Leveraging local neighborhood topology for large scale person re-identification

Svebor Karaman[1,*], Giuseppe Lisanti[1], Andrew D. Bagdanov[2], Alberto Del Bimbo[1]

Abstract

In this paper we describe a semi-supervised approach to person re-identification that combines discriminative models of person identity with a Conditional Random Field (CRF) to exploit the local manifold approximation induced by the nearest neighbor graph in feature space. The linear discriminative models learned on few gallery images provides coarse separation of probe images into identities, while a graph topology defined by distances between all person images in feature space leverages local support for label propagation in the CRF. We evaluate our approach using multiple scenarios on several publicly available datasets, where the number of identities varies from 28 to 191 and the number of images ranges between 1003 and 36 171. We demonstrate that the discriminative model and the CRF are complementary and that the combination of both leads to significant improvement over state-of-the-art approaches. We further demonstrate how the performance of our approach improves with increasing test data and also with increasing amounts of additional unlabeled data.

Keywords: Re-Identification, Conditional Random Field, Semi-supervised, ETHZ, CAVIAR, 3DPeS, CMV100

1. Introduction

Person re-identification is the problem of identifying previously seen individuals on the basis of one or more images captured from one or more cameras. It is an important aspect of modern surveillance systems as it is a way of maintaining identity information about targets in multiple views over potentially long periods of time. Re-identification is normally formulated in terms of a set of *gallery images* for each individual of interest, and a set of *probe images* which should be re-identified by determining the corresponding gallery individual. The problem is difficult due to changes in illumination and pose, occlusions, similarity of appearance, and changes in camera view. All of these render difficult the search for discriminative feature representations that capture identity and are robust to changing imaging conditions.

Re-identification performance is traditionally evaluated as a retrieval problem. The given data at training time is a gallery set consisting of a single or several images of known individuals. At test time, for each probe image or group of probe images of an unknown person, the goal of re-identification is usually to return a ranked list of individuals from the gallery. In the re-identification literature there are three main scenarios:

- **Single-vs-Single (SvsS) re-identification:** in which *exactly one* example image of each person is provided in the gallery and *at least one* instance of each person is present in the probe set.
- **Multi-vs-Multi (MvsM) re-identification:** in which a *group* of M examples of each individual is given in the gallery and a *group* of M examples of each individual is given to be re-identified in the probe set.
- **Multi-vs-Single (MvsS) re-identification:** in which *multiple images* of each person are given as groups in the gallery image set, and *exactly one* example image of each person is given in the probe set. The MvsS re-identification modality is little used and at times misinterpreted in the literature. It is not a realistic reflection of real-world application scenarios and we do not consider it further in this work.

*Corresponding author. Tel.: +39 055 275 1390, Fax: +39 055 275 1396
Email addresses: svebor.karaman@unifi.it (Svebor Karaman), giuseppe.lisanti@unifi.it (Giuseppe Lisanti), bagdanov@cvc.uab.es (Andrew D. Bagdanov), alberto.delbimbo@unifi.it (Alberto Del Bimbo)
[1]Media Integration and Communication Center (MICC), University of Florence, Viale Morgagni 65, Firenze 50134, Italy.
[2]Computer Vision Center, Barcelona, Universitad Autonoma de Barcelona, Bellaterra, Spain.

Preprint accepted for publication in Pattern Recognition

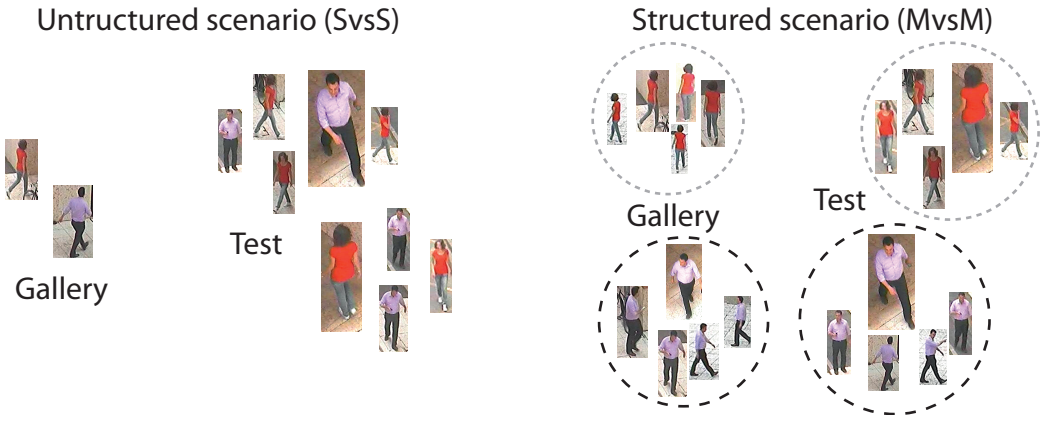


Figure 1: Unstructured and structured re-identification problems. In unstructured scenarios (left) the units of re-identification are individual images, while in structured scenarios (right) re-identification is performed on groups of images of the same person.

This breakdown of re-identification scenarios as used in the literature focuses primarily on the number of examples of each person given in the gallery and probe image sets. We feel that is more useful, instead, to think of re-identification problems as either "structured" or "unstructured". Structured re-identification problems are ones in which knowledge that multiple images depict the same person is given *a priori* in the problem definition. A typical example of *structured re-identification* is the multi-versus-multi (MvsM) scenario in which perfect, hard group structure is known about both the gallery and the test image sets[3]. We refer to the grouping structure in MvsM re-identification as *hard* because the groups of images corresponding to the same person are perfectly known. Note that perfect group knowledge is almost never known in practice, and knowledge of group structure significantly simplifies re-identification problems. Group knowledge in probe image sets can be used to apply voting schemes or similar methods to obtain more robust re-identification than assignment of individual probe image to nearest neighbors, for example.

In contrast, "unstructured" scenarios may contain multiple observations of a person without explicitly giving the corresponding group structure. An example of *unstructured re-identification* is the single-versus-single (SvsS) shot scenario in which there may be multiple images of the same person in the probe image set, but this grouping structure is not known. We sometimes refer to this type of re-identification problem as single-versus-all (SvsAll) to emphasize the fact that the task is to identify *all* individual probe images [1]. Note that unstructured problems are much more difficult than structured ones. Figure 1 illustrates the difference between structured and unstructured re-identification. In the structured case, the unit of re-identification is a *group* of images of the same person, while in the unstructured case the unit of re-identification is each *individual image*.

One of the greatest benefits in structured re-identification is that having multiple aspects of the same person in both the probe and gallery greatly increases the likelihood of finding at least one good match. Our objective with this work is to bring some of the benefits of structured re-identification problems to unstructured ones. To this end, we propose to infer a sort of soft group structure that can then be used to improve re-identification accuracy. We use the nearest-neighbor topology of all available gallery and probe imagery in feature space to accomplish this. This topology, which is essentially an estimation of the local data manifold, is used to adapt simple discriminative models of person identity to better reflect the structure of the data in feature space. A unique advantage of our approach is that we are able to exploit not only gallery imagery, but all available imagery when performing re-identification.

Re-identification has largely been limited to relatively small benchmarks that contain a few hundred or a few thousand images. In this work we also demonstrate how our semi-supervised approach scales well to very large datasets containing tens of thousands of images to label, and we are the first to perform re-identification experiments on such large datasets. Furthermore, in contrast to most discriminative techniques, the performance of our approach

[3]Note that throughout the paper we use *gallery images, training examples, training data* and *training samples* interchangeably to refer to images from the gallery. Similarly, we use *probe images, test images, test samples* and *test data* interchangeably to refer to images from the probe set.

improves with increasing amounts of test data. Even the addition of unlabeled, anonymous images for which we do not desire labels helps improve the performance of our approach due to better manifold sampling in the nearest-neighbor topology.

In the next section we review relevant work from the literature on person re-identification. We describe our approach to combining structure discovery with discriminative models in section 3. In section 4 we report on a series of experiments we performed to explore the performance of our algorithm and compare our approach to the state-of-the-art. Finally, we conclude in section 5 with a discussion of our contribution.

2. Related work

In this section we review the major trends in re-identification research, which can be broadly categorized into approaches that focus mostly on sophisticated appearance models for re-identification and approaches that focus more on learning discriminative classifiers or metrics for re-identification. We give an overview of our proposed approach with respect to previous work in section 2.3.

2.1. Appearance modeling for re-identification

The majority of existing research on person re-identification has concentrated on the development of sophisticated features for describing the visual appearance of targets. In [2] were introduced discriminative appearance-based models using Partial Least Squares (PLS) over texture, gradients and color features. The authors of [3] use an ensemble of local features learned using a boosting procedure, while in [4] the authors use a covariance matrix of features computed in a grid of overlapping cells. The SDALF descriptor introduced in [5] 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 [6] 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 [8] represents a person by a global mean color histogram and recurrent local patterns through epitomic analysis. More recently, the authors of [9] suggested that re-identification can be performed by exploiting small salient regions in each person image. They applied adjacency-constrained patch matching to build dense correspondence between image pairs through an unsupervised saliency learning method that does not require identity labels during training.

Two limitations of appearance-based approaches are that they often attempt to fit complex appearance models to target images of limited quality, and that they typically use aggregate or mean appearance models over multiple observations of the same individual (for multi-shot modalities). Both of these can be serious limitations in practice, since much surveillance imagery is resolution-limited and mean appearance models may not exploit well all available imagery.

2.2. Learning for re-identification

Differently than the approaches mentioned above, learning-based ones concentrate specifically on the classification or ranking technique applied to re-identify probe images. Techniques based on learning can be roughly grouped into metric learning approaches and those that learn strong discriminative models for classification or ranking. The approach in [10] is a supervised technique that uses pairs of similar and dissimilar images and a relaxed RankSVM algorithm to rank probe images. A set-based discriminative ranking approach was also recently proposed which alternates between optimizing a set-to-set geometric distance and a feature space projection, resulting in a discriminative set-distance-based model [11]. The Probabilistic Relative Distance Comparison approach learns a metric under which the probability of an incorrect match having a small distance is less than that of a correct one [12]. Camera transfer approaches have also been proposed that use images of the same person captured from different cameras to learn metrics [13, 14].

The authors of [15] propose a method for person appearance matching across disjoint camera views by learning a model that selects the most descriptive features for a specific class of objects. In particular, learning is performed in a covariance metric space using an entropy-driven criterion. Recently, a new metric learning method was proposed

in [16], where the authors use a feature extraction process based on both HSV and YUV color spaces, followed by unsupervised Principal Component Analysis (PCA) and supervised Local Fisher Discriminant Analysis (LFDA) for dimensionality reduction. Attribute learning has also been applied to re-identification. The authors of [17] define an ontology of basic concepts that can be related to visual characteristics of person images and then they learn (on half of the VIPeR dataset) an attribute-centric, part-based feature representation.

A semi-supervised, manifold ranking approach was recently applied to single-shot re-identification in [18]. However, this work is limited as only configurations with a single gallery image per person and a single test image are used. Its extension to multi-shot modalities and re-identification of multiple probes in "batches" is non-trivial due to the need to include multiple test images in the neighborhood graph and to maintain and propagate multiple rankings corresponding to each probe image. Another semi-supervised approach was recently proposed in [19], in which the authors fuse different features through a multi-feature learning (MFL) strategy that requires at least a single image per person as training data.

Learning-based approaches can achieve higher re-identification accuracy, especially at high ranks, but usually at the cost of setting aside a portion of available labeled data for learning discriminants or metrics. In real situations this implies labeling data for learning (at least a pair of images for each identity to estimate intra- and inter-class variation), and in practice such labeled imagery on which to learn is scarce.

2.3. Our contribution with respect to previous work

The semi-supervised approach we propose in this paper is similar in spirit to Blum et al. [20], and is an extension of our previous works on semi-supervised identity inference [1, 21]. The approach described in [20] uses graph mincuts to exploit labeled and unlabeled examples for binary classification problems. Their approach is applied only to binary labeling problems, however, and uses unary and binary cost functions based on distances in feature space. Though applied to multi-class problems, our previous work in [1] was based on similarities between simple bag-of-words appearance models and was evaluated on a single dataset. This work was extended in [21] by introducing an appearance model based on RGB and HOG histograms (instead of the bag-of-words) and by extending the evaluation to new datasets.

The approach we propose in this paper, in contrast, uses learned discriminative models for unary cost functions, and can be applied to multi-class labeling problems like person re-identification. It extends our previous approaches [1, 21] by explicitly leveraging labeled gallery images by learning discriminative models of each person. Like the identity inference approach, it exploits the structure in feature space, however in this work we show how all model parameters can be directly estimated from all available imagery. Finally, we demonstrate the effectiveness of our approach to leveraging local neighbor topology on a variety of re-identification datasets ranging from a few images per person to thousands of images per person.

Most semi-supervised approaches exploit unlabeled observations to augment the available training data and improve the resulting discriminative models. In contrast, our approach is designed to effectively exploit all available instances of targets in the gallery and probe image sets. As stated previously, learning-based approaches can achieve higher re-identification accuracy but at the cost of setting aside a portion of available labeled data for learning. Our approach learns weak, discriminative models from gallery imagery, and then uses an inference procedure in a CRF at *test time* to exploit all available data instead of learning on data set aside specifically for this purpose.

3. Discriminative model adaptation with CRFs

We begin in the next section with an overview of our proposed approach. In section 3.2 we formally define the re-identification problem. Section 3.3 details how we build weak discriminative models, and section 3.4 how the CRF topology is defined on top of all available imagery. Finally, in section 3.5 we show how all model parameters in our approach are estimated without supervision directly from the data given in a re-identification problem.

3.1. Overview of our approach

Our approach explicitly leverages labeled gallery images by learning discriminative models of each person, and then exploits the structure in feature space induced by the nearest neighbor topology of all available images. Figure 2 gives a visual overview of our semi-supervised approach:

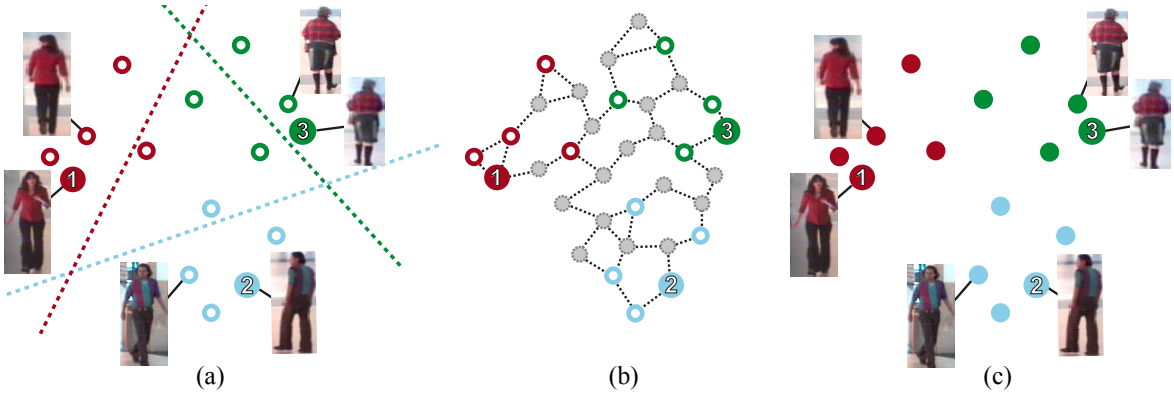


Figure 2: Illustration of our approach. (a) We fit a discriminative linear model to each gallery individual. (b) The k-nearest neighbors in feature space induce a local topology over all images of the re-identification problem: gallery (filled numbered circles), probe (unumbered colored circles) and added unlabeled data (grey filled circles) if available. (c) Inference in a CRF defined in terms of the discriminative models and the local topology results in robust re-identification of many test images. Best viewed in color.

- We first learn a set of weak linear discriminant models (section 3.3) of person identity. These models are learned on the labeled gallery data and, though they do not generalize well, give a coarse estimate of potential identities. In figure 2a, the learned margins separate identities roughly (but not perfectly, as evidenced by those probe images lying on the wrong side of their respective margins).
- At test time, we rely on a nearest-neighbor graph to estimate the local manifold structure around all observations (whether they be labeled, unlabeled test, or unlabeled "anonymous" images for which no label is desired). The topology is built from local similarities in feature space and is illustrated in figure 2b. The nearest-neighbor topology provides an *implicit structure* to an otherwise unstructured re-identification problem.
- Inference in the CRF (section 3.4), relying on the local similarities constraints from the topology, then adapts the outputs of the discriminative models. In figure 2c, probe samples that were lying on the wrong side of the margin of the weak classifiers are correctly classified after inference thanks to local similarity constraints.

Our approach brings some of the advantages of structured problems to unstructured ones. The hard grouping structure of MvsM re-identification problems provides very strong cues that are not given in the unstructured case. Our approach is thus to discover this structure in unstructured problems and leverage it in a principled way.

3.2. The re-identification problem

A re-identification problem is defined by a tuple $(\mathcal{L}, \mathcal{G}, \mathcal{P})$, where $\mathcal{L}$ is a set of N identities $\mathcal{L} = \{1, 2, \ldots, N\}$, $\mathcal{G}$ is a set of gallery images $\mathcal{G} = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_n\}$, and $\mathcal{P}$ is a set of probe images $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m\}$[4]. To simplify notation we use integers to represent identities in $\mathcal{L}$ and subsequently as labels in our CRF formulation below. It will also be useful to define the set $\mathcal{I}$ of all images (both gallery and probe) in a re-identification problem:

$$\begin{aligned} \mathcal{I} &= \mathcal{G} \cup \mathcal{P} \\ &\equiv \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n+m}\}. \end{aligned} \tag{1}$$

Thus, when we wish to refer specifically to an arbitrary gallery image we will use $\mathbf{g}_i$, while $\mathbf{p}_i$ will refer instead to an arbitrary probe image. When we use $\mathbf{x}_i$, however, we refer to an arbitrary image, either gallery or probe, in a re-identification problem.

We further define a function $l : \mathcal{I} \to \mathcal{L}$ that maps images to their corresponding identities. That is, for each image $\mathbf{x}_i \in \mathcal{I}$ its corresponding identity is the label $l(\mathbf{x}_i) \in \mathcal{L}$. For convenience, we also define the set of all gallery images corresponding to person i: $\mathcal{G}_i = \{\mathbf{g} \in \mathcal{G} : l(\mathbf{g}) = i\}$. The labels for all gallery $\mathbf{g}_i$ images are given in the definition of a re-identification problem, while the task of re-identification is to estimate the correct label for probe images $\mathbf{p}_i$.

[4] In practice each $\mathbf{g}_i$ and $\mathbf{p}_j$ are vectors in some feature space rather than images. Nonetheless, we will use the term "image" to describe gallery and probe sets.

3.3. Linear SVMs for re-identification

We begin by constructing a linear discriminant model of person identity using the gallery images $\mathcal{G}$ and their labels in the re-identification problem. With each identity $i \in \mathcal{L}$ in a re-identification problem we associate a set of weights $\mathbf{w}_i$ and a scalar bias b_i. A linear discriminant model is estimated for each person by solving a standard linear SVM optimization problem over the gallery images:

$$(\mathbf{w}_i, \cdot, b_i) = \arg\min_{\mathbf{w}_i, \boldsymbol{\xi}, b_i} \frac{1}{2}||\mathbf{w}_i||^2 + C\sum_{j=1}^{n} \xi_j \tag{2}$$

subject to

$$\begin{aligned} \delta^i_{l(\mathbf{g}_j)}(\mathbf{w}_i^T \mathbf{g}_j + b_i) &\geq 1 - \xi_j, \ \forall\, i \in \{1, \ldots, N\} \ \forall\, j \in \{1, \ldots, n\}, \\ \text{and } \xi_j &\geq 0, \ \forall\, j \in \{1, \ldots, n\}, \end{aligned} \tag{3}$$

where δ^i_j is a modified Kronecker delta function:

$$\delta^i_j = \begin{cases} 1 & \text{if } i = j \\ -1 & \text{otherwise.} \end{cases} \tag{4}$$

The discriminative models $(\mathbf{w}_i, b_i)$ are learned only on gallery images, and the construction of these models is the only supervised component of our approach. We use linear SVMs due to their simplicity and efficiency in evaluating multiple discriminants using a single matrix multiplication. In principle more complex, nonlinear SVMs could be used instead. However, given that re-identification problems typically have very few examples for each person in the gallery, linear discriminative models are less likely to overfit to the limited training data than more complex, kernel-based SVMs.

In some problems there may be only a single image for each person in the gallery, and in this case our formulation is equivalent to the exemplar SVM [22]. A critical step in the original exemplar SVM formulation is the calibration of SVM outputs using a set of judiciously selected positive and negative samples from a validation set. In re-identification, however, we do not have access to an abundance of observations suitable for calibration and must find an alternative method for improving the generalization of our linear models on unseen data. In the next section we use a conditional random field on top of SVM outputs to adapt labeling decisions according to the local topology of feature space.

3.4. Structuring feature space with CRFs

Real-world re-identification scenarios are often characterized by few gallery images and an abundance of unlabeled test images to be re-identified. Here we present an unsupervised approach to inducing a local topology over all images in a re-identification problem. When combined with the discriminative models of the previous section, the result is a semi-supervised approach to person re-identification.

Given an instance $(\mathcal{L}, \mathcal{G}, \mathcal{P})$ of a re-identification problem, we construct a CRF whose topology models the structure of all observations (both labeled gallery and probe images of unknown identity) in feature space. Our approach is independent of the feature descriptor used, and could be for example any of the appearance features described in section 2.1. We prefer to use a feature descriptor less sensitive to the quality of imagery, and describe the one we use for all experiments explicitly in section 4.1.

A CRF is defined by a graph $G = (\mathcal{V}, \mathcal{E})$ and a label set $\mathcal{L}$ which, for the purposes of our construction, is equivalent to the set of identities $\mathcal{L}$ in the re-identification problem modeled by the CRF. We create one vertex in $\mathcal{V}$ to represent each image in the re-identification scenario (all gallery and probe images):

$$\mathcal{V} = \{v_1, v_2, \ldots, v_{n+m}\}. \tag{5}$$

Node v_i in the vertex set $\mathcal{V}$ corresponds to image $\mathbf{x}_i \in \mathcal{I}$ in the re-identification problem.

The graph topology is then defined using the group structure of probe images, if given as in structured re-identification problems. Otherwise, we use the structure induced by the nearest neighbor topology in feature space

using all images of the re-identification problem. Explicitly, we create edges between images v_i and v_j if one is in the k-nearest neighbors of the other:

$$\mathcal{E} = \{(v_i, v_j) : \mathbf{x}_i \in \text{kNN}(\mathbf{x}_j) \vee \mathbf{x}_j \in \text{kNN}(\mathbf{x}_i)\}, \tag{6}$$

where $\text{kNN}(\mathbf{x})$ is the set of the k nearest neighbors to $\mathbf{x}$ in $\mathcal{I}$. We use Euclidean distances to define neighbors in feature space.

Given a hypothetical labeling $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_{|V|})$ assigning a label $\hat{y}_i \in \mathcal{L}$ to each vertex $v_i \in \mathcal{V}$ in graph G, we define the energy function of $\hat{\mathbf{y}}$ as:

$$E(\hat{\mathbf{y}}) = \sum_{i \in \mathcal{V}} \phi_i(\hat{y}_i) + \lambda \sum_{(v_i, v_j) \in \mathcal{E}} \psi_{ij}(\hat{y}_i, \hat{y}_j), \tag{7}$$

where $\phi_i(\hat{y}_i)$ is a unary data cost expressing how compatible the label $\hat{y}_i$ is with the observed data $\mathbf{x}_i$ at vertex v_i, and ψ_{ij} is a binary smoothness cost expressing how compatible the labeling of $\hat{y}_i$ and $\hat{y}_j$ are at nodes v_i and v_j, respectively. The parameter λ controls the tradeoff between data and smoothness costs. Intuitively, the unary cost defines a distribution over labels (identities) at each node, while the smoothness penalty should encourage that nodes corresponding to similar observations take similar labels.

The unary data costs are defined in terms of the linear models learned for each identity in the gallery. We define the unary penalty ϕ_i at vertex v_i in the CRF as:

$$\phi_i(\hat{y}_i) = e^{-(\mathbf{w}_{\hat{y}_i} \mathbf{x}_i + b_{\hat{y}_i})}. \tag{8}$$

The unary data cost for the identity $\hat{y}_i$ at vertex v_i is equal to the negative exponential of the output of the corresponding linear model $(\mathbf{w}_i, b_i)$ learned using equation 2. For gallery images, $\phi_i(l(i))$ is set to 0.

The unary penalty ϕ_i can also be defined directly using distances between gallery features of $\hat{y}_i$ and the feature $\mathbf{x}_i$ associated to vertex v_i:

$$\phi_i(\hat{y}_i) = \min_{\mathbf{g} \in \mathcal{G}_{\hat{y}_i}} ||\mathbf{g} - \mathbf{x}_i||_2. \tag{9}$$

The smoothness cost we use is defined in terms of Euclidean distances in feature space:

$$\psi_{ij}(\hat{y}_i, \hat{y}_j) = \psi(\hat{y}_i, \hat{y}_j) e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||_2}{\sigma^2}}, \tag{10}$$

where σ^2 is the variance of the distances between all connected features in the graph and $\psi(\hat{y}_i, \hat{y}_j)$ is the average distance between gallery images of identity $\hat{y}_i$ and $\hat{y}_j$:

$$\psi(\hat{y}_i, \hat{y}_j) = \frac{1}{|\mathcal{G}_{\hat{y}_i}||\mathcal{G}_{\hat{y}_j}|} \sum_{\mathbf{g} \in \mathcal{G}_{\hat{y}_i}} \sum_{\mathbf{g}' \in \mathcal{G}_{\hat{y}_j}} ||\mathbf{g} - \mathbf{g}'||_2. \tag{11}$$

The exponential factor in equation (10) tends to enforce labelling consistency by increasing the cost of giving different labels to similar images.

In section 4 we report on experiments using either the unary cost defined over SVM outputs (8), referred to as **SVM+CRF**, or the one defined directly in terms of distances in feature space (9), referred to as **FEAT+CRF**. In both cases the unary penalties are L1-normalized so that each ϕ_i is a probability distribution over the set of labels $\mathcal{L}$. Label inference is performed using α-expansion moves in the graph-cut formulation of energy-minimization [23]. As usual when dealing with multi-label problems, the α-expansion is applied successively to each label until convergence, iteratively solving N binary labelling problems.

3.5. Parameter estimation

There are several parameters to estimate before solving a re-identification problem with our approach: the SVM weights and biases $(\mathbf{w}_i, b_i)$ for each identity i, the cost C in equation (2) that penalizes misclassifications in the gallery set, the λ parameter in equation (7) balancing the tradeoff between smoothness and data penalties in the CRF energy,

and the number of neighbors k to use when constructing the edge topology of the CRF. Here we describe how we estimate all parameters directly from gallery image sets and the total number of images. Note that this removes the need for cross-validation since all parameters are estimated from gallery data and information extracted from probe images at test time (the total number of images).

Linear SVM parameters We use `libsvm`[5] to estimate the linear SVM models $(\mathbf{w}_i, b_i)$ for each person in the gallery. Since the energy function in equation (7) is a combination of distances in feature space (the smoothness penalty ψ_{ij}) and linear SVM outputs (the data cost ϕ_i), it is important to scale SVM outputs so that the CRF can correct errors when enough local evidence is present. Our intuition is that we want the positive examples in the gallery to have, on average, positive outputs from their corresponding linear model. At the same time we desire negative outputs, on average, from the other linear models on these same images. The C value in the SVM objective in equation (2) allows us to do just this by selecting the least C yielding linear SVM ensembles $(\mathbf{w}_i, b_i)$ such that:

$$\frac{\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{L}}\sum_{\mathbf{g}\in\mathcal{G}_i}\mathbf{w}_i^T\mathbf{g}+b_i}{\frac{1}{|L|}\sum_{i\in\mathcal{L}}\frac{1}{|\mathcal{G}|-|\mathcal{G}_i|}\sum_{\mathbf{g}\in\mathcal{G}\setminus\mathcal{G}_i}\mathbf{w}_i^T\mathbf{g}+b_i} < 0 \tag{12}$$

This ratio is a sort of average "contrast" between positive and negative SVM outputs on gallery images. Using this ratio of averages helps to have linear models that generalize well and avoid overfitting to the limited training examples in the gallery. A single C value common to all models is estimated for each trial through a coarse-to-fine approach. The optimal C intuitively depends on the number of persons, the number of gallery images, and the quality of images. In practice a single C could be estimated once for a given scenario as intra-dataset C-variance is very small. Label inference in the CRF defined on top of these outputs can correct errors of the SVMs on test examples thanks to the induced smoother decisions.

Number of neighbors and λ The number of neighbors used to create the topology of each CRF must be selected wisely to be robust to potential imbalance (different number of images per person) and to yield meaningful local connectivity. We desire a topology that provides local support for label consistency. Though the expected number of images per label will be approximately $\frac{|\mathcal{V}|}{|\mathcal{L}|}$, we do not build the topology setting k to this value since it implies a perfectly balanced dataset, and further that nearest neighbors are nearly perfect (i.e. that the k nearest neighbors always belong to the same class). For these reasons we use a sublinear function of the expected number of images per identity and set $k = \left\lceil \sqrt{\frac{|\mathcal{V}|}{|\mathcal{L}|}} \right\rceil$.

In our preliminary experiments using CRFs to structure re-identification problems we noticed a strong connection between the number of neighbors k and the λ factor in (7) controlling the balance between the data and smoothness cost. Indeed, as the CRF topology does not use a fixed grid, as in image segmentation for example, but relies on the structure induced by the nearest neighbors in feature space, the smoothness cost can easily overpower the unary penalty at vertices of high degree. Using a fixed value of λ and increasing the number of neighbors gives more weight to the smoothness cost. We therefore adapt λ to the topology of each re-identification problem by setting it to $\lambda = \frac{|\mathcal{V}|}{|\mathcal{E}|}$, which has the effect of automatically balancing the influence of unary and binary costs.

4. Experimental results

In this section we report on a number of experiments performed on structured and unstructured re-identification scenarios to compare our approach to a number of baselines and the state-of-the-art on publicly available datasets.

4.1. Datasets and feature descriptor

We evaluated the performance of our approach on several publicly available re-identification datasets, and for all experiments we use a simple descriptor of person appearance.

[5] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[4] For 3DPeS we have 1003 labeled images and 62 796 unlabeled "anonymous" images.

	ETHZ1 [2]	ETHZ2 [2]	ETHZ3 [2]	CAVIAR [6]	3DPeS [24]	CMV100 [25]
Environment	Outdoor	Outdoor	Outdoor	Indoor	Outdoor	Indoor
Cameras	1	1	1	2	8	5
Identities	83	35	28	72	191	100
Min/Avg/Max images per person	7/58/226	6/56/206	5/62/356	10/17/20	2/5/26	7/361/1245
Total images	4857	1961	1762	1220	1003 [4]	36171
Average detection size	132×60	135×63	148×66	81×34	158×74	224×75

Table 1: Characteristics of re-identification datasets.

Datasets Most publicly available re-identification datasets contain very few images per person. The VIPER [3] dataset, for example, only provides a pair of images for each identity and thus no meaningful structure can be discovered. The most interesting datasets for our approach are CAVIAR [6], ETHZ [2], 3DPeS [24] and CMV100 [25]. Here we briefly describe each of these datasets (see table 1 for a summary of their characteristics).

- **CAVIAR** The CAVIAR [6] dataset consists of several sequences recorded in a shopping center. It contains 72 persons observed from two views. 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 low average resolution of 34×81 pixels makes it difficult to extract discriminative features.

- **ETHZ** The ETHZ dataset [2] consists of three distinct datasets, each corresponding to a different outdoor sequence in which different persons appear. The ETHZ datasets are interesting because of the relatively high image resolution and the number of images per individual.

- **3DPeS** The 3DPeS dataset was acquired from 8 cameras monitoring a university campus where person images exhibit strong variations in pose, size and lighting conditions [24]. The 3DPeS re-identification dataset is a subset of the whole 3DPeS dataset and contains 1003 images of a total of 191 persons[7]. We used the detector of [26] on the entire dataset (177 904 frames in total) to obtain an additional 62 796 unique unlabeled person images. We use these "anonymous" detections to add many observations to our CRF topology and evaluate how it influences label propagation.

- **CMV100** The CMV100 dataset contains 100 persons in images acquired indoor from 5 cameras [25]. The annotation of the dataset given by the authors corresponds to the results of a tracking algorithm using background subtraction, hence the images may not be of the full body of each person. As with 3DPeS, we used the detector of [26] to extract a total of 36 171 person detections for an average of 362 images per person.

Feature descriptor To describe the visual appearance of a person we use a descriptor based on both color and texture information. We do not rely on symmetry, parts, or background/foreground modeling and we do not accumulate an averaged descriptor when given multiple gallery images. Given an input image of a target (that is a rectangular subimage containing the target), it is resized to 128×64 pixels. It is then divided into overlapping horizontal stripes of 16 pixels in height. From each stripe we extract both a Hue-Saturation (HS) histogram of 8×8 bins and an RGB histogram of $4 \times 4 \times 4$ bins.

To reduce the influence of background clutter, some re-identification techniques exploit complex methods to model background information [5] in order to exclude it from the final descriptor, or fit part-based models [6] that determine pooling regions for feature extraction. However, re-identification is often performed on very small images and we believe that fitting complex background or part models on very few pixels is unuseful. For this reason we exploit a more simple and straightforward approach obtained by weighting the contribution of each pixel to its corresponding histogram bins according to an Epanechnikov kernel centered on the target image. Finally, to include shape information we concatenate an Histogram of Oriented Gradients (HOG) descriptor, computed on a grid over the image as

[7] We removed ID 193 from 3DPeS since it is a duplicate of ID 10. ID 139 was also removed since it has only a single image.

Structured	ETHZ1			ETHZ2			ETHZ3			CAVIAR		
M =	2	5	10	2	5	10	2	5	10	2	3	5
FEAT	88.5	96.4	98.5	85.9	93.3	97.3	94.8	98.8	99.8	42.2	50	60.3
FEAT+Group	93.3	99.4	**99.6**	92.6	99.1	**100**	98.9	**100**	**100**	50.7	65	84.2
FEAT+CRF	93.5	99.4	**99.6**	92.3	99.1	**100**	98.9	**100**	**100**	50.7	65.8	85.3
SVM	89.9	96.4	97.9	87.7	92.8	95.9	95.6	98.4	99.5	48	55.7	66.3
SVM+Group	95.5	**99.5**	99.3	93.1	99.1	**100**	**99.6**	**100**	98.6	**61.7**	**77.1**	**93.5**
SVM+CRF	**95.7**	**99.5**	99.3	**93.7**	**99.4**	**100**	**99.6**	**100**	98.6	60.7	76.1	93.2

Table 2: Baseline comparison of Rank-1 accuracy for structured scenarios on ETHZ and CAVIAR datasets.

in [27], to the HS and RGB histograms. The final descriptor dimensionality is 2960. The RGB histogram captures discriminative color information about each target, the HS histogram guarantees a degree of illumination invariance, and the HOG captures local structure and texture.

4.2. Baseline evaluation

We evaluated a number of baseline algorithms as well as a number of different configurations of our approach:

- **FEAT:** A nearest neighbor classifier in feature space.
- **FEAT+Group:** Applied only to structured scenarios, it assigns to each group of test images the label for which the average distance between test images of that group and gallery images of that label is minimal.
- **FEAT+CRF:** The CRF approach using the feature data penalty defined in equation (9).
- **SVM:** Classification is based on the highest score from all identity SVMs.
- **SVM+Group:** Applied only to structured scenarios, it assigns to each group of test images the label for which the average SVM score on the group of test images from the SVM corresponding to that label is the highest.
- **SVM+CRF:** The CRF approach using the SVM data cost defined in equation (8).

In the following we use the terminology *structured scenarios* to refer to MvsM protocols, and *unstructured scenarios* for protocols where a few gallery images are given per person and all remaining images form the probe set. In structured scenarios the grouping structure of probe images is given. We feel that classification accuracy is the most important metric for real-world applications and for all experiments we report Rank-1 re-identification results. All results for our approach and baselines are averaged over ten random splits of gallery and probe image sets.

Structured scenarios and hard grouping From table 2 it is clear that hard grouping knowledge significantly helps to improve results. This can be seen by comparing the approach **FEAT** that discards the grouping knowledge with its equivalents making active use of group knowledge **FEAT+Group** and **FEAT+CRF**. The same holds for the **SVM** approach and its counterparts **SVM+Group** and **SVM+CRF** that reap too the benefits given by the hard grouping knowledge. Results on the ETHZ dataset are nearly saturated, in fact all structured results on ETHZ in table 2 for group-aware approaches are higher than 92.3%, and if we look only at results with $M \geq 5$ the minimum performance of these approaches is 98.6%. On CAVIAR, the results of the **FEAT** and **SVM** are lower and the gain of group-aware approaches is even higher. Using grouping knowledge with the SVM improves results by 13.7%, 15.4% and 27.2% for $M = 2$, 3 and 5 respectively. Selecting the label assigned to test images relying on the grouping knowledge and not considering each test image independently clearly makes the decision more robust. Group classification procedures and the CRF yield similar performance, however the group structure being unavailable for unstructured scenarios the **FEAT+Group** and **SVM+Group** cannot be applied. In the next section, we evaluate the performance of our CRF model extended to unstructured scenarios by using a k-nearest neighbor topology (see eq. 6).

Unstructured scenarios and soft grouping In table 3 we report a baseline comparison on unstructured scenarios on the ETHZ, CAVIAR and 3DPeS datasets. Note that the **FEAT+Group** and **SVM+Group** baselines cannot be evaluated on these scenarios since no explicit group structure is given. For unstructured scenarios we select M images for each person as gallery, and *all* remaining images are used as probes.

The first thing one notices from these results is that unstructured re-identification scenarios are indeed much more challenging than structured ones, as is evidenced by the significantly lower performance of all techniques with

Unstructured	**ETHZ1**				**ETHZ2**				**ETHZ3**				**CAVIAR**				**3DPeS**		
M =	1	2	5	10	1	2	5	10	1	2	5	10	1	2	3	5	1	2	3
FEAT	76.2	86.9	95.7	97.8	71.9	84	93.1	97.3	83.5	91.7	97.6	99.3	26.8	36.2	43.5	52.9	40.1	52.1	59.1
FEAT+CRF	79	89.9	96.9	**98.4**	76.3	87.8	**95**	**98**	85.4	92.9	99.3	**99.6**	27.1	36.7	44.3	53.8	41.1	53.4	61.8
SVM	80.6	89	95.8	97.4	74	85.1	92.5	95.5	85.1	93.8	98.3	99.1	30.7	42	49.4	59.6	43.5	57.7	63.7
SVM+CRF	**84.9**	**92.1**	**97.2**	98.2	**78.9**	**89.1**	94.8	97	**88.3**	**96.9**	**99.6**	99.5	**31.7**	**43.3**	**50.4**	**61.7**	**45.5**	**60.1**	**66.1**

Table 3: Baseline comparison of Rank-1 accuracy for unstructured scenarios on ETHZ, CAVIAR and 3DPeS datasets.

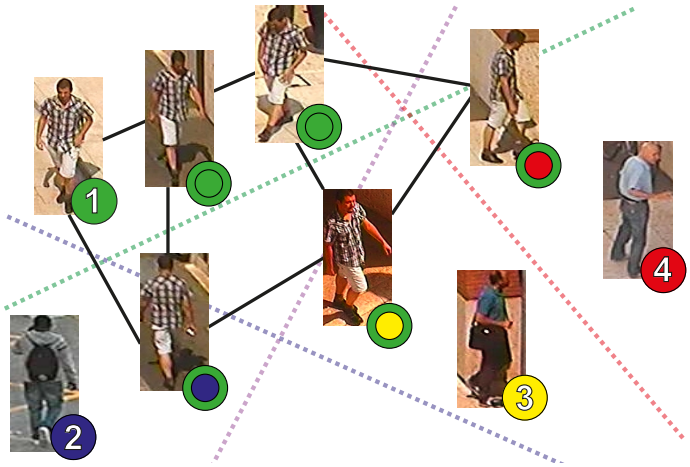


Figure 3: An example from 3DPeS: the outer circles on nodes indicate the correct class label, and the center of each circle the label given by the SVMs. Filled numbered circles are gallery images and dashed lines indicate the SVM margins. The SVM errors are all corrected by the CRF thanks to the local topology (black lines).

respect to structured results presented in table 2. Unstructured problems are more difficult of course due to the lack of structure in the probe set that is given in structured problems, but also due to the fact that more images must be correctly re-identified.

In the literature, apart from the IDINF [1, 21] approaches, only single-versus-single (i.e. $M = 1$) unstructured scenarios are considered. In table 3 we consider a range of gallery set sizes: $M \in \{1, 2, 5, 10\}$ for ETHZ, $M \in \{1, 2, 3, 5\}$ for CAVIAR and $M \in \{1, 2, 3\}$ for 3DPeS. Note how the addition of more gallery images consistently and significantly improves re-identification accuracy for all baselines. The availability of multiple examples of each person can be exploited to improve re-identification in unstructured scenarios. Moreover, the CRF approaches are able to uncover and exploit structure in the unlabeled and unstructured probe image sets. The **FEAT+CRF** and **SVM+CRF** approaches, even with only a few gallery images per person, begin to approach the Rank-1 performance of their structured counterparts on the ETHZ datasets (see tables 2 and 3). Note that unstructured and structured Rank-1 performance are not strictly comparable. In fact the unstructured problems being solved by **FEAT+CRF** and **SVM+CRF** are much more challenging than their structured counterparts as many more images are being labeled.

An illustration of SVM error correction by the CRF is given in figure 3. In this example three test images are misclassified by the SVM, indicated by the images lying on the wrong side of the correct class margin. However, exploiting the local topology these labels are corrected by the CRF. Note that this is a visualization of actual results from one SvsS trial (191 gallery and 812 test images) on 3DPeS. In section 4.4 we will elaborate on this point and show how even the addition of more unlabeled imagery actually aids our approach by enabling better estimation of the local data manifold.

4.3. Comparison with the state-of-the-art

After establishing the benefits of our CRF approach over simpler baselines, as discussed in the previous section, we performed a number of experiments to compare our approach to the current state-of-the-art. We compare to state-of-the-art results given by the authors for IDINF [1], SDALF [5], HPE [8], AHPE [8] and CPS [6]. For the

Structured	**ETHZ1**			**ETHZ2**			**ETHZ3**			**CAVIAR**		
M =	2	5	10	2	5	10	2	5	10	2	3	5
AHPE [8]	-	91	-	-	90.6	-	-	94	-	7	8	7.5
CPS [6]	-	97.7	-	-	97.3	-	-	98	-	-	13	13
HPE [8]	77	84	85	77	79	81	83	86.5	83	-	-	-
SDALF [5]	78	90.2	89.6	85	91.6	89.6	86.5	93.7	89.6	-	8.5	8.3
IDINF [1]	87	92	99	80.7	94.3	95.9	85.3	92.2	96.1	-	-	-
FEAT+CRF	93.5	99.4	**99.6**	92.3	99.1	**100**	98.9	**100**	**100**	50.7	65.8	85.3
SVM+CRF	**95.7**	**99.5**	99.3	**93.7**	**99.4**	**100**	**99.6**	**100**	98.6	**60.7**	**76.1**	**93.2**

Table 4: Comparison with the state-of-the-art for structured scenarios on ETHZ and CAVIAR datasets.

Unstructured	**ETHZ1**				**ETHZ2**				**ETHZ3**				**CAVIAR**
M =	1	2	5	10	1	2	5	10	1	2	5	10	1
SDALF [5]	64.8	-	-	-	64.4	-	-	-	77	-	-	-	-
AHPE [8]	-	-	-	-	-	-	-	-	-	-	-	-	7
CPS [6]	-	-	-	-	-	-	-	-	-	-	-	-	8.5
MR-L_u k=4 [18]	78.8	-	-	-	73.7	-	-	-	85.1	-	-	-	28.1
MR-L_u k=15 [18]	78.1	-	-	-	73.3	-	-	-	84.8	-	-	-	27.7
MFL opt. [19]	-	-	-	-	-	-	-	-	-	-	-	-	8.2
IDINF [1]	69.7	83.3	92.2	96.1	65.7	80.4	89.5	89.5	88.1	93.6	98.4	98	-
FEAT+CRF	79	89.9	96.9	**98.4**	76.3	87.8	**95**	**98**	85.4	92.9	99.3	**99.6**	27.1
SVM+CRF	**84.9**	**92.1**	**97.2**	98.2	**78.9**	**89.1**	94.8	97	**88.3**	**96.9**	**99.6**	99.5	**31.7**

Table 5: Comparison with the state-of-the-art for unstructured scenarios on ETHZ and CAVIAR datasets.

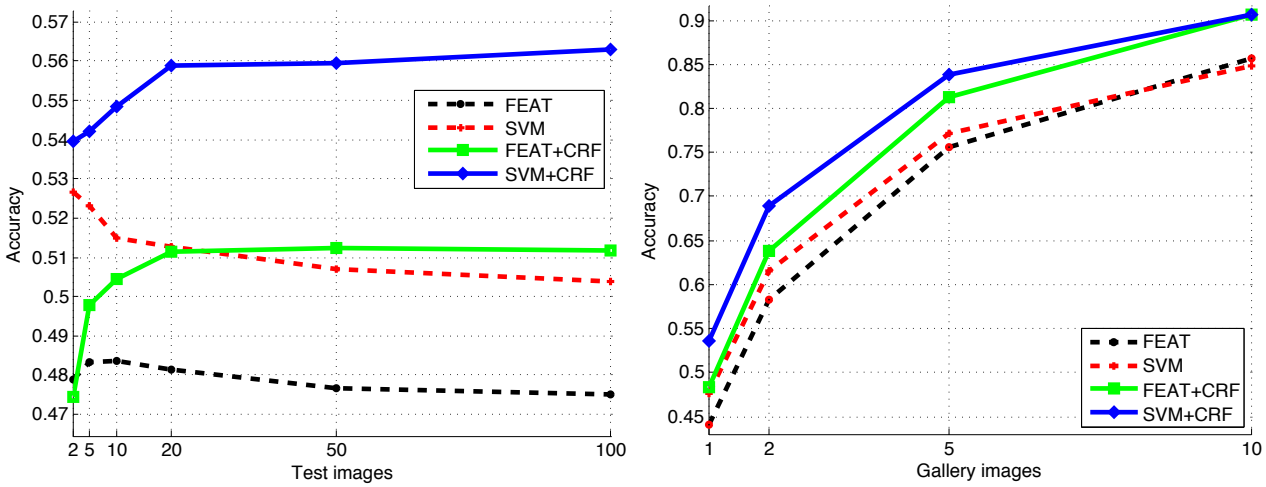
semi-supervised manifold ranking approach, referred to as "MR-L_u" in table 5, we re-implemented the unnormalized Laplacian method described in [18] using our descriptor and report results for the number of neighbors ($k = 15$) used by the authors and also for $k = 4$ which was found to be the best value of k from the set $\{2, 4, 8, 15\}$. The β parameter controlling the stability of manifold ranking was set to 100, which was found to be the best value in from the set $\{0.01, 1, 100, 10000\}$. Note that, as in the original work [18], manifold ranking is only evaluated for unstructured re-identification with a single gallery image as it is not directly applicable to structured or multi-shot scenarios.

Structured scenarios and hard grouping Here we evaluate performance on structured scenarios using the ETHZ and CAVIAR datasets. In these scenarios a fixed number of M gallery and M probe images per subject are selected. Rank-1 results for our approaches and the state-of-the-art are given in table 4 for the ETHZ and CAVIAR datasets.

Our method outperforms all state-of-the-art approaches on ETHZ and CAVIAR. The improvement over our previous work using a CRF over a bag-of-words representation [1] is more noticeable for values of $M \leq 5$, which is due both to the feature used in this work, that is more powerful and adapted for re-identification, and also to the use of discriminative models. The comparison with the other approaches proposed in the literature shows a significant performance improvement that also tends to get wider with higher M values. One explanation is that most approaches compute an average or accumulated representative for each gallery and probe group, while **FEAT+CRF** maintains all individual feature descriptors in both gallery and probe sets, and **SVM+CRF** exploits all gallery images per identity to learn a single SVM but keeps multiple features in the probe set. The performance gap is even wider on the CAVIAR dataset. Our approach obtains 60.7%, 76.1% and 93.2% with M=2, 3 and 5, respectively. The best state-of-the-art performance up to now were 7%, 13% and 13%, again for M=2, 3 and 5, respectively. This is likely due to the fact that state-of-the-art approaches often fit complex appearance models (identifying axis of symmetry and asymmetry for SDALF and AHPE, or part-based models for CPS) on low-resolution images (see table 1).

Unstructured scenarios and soft grouping For ETHZ and CAVIAR we report results in table 5. On all datasets the **SVM+CRF** approach performs best most of the time, while **FEAT+CRF** gives the best performance on ETHZ when many gallery images are available. This is likely due to the fact that the **SVM** learns a single model while the **FEAT** approach maintains all gallery descriptors independently until test time.

Our approach outperforms other semi-supervised methods such as IDINF [1], MFL [19] (with optimized param-



(a) CMV100 unstructured scenarios varying the number of test images while using a single gallery image per person.

(b) CMV100 unstructured scenarios varying the number of gallery images using all remaining images as test images.

Figure 4: Results for unstructured scenarios on CMV100.

eters) and MR-L_u [18] on all datasets. In the case of IDINF, this is due to our use of discriminative models for data costs and to the use of a more discriminative feature. In the case of MFL with optimized parameters, we outperform them on CAVIAR, which is likely due to the fact that they extract many features but from a pictorial structure model fit to images of very low resolution. In the case of MR-L_u, however, it is due to the ability of our CRF-based formulation to exploit all available test data during inference. In contrast, MR-L_u labels each probe image independently and is only able to exploit gallery images and the single probe image being labeled.

4.4. Large scale person re-identification

In this section we report on a series of experiments we performed on the CMV100 and 3DPeS datasets to explore issues related to large scale person re-identification and to quantify the performance of our approach on large datasets. The CMV100 dataset has a total of 36 171 images depicting 100 persons. The 3DPeS dataset has 1003 labeled images of 191 persons and we extracted a total of 62 976 unlabeled "anonymous" detections. See table 1 for a summary of the characteristics of these two datasets. To the best of our knowledge, we are the first to consider such large re-identification problems.

Re-identification over very large probe sets In figure 4a we report unstructured re-identification results on CMV100 for a single probe image and the number of probe images per person varying from two to one hundred. Each curve represents one of the baseline configurations defined in section 4.2. Note how the performance of our CRF approaches consistently improve with increasing number of probe images, indicating that CRF is effective in leveraging the hidden feature space structure to improve accuracy. The performance of the **SVM** and **FEAT** baselines, however, decreases with added probe imagery. Specifically, while the SVM performance decreases when adding more test images (going down from 52.7% Rank-1 accuracy for 2 test images to 50.4% for 100 test images per person), the SVM+CRF approach follows the exact opposite trend going from 54% up to 56.3%. With more test images, better local structure can be discovered.

Figure 4b instead illustrates the Rank-1 unstructured re-identification accuracy of our approaches as a function of increasing number of gallery images. For these experiments the scenario could be called "multi-versus-all" re-identification since the task is, given M gallery images per person, to label *all remaining images in the dataset*. From figure 4b we can see that all of our approaches benefit from additional exemplars of each person in the gallery. Note also that the discriminative models of the **SVM+CRF** approach are most useful for limited numbers of gallery images. At about ten gallery images per person, the **FEAT+CRF** approach performs approximately equivalently.

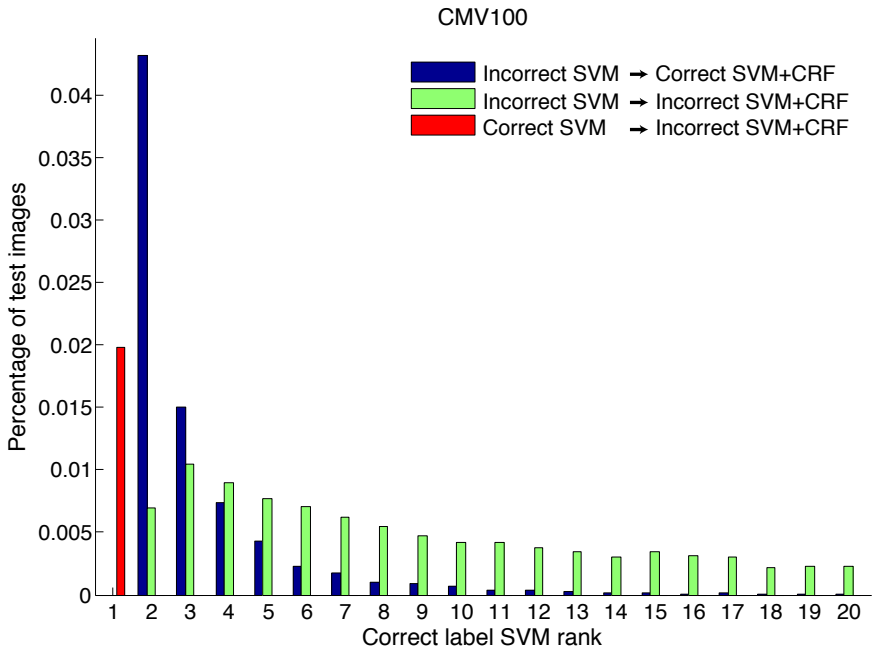


Figure 5: Label switch analysis between SVM and SVM+CRF approaches on CMV100. Results are averaged over 10 trials with a single gallery image and 100 test images.

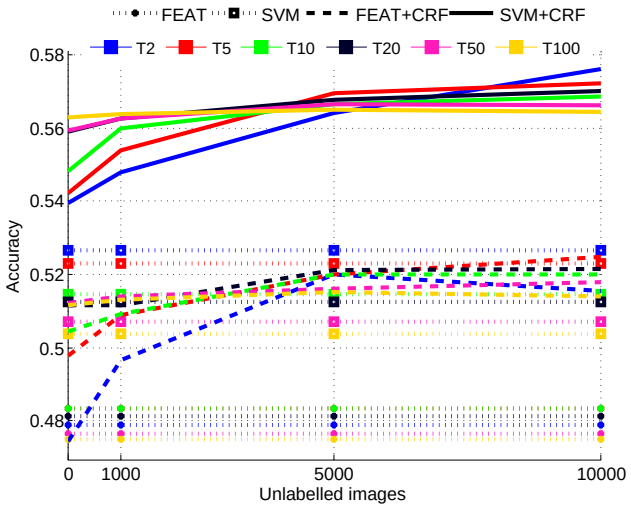
There are no directly comparable state-of-the-art results on the CMV100 dataset, but considering a sequence-based matching task, [25] reports a 28% Rank-1 accuracy. Using 5 images per person (for CMV100 this implies 500 images for the gallery), our approach correctly labels 30 745 images out of the 35 671 test images, or a Rank-1 accuracy of 85%.

Figure 5 shines more light on the contribution of the CRF to re-identification on large probe sets. In this figure we plot the proportion of test images whose labels switch when applying the CRF on top of SVM outputs. Switch proportions are further broken down in this plot according to the rank at which the correct identity is returned according to the SVM scores. There are three possibilities for these labels switches:

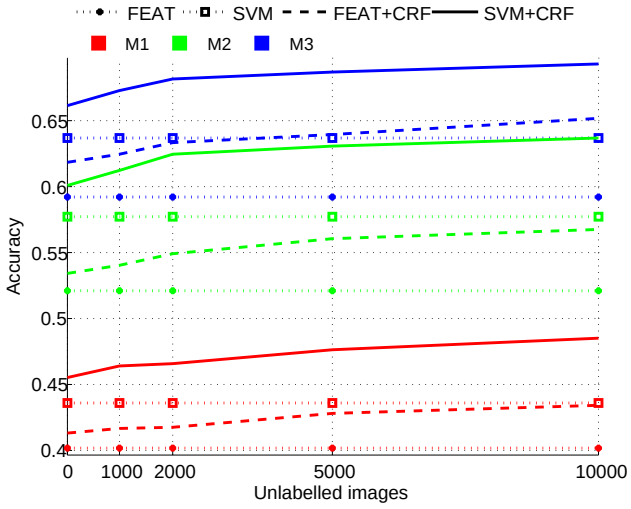
- The SVM gives the correct label but the SVM+CRF switches it to an incorrect one (red bars in figure 5), note that this is possible only if the correct label is ranked first by the SVM.
- The SVM+CRF switches an incorrect label given by the SVM to the correct one (blue bars in figure 5).
- The SVM+CRF switches an incorrect label given by the SVM to another incorrect one (green bars in figure 5).

From this figure we see that errors even at high ranks can be corrected by the CRF, and that only a small portion of test images switch to incorrect labels due to label inference with local constraints. The total gain in accuracy attributable to the CRF is the sum of all blue bars minus the red bar.

The benefits of anonymous unlabeled imagery Re-identification datasets are often constructed by sub-sampling larger video surveillance datasets. We believe that this results in datasets posing more difficult problems than what would arise in the real-word applications they are supposed to model. We therefore investigate what happens if we incorporate discarded images in the testing phase on 3DPeS and CMV100. These additional images are acquired by running a person detector on all frames of the original dataset. One of the unique features of our CRF approaches is that we are able to leverage all available imagery: gallery images, probe images, and when available also what we call "anonymous" unlabeled person images for which no labels are desired. The idea is that these unlabeled images help to better sample the local manifold and thus to improve the propagation of labels during inference. Results on CMV100 and 3DPeS are given in figures 6a and 6b. Note that the first column corresponds to standard unstructured



(a) CMV100 unstructured scenarios using one gallery image per person while fixing the number of test images (T) and adding unlabeled images. Best viewed in color.



(b) 3DPeS unstructured scenarios fixing the number of gallery images (M) and adding unlabeled images.

Figure 6: Results for unstructured scenarios with additional unlabeled images on CMV100 and 3DPeS.

scenarios as no unlabeled images are added and thus are strictly equal to results given in figure 4a for CMV100 and table 3 for 3DPeS.

These results demonstrate that adding unlabeled images generally results in performance gains. This is particularly true when the number of test images is low as in 3DPeS or when using small T values on CMV100. The addition of unlabeled data results in a denser sampling in feature space and thus a better approximation of the manifold structure by our nearest-neighbor topology in the CRF. This enhanced topology can help to obtain a better final labelling thanks to more meaningful local constraints. In figure 7 we illustrate this with an example from 3DPeS. In the top of figure 7 one probe example of identity **1** is incorrectly labeled as **2**. In the bottom example, however, the addition of unlabeled images creates a better connectivity, as connections between the mislabeled probe and images taking the correct label are enforced while connections with other labels are discarded, and thus enabling the inference procedure to correct these errors.

As with CMV100, for 3DPeS there are no state-of-the-art re-identification results directly comparable to our approach. The authors of [24], however, report a Rank-1 accuracy of 37.5% using a subset of 100 persons, 3 to 5 gallery images per person, and a single test image. From figure 6b we see in the first column that with only two gallery images per person, we are able to label correctly 60.1% of *all* remaining images of the 3DPeS dataset without the addition of anonymous unlabeled imagery.

Computational cost To give some insight into the computational cost of our approach we report some timing results (minus the feature extraction process) on the 3DPeS dataset on a computer with an Intel Xeon X5650 CPU and 12GB of RAM. When applying our method to a limited number of images (191 gallery, 812 test), the total computational cost of one run (9.7s) is dominated by the SVM training time (7.3s). However, after adding 5000 unlabeled images the total cost of a run (24.4s) is dominated by the construction of the nearest neighbor topology (10.6s), since our current naive implementation is quadratic in the number of nodes. However, finding the k-nearest neighbors is a common problem and more efficient approximate methods such as [28] could be employed if very large scale problems were to be solved with our approach.

5. Conclusions

In this article we described a semi-supervised approach to combining discriminative models of person identity with a conditional random field model of local feature-space topology to solve re-identification problems. Both the

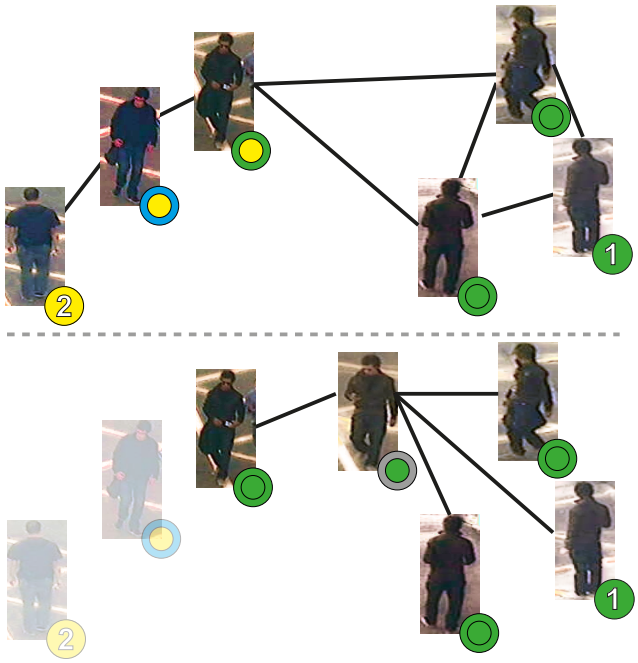


Figure 7: Illustration of how adding unlabeled samples (nodes with outer circle in grey) improves the topology in the CRF (black lines) and improves labelling. This example is based on actual labelling results of the **SVM+CRF** approach on one trial from 3DPeS without unlabeled samples (top row) and with unlabeled samples (bottom row).

discriminative models and the CRF improve re-identification results over simpler baselines, and the combination of both exhibit the most significant improvement. Our approach can efficiently solve structured re-identification problems, but the combination of linear SVMs and CRF excels particularly in the case of more difficult, unstructured re-identification problems where no group structure is given. Experimental results demonstrate that, even in cases of re-identification of very many probe images on the basis of very few gallery images, our approach performs very well. We further demonstrated that adding additional, unlabeled images to re-identification problems increases the performance of our approach.

The performance of our approach scales very well with increasing amounts of test data due to more dense sampling of the manifold in feature space. This is the opposite of what occurs for discriminative models like SVMs whose performance degrades when adding test data without additional training data. Our approach is able to make the most out of *all* available data at re-identification time, while discriminative models and existing semi-supervised approaches like manifold ranking are limited to exploiting only images from the gallery set. We believe that in real-world re-identification scenarios there will be many times more unlabeled images available than labeled ones, and in this work we have demonstrated the effectiveness of our approach on very large re-identification problems involving tens of thousands of probe images.

The combination of weak discriminative models and the relatively weak topology in feature space allows our technique to achieve some of the benefits of stronger discriminative models or metric learning. Our improved performance, however, does not come at the cost of setting aside a portion of available data for learning strong discriminants or metrics. We finally note that the approach we describe in this paper is not limited to re-identification problems and can be profitably and more broadly applied to other recognition tasks [29]. It will also be interesting to investigate the relationship between the techniques proposed in this article and more traditional approaches to metric learning.

Acknowledgements

This work was partially supported by Thales Italia and the MNEMOSYNE project (POR-FSE 2007-2013, A.IV-OB.2). Andrew D. Bagdanov acknowledges the support of Ramon y Cajal Fellowship RYC-2012-11776.

References

[1] S. Karaman, A. D. Bagdanov, Identity inference: generalizing person re-identification scenarios, in: Proceedings of ECCV - Workshops and Demonstrations, 2012, pp. 443–452. 2, 4, 11, 12

[2] W. Schwartz, L. Davis, Learning discriminative appearance-based models using partial least squares, in: Proceedings of SIBGRAPI, 2009, pp. 322–329. 3, 9

[3] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: Proceedings of the European Conference on Computer Vision (ECCV), 2008, pp. 262–275. 3, 9

[4] S. Bak, E. Corvee, F. Bremond, M. Thonnat, Multiple-shot human re-identification by mean riemannian covariance grid, in: Proceedings of AVSS, 2011, pp. 179–184. 3

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2360–2367. 3, 9, 11, 12

[6] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification, in: Proceedings of the British Machine Vision Conference (BMVC), 2011, pp. 68.1–68.11. 3, 9, 11, 12

[7] Y. Cai, M. Pietikäinen, Person re-identification based on global color context, in: Proceedings of ACCV Workshops, 2011, pp. 205–215. 3

[8] L. Bazzani, M. Cristani, A. Perina, V. Murino, Multiple-shot person re-identification by chromatic and epitomic analyses, Pattern Recognition Letters 33 (7) (2012) 898–903. 3, 11, 12

[9] X. W. Rui Zhao, Wanli Ouyang, Unsupervised salience learning for person re-identification, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3586 – 3593. 3

[10] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, Person re-identification by support vector ranking, in: Proceedings of the British Machine Vision Conference (BMVC), 2010, pp. 21.1–11. 3

[11] Y. Wu, M. Minoh, M. Mukunoki, S. Lao, Set based discriminative ranking for recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), 2012, pp. 497–510. 3

[12] W.-S. Zheng, S. Gong, T. Xiang, Person re-identification by probabilistic relative distance comparison, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 649–656. 3

[13] T. Avraham, I. Gurvich, M. Lindenbaum, S. Markovitch, Learning implicit transfer for person re-identification, in: Proceedings of ECCV - Workshops and Demonstrations, 2012, pp. 381–390. 3

[14] M. Hirzer, P. M. Roth, M. Köstinger, H. Bischof, Relaxed pairwise learned metric for person re-identification, in: Proceedings of the European Conference on Computer Vision (ECCV), 2012, pp. 780–793. 3

[15] S. Bak, G. Charpiat, E. Corvee, F. Bremond, M. Thonnat, Learning to Match Appearances by Correlations in a Covariance Metric Space, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2012, pp. 806–820. 3

[16] S. Pedagadi, J. Orwell, S. Velastin, Local Fisher Discriminant Analysis for Pedestrian Re-identification, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3318 – 3325. 4

[17] R. Layne, T. Hospedales, S. Gong, Person re-identification by attributes, in: Proceedings of the British Machine Vision Conference (BMVC), 2012. 4

[18] C. C. Loy, C. Liu, S. Gong, Person re-identification by manifold ranking, in: Proceedings of IEEE International Conference on Image Processing, 2013, pp. 3567 – 3571. 4, 12, 13

[19] D. Figueira, L. Bazzani, H. Q. Minh, M. Cristani, A. Bernardino, V. Murino, Semi-supervised multi-feature learning for person re-identification., in: Proceedings of AVSS, 2013, pp. 111–116. 4, 12

[20] A. Blum, S. Chawla, Learning from labeled and unlabeled data using graph mincuts, in: Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 19–26. 4

[21] S. Karaman, G. Lisanti, A. D. Bagdanov, A. Del Bimbo, From re-identification to identity inference: Labeling consistency by local similarity constraints, in: S. Gong, M. Cristani, S. Yan, C. C. Loy (Eds.), Person Re-Identification, Springer, 2014, pp. 287–307. 4, 11

[22] T. Malisiewicz, A. Gupta, A. A. Efros, Ensemble of exemplar-svms for object detection and beyond, in: Proceedings of the International Conference on Computer Vision (ICCV), 2011, pp. 89–96. 6

[23] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (11) (2001) 1222–1239. 7

[24] D. Baltieri, R. Vezzani, R. Cucchiara, 3dpes: 3d people dataset for surveillance and forensics, in: Proceedings of the First International ACM Workshop on Multimedia access to 3D Human Objects, 2011. 9, 15

[25] V. Takala, M. Pietikainen, Cmv100: A dataset for people tracking and re-identification in sparse camera networks., in: Proceedings of ICPR, 2012, pp. 1387–1390. 9, 14

[26] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, D.m.: Cascade object detection with deformable part models, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2241–2248. 9

[27] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 886 – 893. 10

[28] W. Dong, C. Moses, K. Li, Efficient k-nearest neighbor graph construction for generic similarity measures, in: Proceedings of WWW, 2011, pp. 577–586. 15

[29] S. Karaman, L. Seidenari, A. D. Bagdanov, A. Del Bimbo, L1-regularized logistic regression stacking and transductive crf smoothing for action recognition in video, in: Results of the THUMOS 2013 Action Recognition Challenge with a Large Number of Classes, 2013. 16