RE-IDENTIFICATION OF PERSONS IN VIDEOS AND IMAGES

Vittorio Murino [1,2] and Marco Cristani [2]

Problem definition, performance metrics, and datasets

Person re-identification (*re-id*) is defined as the problem of recognizing an individual captured in different times and/or locations over several non-overlapping camera views, considering a large set of candidates. This is a very difficult problem as most of the time people are captured by diverse low resolution cameras, under occlusions conditions, badly (and different from view to view) illuminated, and in varying poses. In this context, a robust modeling of the entire body appearance of a person is mandatory, especially when other classical biometric cues (face, gait) are not available or difficult to catch, due to the sensors' scarce resolution or low frame-rate.

Re-id algorithms are typically tested considering *ad hoc* datasets composed by a fixed number of individuals, each one represented by several (minimum two) image windows containing or bounding boxes surrounding a person, typically in different poses (e.g., frontal, back, profile, etc.). The actual re-identification is performed by considering one instance of an individual, the so-called "probe", which is compared and matched against all the other instances of all subjects in the dataset, the so-called "gallery" set. Hence, given a set of probes, an re-id method provides a ranking of a certain number of subjects in the gallery set, typically ranging from 1 to 25, 50 or 100, depending on the used dataset. Algorithms' performance is commonly calculated by the recognition rate estimated by the Cumulative Matching Characteristic (CMC) curve, and the normalized Area Under Curve (nAUC) score for the CMC curve. The CMC curve is a plot of the recognition performance vs. the ranking score and represents the expectation of finding the correct match in the top n matches; nAUC gives an overall score of how well a re-identification method performs overall.

The most used image datasets are VIPER, iLIDS, ETHZ, and CAVIAR4REID, which differs for a number of characteristics: number of subjects considered (in the order of hundreds), number of instances per subject (from 2 to 10), pose and lighting variations, severity of the occlusions, number of viewpoints (cameras), window resolution (from 17×39 up to 72×144, 48×128, 64×128, etc.). Furthermore, 3D datasets are now becoming available for re-id research.

All the following developed methods proved to perform better than state of the art at the moment of the publication.

State of the art

Given the datasets as composed by image crops containing a pedestrian, the re-id problem becomes "simply" to identify suitable features to build a discriminant descriptor or signature, and how to match such descriptors to find the correct correspondence. However, in actual applications, the scenario is much more complex implying the detection (and possibly tracking) of a person, which is not perfect in many cases, often preceded by a background subtraction stage which may affect the detection as well.

Nevertheless, disregarding real scenarios, the re-id literature is mainly characterized by appearance-based methods which can be categorized in direct and learning-based approaches, and in single- and multiple-shot techniques. *Learning-based* techniques are characterized by the use of a training dataset of different individuals where the features and/or the policy for combining them are analyzed to ensure high re-identification accuracy. The underlying assumption is that the knowledge extracted from the training set can generalize to unseen samples, so allowing the learning of a discriminative model or the distance metrics from data. Binary Support Vector Machines (SVM) [1], multi-class SVM [2], nearest neighbor classifier [3], partial least square reduction [4], boosting [5, 6], distance learning [7, 8, 9], descriptor learning [10], and ensemble RankSVM [12] have been customized for the re-identification problem.

Direct methods do not consider any training set as they are usually focused on finding discriminant parts of the human appearance, and on manually designing features that perform very well on a particular re-id scenario. The framed person is typically subdivided into horizontal stripes [13], symmetrical and asymmetrical parts [14], semantic parts [15, 16], regions clustered by color [17], concentric rings [18], or a grid of localized

[1] Pattern Analysis & Computer Vision (PAVIS), Istituto Italiano di Tecnologia, Italy.

[2] Dipartimento di Informatica, University of Verona, Italy.

patches [19]. Several types of features can be extracted from these regions: color histograms or other statistics [13, 14, 16], maximally stable color regions [12], depth features [11], histogram of oriented gradients [4], Gabor and Schmid filters [12], interest points [20], covariance matrices [7, 21], attributes [22] and Haar-like features [5] have been exhaustively tested in the literature.
Single-shot approaches focus on associating pairs of images, each containing one instance of an individual (e.g., [2, 3, 4, 12]). Multi-shot methods employ multiple images of the same person as probe and/or gallery elements (e.g., [14, 15, 19, 23]). The assumption of the multi-shot methods is that individuals can be tracked so that it is possible to gather a lot of images, and use this richer information for, e.g., calculating more robust features and/or descriptors.
Our experience in re-id lies mainly in direct methods, and we recently investigated a learning-based method. Both single- and multiple-shot modality algorithms have been addressed.

The proposed methods
One of the first and most performing approaches is SDALF, Symmetry-Driven Accumulation of Local Features [A][H]. This method exploits the symmetry of the human body by weighting differently areas close to the symmetry axes with respect to regions far away from them. After a pre-processing phase, salient parts of the body figure are extracted by adopting perceptual principles of symmetry and asymmetry. First, we find two horizontal axes of asymmetry that isolate three main body regions, usually corresponding to head, torso and legs. Head is discarded and for the other two parts, a vertical axis of appearance symmetry is estimated. Then, complementary aspects of the human body appearance are detected on each part, highlighting: i) the general chromatic content via HSV histogram; ii) the per-region color displacement through Maximally Stable Colour Regions (MSCR); iii) the presence of Recurrent Highly Structured Patches (RHSP), estimated through a novel per-patch similarity analysis. The extracted features are weighted considering the distance with respect to the vertical axis, so that the effects of pose variations are minimized. The matching between the candidates is estimated by calculating a weighted sum of the matching distances per each feature type, adopting Euclidean or Bhattacharyya distance. This approach applies to both the cases where a single image for each candidate is present, and the cases where multiple images for each individual (not necessarily temporally adjacent) are available, by properly accumulating the local features in a single signature.

The subsequent work focused on the multiple-shot case, presenting an appearance-based direct algorithm based on the extraction and matching of a signature that embeds global and local appearance features [B][C]. After a selection of the foreground image areas (likely containing a person) via clustering and person classification (the SCA model by Jojic et al. in CVPR 2009), complementary aspects of the human appearance are extracted highlighting: 1) the global chromatic content via a mean color histogram, and 2) the presence of recurrent local patterns through epitomic analysis proposed by Jojic et al. in CVPR 2003. The former captures all the chromatic information of an individual's appearance. The latter is supported by the paradigm of object recognition by local features encoding the pixels' local spatial layout with a set of frequently visible patches, also allowing to properly accumulate images in a multi-shot descriptor. This signature is called Histogram Plus Epitome, HPE. A variant consisting on the asymmetry-based segmentation proposed in SDALF was also proposed giving rise to the Asymmetry-based HPE.

Taking benefit of the SDALF experience and further pushing the idea of characterizing significant body parts, the subsequent work focused on a re-id method based on Pictorial Structures (PS) for human body pose estimation [D]. We build upon a standard PS framework where general part detectors localize the body parts, and a kinematic tree prior captures the whole structural knowledge. The idea here was inspired from the way humans perform re-id, i.e., by concentrating the attention to certain peculiar ("salient") body parts, looking for part-to-part correspondences. After fitting a PS on all images, from each localized part we extract an ensemble of features, encoding complementary aspects, such as the chromatic content and the spatial arrangement of colors. The first aspect is captured by HSV histograms, while the second aspect is codified by MSCR, also previously adopted in SDALF. The features of each part are subsequently combined into a single ID signature. Matching between signatures is carried out by standard distance minimization strategies, and the method applied to both single- and multi-shot cases. For the latter, we proposed a strategy to improve the PS fitting on images of the same subject, consisting in learning the local appearance of each part in a given subject so that *ad hoc* appearance part detectors can provide more accurate PS fitting. This new model was called Custom

Pictorial Structure (CPS): once CPS is fitted on data, features are extracted from each instance as in the single-shot case, and the individual signatures are pooled together to obtain a multi-shot ID signature.

To make more robust the image appearance-based methods, possibly relaxing the underlying hypothesis of maintenance of the same person clothing, we developed a new approach that uses soft biometric cues as features, as alternative or complementary to classical re-id [G]. In this method, discriminant cues are extracted from range data acquired using RGB-D cameras, such as the MS Kinect now available, to acquire depth information in a fast and affordable way. The idea here is to consider features capable to embody more implicit human body characteristics, i.e., related to specific anthropometric measurements. We introduced two distinct sets of features: the first one represents cues computed from the fitted skeleton to depth data, i.e. the Euclidean distance between selected body parts such as legs, arms and the overall height; the second set contains features computed on the surface given by the range data, coming in the form of geodesic distances computed from a predefined set of joints (e.g., from torso to right hip). The latter measure gives an indication of the curvature (and, by approximation, of the size) of specific regions of the body. We analyzed the effectiveness of each feature separately and how they have to be weighted in order to maximize the re-identification performance. We obtained promising results on a dataset of 79 persons, acquired over different places and intervals of days.

As for learning-based methods, we recently proposed a technique [I] that subsumes the best aspects of direct and learning-based methods. This technique allows the exploitation of multiple features independently of their nature and, at the same time, does not require the classifier training for each pair of images containing the same person (since unavailable in many real cases). Our approach casts re-id as a semi-supervised multi-class recognition problem, where each class corresponds to the identity of one individual. In particular, we exploit the general framework of multi-view learning [J] with manifold regularization in vector-valued Reproducing Kernel Hilbert Spaces (RKHS). In this setting, each feature is associated with a component (view) of a vector-valued function in an RKHS. Unlike multi-kernel learning, all components of a function are forced to map in the same fashion, i.e., to distinguish in a coherent way the different individuals. The desired final output is given by their combination which is a fusion mechanism joining together the different features. This approach trains a classifier from a labeled (gallery) set of P different individuals, exploiting the structure of unlabeled data that can be the probe set or other images possibly acquired during tracking. In other words, it does not require to have inter-camera image pairs of the same person, but only a single labeled image per person, thus making our approach truly applicable in real scenarios.

The road ahead

Much work has been done for re-identification but we are not yet in a position to deploy a method in a real scenario with a certain effectiveness and reliability for several reasons. First of all, most of the techniques are now tested on datasets only, overlooking the problems derived by a non optimal person detection and tracking, and this may heavily affect re-id performance. Second, a real scenario likely implies the online construction of the dataset leading to problems of efficient storage and retrieval of the person signature, which may also possibly reach huge dimensions hence requiring effective retrieval strategies. To date, these issues have not yet been sufficiently addressed.

The road ahead is anyway paved. Thanks to the new sensor technologies, in particular RGB-D cameras, the use of soft biometric cues can aid the image-based methods and make re-id more robust and valuable.

Another interesting research line regards the use of a pan-tilt-zoom (PTZ) camera for re-id [K]. The idea here is to build a person descriptor which contains both typical information (full body, specific areas) at a certain resolution, and also specific peculiar human regions at higher resolution (e.g., a leg or an arm), to capture discriminant characteristics (e.g., a tattoo) able to uniquely distinguish among the subjects.

A new recent trend, slightly deviating from the surveillance aspects, consists in the re-identification of human subjects on the basis of their behavior in the use of social media. An example consists in analyzing Skype chats [F][L] and extract a descriptor containing non-semantic features like e.g., average number of character per word, misspelled words, used emoticons, turn duration, to cite a few. If this can be done for a number of chats per person (i.e., we get a reliable training set), it is possible to build a signature which has been proved to be effective to re-identify a subject when engaged in a new chat. This is interesting trend given the large diffusion of social media nowadays and may support enhanced security in the cyberworld making user identity recognition and verification more effective in a fully transparent way.

References (authors work)

[A] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani: Person re-identification by symmetry-driven accumulation of local features. CVPR 2010: 2360-2367

[B] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, V. Murino: Multiple-Shot Person Re-identification by HPE Signature. ICPR 2010: 1413-1416

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

[D] D.S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino: Custom Pictorial Structures for Re-identification. BMVC 2011: 1-11

[E] R. Satta, G. Fumera, F. Roli, M. Cristani, V. Murino: A Multiple Component Matching Framework for Person Re-identification. ICIAP (2) 2011: 140-149

[F] M. Cristani, G. Roffo, C. Segalin, L. Bazzani, A. Vinciarelli, V. Murino: Conversationally-inspired stylometric features for authorship attribution in instant messaging. ACM Multimedia 2012: 1121-1124

[G] I.B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, V. Murino: Re-identification with RGB-D Sensors. 1st Workshop on Re-identification (ECCV 2012 Workshop), 2012: 433-442

[H] L. Bazzani, M. Cristani, V. Murino: Symmetry-driven accumulation of local features for human characterization and re-identification. Computer Vision and Image Understanding 117(2): 130-144 (2013)

[I] D. Figueira, L. Bazzani, Ha Quang Minh, M. Cristani, A. Bernardino, V. Murino: Semi-supervised multi-feature learning for person re-identification. AVSS 2013: 111-116

[J] H.Q. Minh, L. Bazzani, V. Murino: A unifying framework for vector-valued manifold regularization and multi-view learning. ICML 2013: 100-108

[K] P. Salvagnini, L. Bazzani, M. Cristani, V. Murino: Person Re-Identification with a PTZ Camera: An Introductory Study. ICIP 2013: 3552-3556

[L] G. Roffo, C. Segalin, A. Vinciarelli, V. Murino, M. Cristani: Reading between the turns: Statistical modeling for identity recognition and verification in chats. AVSS 2013: 99-104

References (from the state of the art)

[1] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning implicit transfer for person re-identification. In ECCV Workshops, 2012.

[2] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. Pattern Recognition Letters, 36(9):1997–2006, 2003.

[3] Z. Lin and L. S. Davis. Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In Advances in Visual Computing, 2008.

[4] W.R. Schwartz and L.S. Davis. Learning discriminative appearance-based models using partial least squares. In Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing, 2009.

[5] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person Re-identification Using Haar-based and DCD-based Signature. In AMMCSS, 2010.

[6] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In ECCV, 2008.

[7] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In ECCV, 2012.

[8] M. Kostinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In CVPR, 2012.

[9] W. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. IEEE Trans. Pattern Anal. Mach. Intell., 35(3), 2012.

[10] L. Bazzani, M. Cristani, A. Perina, and V. Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recognition Letters, 2011.

[11] I.B. Barbosa, M. Cristani, A. Bue, L. Bazzani, and V. Murino. Re-identification with RGB-D sensors. In ECCV workshop on Re-ID, 2012.
[12] B. Prosser, W. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In BMVC, 2010.
[13] N. Bird, O. Masoud, N. Papanikolopoulos, and A. Isaacs. Detection of loitering individuals in public transportation areas. IEEE Trans. on Intelligent Transportation Systems, 6(2):167 – 177, 2005.
[14] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. Computer Vision and Image Understanding, 117(2):130 – 144, 2013.
[15] D.S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In BMVC, 2011.
[16] D. Figueira and A. Bernardino. Re-Identification of Visual Targets in Camera Networks a comparison of techniques. In ICIAR, 2011.
[17] X. Wang, G. Doretto, T. B. Sebastian, J. Rittscher, and P. H.Tu. Shape and appearance context modeling. In ICCV, 2007.
[18] W.S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In BMVC, 2009.
[19] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Multiple-shot human re-identification by mean Riemannian covariance grid. In AVSS, 2011.
[20] N. Gheissari, T. B. Sebastian, P. H. Tu, J. Rittscher, and R. Hartley. Person reidentification using spatiotemporal appearance. In CVPR, 2006.
[21] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Boosted human re-identification using Riemannian manifolds. Image Vision Comput., 30(6-7):443–452, June 2012.
[22] R. Layne, T. Hospedales, and S. Gong. Person re-identification by attributes. In BMVC, 2012.
[23] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In BMVC, 2006.