDS MvsM (b) i-LIDS {M,S}vsAll

Fig. 5: MvsM (left) and SvsAll and MvsAll (right) re-identification accuracy on i-LIDS.

5.2.4 Summary of multi-shot re-identification results

From these experiments on multi-shot person re-identification it is evident that significant improvement is obtained by exploiting group structure in the probe image set. The simple GroupNN rule and our CRF approach yield similar performance on MvsM re-identification scenarios on the ETHZ datasets where many images of each person are available. In combination with the discriminative power of our descriptor, our approach outperforms the state-of-the-art on the three ETHZ datasets. On the other hand, performance gains on the i-LIDS dataset are limited by the low number of images available for each person in the dataset.

Group structure in the probe image set yields a large boost in multi-shot re-identification performance also on the CAVIAR dataset, with an improvement of almost 30% between the simple NN classification and our CRF formulation exploiting group knowledge (note also the large improvement with respect to state-of-the-art methods). This is likely due to the fact that our approach does not compute mean or aggregate representations of groups and that our descriptor does not fit complex background or part models to the resolution limited images in the CAVIAR dataset.

5.3 Identity inference results

To evaluate the performance of our approach in comparison with other state-of-the-art methods [11, 8, 3], we performed experiments on all datasets using the SvsAll and MvsAll modalities described above. For the general identity inference case, unlike MvsM person re-identification, we have no information about relationships between test images. In the CRF model proposed in section 4 for identity inference, the local neighbourhood structure is determined by the K nearest neighbours to each image in feature space. For all experiments we tested $K \in \{2,4,8\}$. We set $\lambda = \frac{|\mathcal{V}|}{|\mathcal{E}|}$ in equa-

Fig. 6: Identity inference results (SvsAll). First row: test image, second row: incorrect NN result, third row: correct result given by our CRF approach.

tion (1) for the SvsAll and the MvsAll scenarios. Since there may be up to four times more smoothness terms than unary data cost terms in equation (1), setting λ in this way prevents smoothness from dominating the energy function. In identity inference, gallery images are randomly selected and *all remaining images* define the test set. All reported results are averages over ten trials as before. Results for identity inference are presented in figures 3b, 3d and 3f for ETHZ, figure 4b for CAVIAR and figure 5b for i-LIDS. We will now analyze the results for each dataset and then draw general conclusions.

5.3.1 Identity inference results on ETHZ

For the MvsAll configuration on ETHZ we tested $M \in \{2,5,10\}$.

ETHZ1: On ETHZ1 (figure 3b) we can observe that on the SvsAll modality the NN baseline using our descriptor yields an accuracy of 69.7%, while SDALF obtains 64.8%. The CRF approach improves this performance to 72% using 2 neighbours and 73.7% using 8 neighbours. Using 2 gallery images per person increases CRF results to a rank-1 accuracy ranging from 84.2% to 85.6%. Adding more gallery images yields continued improvement, reaching 97.7% accuracy with 10 gallery images per person and our CRF approach with 8 neighbours. Performance of the CRF with different neighbourhood sizes seems to converge at this point.

ETHZ2: On ETHZ2 (figure 3d), which is the most challenging of the ETHZ datasets, we can see that performance is slightly lower. The NN classification on the SvsAll modality obtains an accuracy of 66.9% compared to the SDALF performance of 64.4%, while our approach yields 69.3% and 71.3% accuracy using, respectively, 2 and 8 neighbours. With 2 gallery images, the gap between the NN baseline (79.5%) and the CRF

(84%) slightly widens. Using 10 model images, the performances stabilizes at 97.7% and we observe the same convergence as on ETHZ1.

ETHZ3: Finally, on ETHZ3 (figure 3f) the NN baseline and SDALF obtain the same accuracy of 77%, while the performance of our CRF approach ranges from 79.2% to 81% depending on the neighbourhood size. The performance quickly saturates with a maximum accuracy of 97.7% using 5 training images and 99% with 10 images.

5.3.2 Identity inference on CAVIAR

Identity inference on the CAVIAR dataset is significantly more challenging than on ETHZ. We evaluate performance on the SvsAll modality and on MvsAll modalities for $M \in \{2,3,5\}$ (see figure 4b). With only one gallery images per person, both the nearest neighbor and CRF approaches yield a rank-1 accuracy of about 30%. This is significantly higher than the state-of-the-art of about 8% for SDALF, AHPE and CPS, which is likely due to the simplicity of our descriptor and its robustness to occlusion and illumination changes. For all MvsAll modalities we note a significant gain in accuracy when adding more gallery images per person, with performance peaking at about 59% for the $M = 10$ case. This demonstrates that our CRF approach is able to effectively exploit multiple gallery examples.

5.3.3 Identity inference on i-LIDS

Due to the relatively small average number of images per person in the i-LIDS dataset, we only test the MvsAll modality for $M \in \{2,3\}$. Results are summarized in figure 5b. For only one training example per gallery individual, our approach yields a rank-1 accuracy of about 31%, which is comparable to the state-of-the-art result of 28% reported for SDALF. Adding more examples, as with the CAVIAR dataset, consistently improves rank-1 performance.

5.3.4 Summary of identity inference results

Using the CRF framework proposed in section 4 clearly improves accuracy over the simple NN re-identification rule. With our approach it is possible to label a very large number of probe images using very few gallery images for each person. For example, on the ETHZ3 dataset we are able to correctly label 1553 out of 1706 test images *using only two model images per person*. The robustness of our method with respect to occlusions and illumination changes is shown in the qualitative results in figure 6. The CRF approach yields correct labels even in strongly occluded cases thanks to the neighbourhood edges connecting it to less occluded, yet similar, images. This property of our descriptor pays off particularly well for the resolution-limited CAVIAR dataset, for which we outperform the state-of-the-art already for the SvsAll case.

6 Discussion

In this chapter we introduced the identity inference problem which we propose as a generalization of the standard person re-identification scenarios described in the literature. Identity inference can be thought of as a generalization of the single-versus-all person re-identification modality, and at the same time as a relaxation of the multi-versus-multi shot case. Instances of identity inference problems do not require hard knowledge about relationships between test images (e.g. that they correspond to the same individual). We have also attempted to formalize the specification of person re-identification and identity inference modalities through the introduction of a set-theoretic notation for precise definition of scenarios. This notation is useful in that it establishes a common, unambiguous language for talking about person re-identification problems.

We also proposed a CRF-based approach to solving identity inference problems. Using feature space similarity to define the neighbourhood topology in the CRF, our approach is able to exploit the soft-grouping structure present in feature space rather than requiring explicit group information as in classical MvsM person re-identification. Our experimental results show that the CRF approach can efficiently solve standard re-identification tasks, achieving classification performance beyond the state-of-the-art rank-1 results in the literature. The CRF model can also be used to solve more general identity inference problems in which no hard grouping information and very many test images are present in the probe set.

It is our opinion that in practice it is almost always more common to have many more unlabeled images than labeled ones, and thus that the standard MvsM formulation is unrealistic for most application scenarios. Further exploration of identity inference requires datasets containing many images of many persons imaged from many cameras. Most standard datasets like CAVIAR and i-LIDS are very limited in this regard.

References

1. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Multiple-shot human re-identification by mean riemannian covariance grid. In: Proceedings of AVSS, pp. 179–184 (2011)
2. Bazzani, L., Cristani, M., Perina, A., Farenzena, M., Murino, V.: Multiple-shot person re-identification by hpe signature. In: 20th International Conference on Pattern Recognition (ICPR), pp. 1413–1416 (2010)
3. Bazzani, L., Cristani, M., Perina, A., Murino, V.: Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recognition Letters **33**(7), 898–903 (2012)
4. Boix, X., Gonfaus, J.M., van de Weijer, J., Bagdanov, A.D., Serrat, J., Gonzàlez, J.: Harmony potentials. International Journal of Computer Vision **96**(1), 83–102 (2012)
5. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(11), 1222–1239 (2001)
6. C. C. Loy, C.L., Gong, S.: Person re-identification by manifold ranking. In: Proceedings of IEEE International Conference on Image Processing (2013)
7. Cai, Y., Pietikäinen, M.: Person re-identification based on global color context. In: Proceedings of the Asian Conference on Computer Vision Workshops, pp. 205–215 (2011)
8. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: Procedings of the British Machine Vision Conference, vol. 2, p. 6 (2011)

9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893 (2005)
10. Dikmen, M., Akbas, E., Huang, T.S., Ahuja, N.: Pedestrian recognition with a learned metric. In: Proceedings of the Asian conference on Computer Vision, pp. 501–512 (2011)
11. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2360–2367 (2010)
12. Felzenszwalb, P., Huttenlocher, D.: Efficient belief propagation for early vision. International Journal of Computer Vision **70**(1), 41–54 (2006)
13. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of the European Conference on Computer Vision, pp. 262–275 (2008)
14. Karaman, S., Bagdanov, A.D.: Identity inference: generalizing person re-identification scenarios. In: Computer Vision–ECCV 2012. Workshops and Demonstrations, pp. 443–452. Springer Berlin Heidelberg (2012)
15. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(2), 147–159 (2004)
16. Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2012)
17. Prosser, B., Zheng, W.S., Gong, S., Xiang, T., Mary, Q.: Person re-identification by support vector ranking. In: Proceedings of the PBritish Machine Vision Conference, vol. 2, p. 6 (2010)
18. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision **47**(1), 7–42 (2002)
19. Schwartz, W.R., Davis, L.S.: Learning discriminative appearance-based models using partial least squares. In: Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI), pp. 322–329. IEEE (2009)
20. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C.: A comparative study of energy minimization methods for markov random fields with smoothness-based priors. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(6), 1068–1080 (2008)
21. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(3), 480–492 (2012)
22. W. Zheng, S.G., Xiang., T.: Transfer re-identification: From person to set-based verification. In: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2012)
23. Wainwright, M., Jaakkola, T., Willsky, A.: Map estimation via agreement on trees: message-passing and linear programming. IEEE Transactions on Information Theory **51**(11), 3697–3717 (2005)
24. Zheng, W., Gong, S., Xiang, T.: Re-identification by relative distance comparison. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**(99), 1 (2012)
25. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: Proceedings of the PBritish Machine Vision Conference (2009)
26. Zhou, D., Weston, J., Gretton, A., Bousquet, O., Schölkopf, B.: Ranking on data manifolds. In: Advances in neural information processing systems, vol. 16, pp. 169–176 (2003)