Custom Pictorial Structures for Re-identification

Dong Seon Cheng[1]
cheng.dong.seon@gmail.com
Marco Cristani[1,2]
marco.cristani@univr.it
Michele Stoppa[2]
michele.stoppa@iit.it
Loris Bazzani[1]
loris.bazzani@univr.it
Vittorio Murino[1,2]
http://profs.sci.univr.it/~swan

[1] Dipartimento di Informatica
University of Verona
Italy

[2] Istituto Italiano di Tecnologia
Via Morego, 30
16163 Genova, Italy

Abstract

We propose a novel methodology for re-identification, based on Pictorial Structures (PS). Whenever face or other biometric information is missing, humans recognize an individual by selectively focusing on the body parts, looking for part-to-part correspondences. We want to take inspiration from this strategy in a re-identification context, using PS to achieve this objective. For single image re-identification, we adopt PS to localize the parts, extract and match their descriptors. When multiple images of a single individual are available, we propose a new algorithm to customize the fit of PS on that specific person, leading to what we call a Custom Pictorial Structure (CPS). CPS learns the appearance of an individual, improving the localization of its parts, thus obtaining more reliable visual characteristics for re-identification. It is based on the statistical learning of pixel attributes collected through spatio-temporal reasoning. The use of PS and CPS leads to state-of-the-art results on all the available public benchmarks, and opens a fresh new direction for research on re-identification.

1 Introduction

Human re-identification (re-id) consists in recognizing a person in different locations over various non-overlapping camera views. It is commonly assumed that individuals do not change their clothing within the observation period, and that finer biometric cues (face, fingerprint, gait) are unavailable. We consider here the *appearance-based* re-id, *i.e.*, we exploit solely the appearance of the body. Re-id is an important problem: it has been the focus of intense research in the last five years, due to the distribution of challenging datasets (VIPeR [14], iLIDS Multi Camera Tracking Scenario [17], ETHZ [7]), but its roots lie farther back, in object model design for human tracking [25]. It is also pervasive, extending from the original video-surveillance field to the most recent photo-tagging domain [23].

BMVC 2011 http://dx.doi.org/10.5244/C.25.68

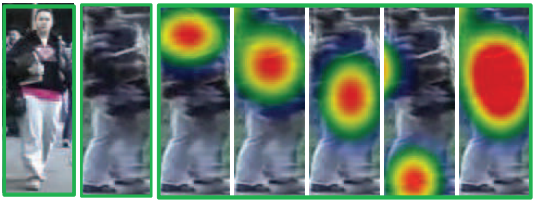

Figure 1: Re-id performed by a human subject: (a) the test probe we asked to match; (b) the correct match in the gallery; (c) the fixation heat maps from eye-tracking over consecutive 1s intervals - the hotter the color, the longer the time spent looking at that area.

In this paper, we present a novel methodology for human re-id, based on Pictorial Structures (PS) for human body pose estimation. PS essentially rely on two components: one capturing the local appearance of body parts, and the other representing an articulated body structure [9]. Inference in a PS involves finding the MAP spatial configuration of the parts, *i.e.*, the body pose. We build upon the PS framework of [1], where general part detectors localize the body parts, and a kinematic tree prior captures the structural knowledge.

Our proposal takes inspiration from how humans perform re-id, assuming they operate under the same hypotheses described at the beginning of this section. Taking a subset of the VIPeR dataset (45 pedestrians), we set up a simple re-id experiment where subjects were asked to match test probes to candidate galleries (5×2 cm each, pooled together in a single screen), while being monitored with an eye-tracker system (SMI Red 120Hz). This allowed us to obtain fixation heat maps, showing where the subjects concentrated their attention. As shown in Fig. 1 for a particular trial, the fixation maps indicate a tendency to scan salient parts of the body, looking for part-to-part correspondences. We think that encoding and exploiting the human appearance per parts is a convenient strategy for re-id, and PS are the best tool for this task. PS are usually fitted on individual images, as independent entities, and we exploit this setting for *single-shot* re-id, which consists in matching pairs of images, a probe and a gallery image for each subject. 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 Maximally Stable Color Regions (MSCR) [11], previously adopted for re-id in [8]. The features of each part are subsequently combined into a single ID signature. Matching between signatures is carried out by standard distance minimization strategies. On the other hand, *multi-shot* re-id occurs when each subject has multiple images, either in the gallery and/or the probe set, which can be exploited to accumulate more visual information and ensure higher re-id accuracy. In this case, we propose a strategy to improve the PS fitting on images of the same subject. This task has received little attention in the literature, like [21], where a large number of consecutive images per person was used, whereas a re-id task often provides only few (2-5) non-consecutive images. Our idea is to learn the local appearance of each part in a given subject so that ad-hoc appearance part detectors can provide more accurate PS fitting. The detectors we used are multidimensional Gaussian filters capturing the appearance of every pixel in each part. Moreover, chances of bad learning due to the scarcity of samples per person is mitigated by employing spatial reasoning, *i.e.*, by augmenting the statistics of a pixel with similar neighboring pixel values in a surrounding region, identified through non-parametric Mean Shift segmentation [5]. We have called this new model Custom Pictorial Structure (CPS). Once CPS is fitted on data, features are extracted from each instance as in the single-shot case,

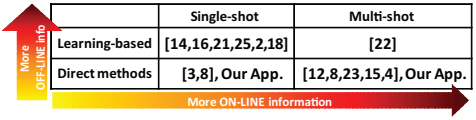

Figure 2: Taxonomy of re-identification methods.

and the individual signatures are pooled together to obtain a multi-shot ID signature. The matching policy between multi-shot signatures is based on finding, among the individual signatures of the compared individuals, the pair whose distance is minimal.

Summarizing, in this paper we propose:

- the adoption of PS for single-shot re-id, and appropriate per-part feature extractors;
- a novel fitting strategy for PS, specifically suited for few, non-consecutive images of the same subject, and a strategy for matching multi-shot signatures.

Experiments have been carried out on all the available re-id datasets (iLIDS, ETHZ1,2,3, VIPeR) with convincing results in all modalities and outright best performances in the multi-shot case. Moreover, we created a novel dataset starting from the CAVIAR tracking data[1], presenting a unique combination of challenges. We finally evaluate our method against the human capability of re-id on a subset of the VIPeR dataset.

The paper is organized as follows. Sec. 2 analyzes related work, which lies both in the PS and and in the re-id literature. Our approach is detailed in Sec. 3, and the related experiments are discussed in Sec. 4. Finally, Sec. 6 wraps up with remarks and future perspectives.

2 State of the art

Pictorial structures. The literature on PS is large and multifaceted. Here, we briefly review the studies that focus on the appearance modeling of body parts. We can distinguish two types of approaches: the single-image and multiple-image methods. In the former case, a PS processes each image individually. In [20], a two-step image parsing procedure is proposed, that enriches an edge-based model by adding chromatic information. In [6], a learning strategy estimates relations between body parts and a shared color-based appearance model is used to deal with occlusions. In the other case, several images *representing a single person* are available. Very few methods deal with this situation. In [21], two approaches for building PS have been proposed for tracking applications. A top-down approach automatically builds people models starting by convenient key poses detections; a bottom-up method groups together candidate body parts found along the considered sequence exploiting spatio-temporal reasoning. This technique shares some similarities with our approach, but it requires a high number of temporally consecutive frames (50-100). In our setting, few ($\leq$5), unordered images are instead expected. In a photo-tagging context, PS are grown over face detections to recognize few people [23], modeling the parts with Gaussian distributions in the color space.

Re-identification. A taxonomy of recent appearance-based techniques is shown in Fig. 2, displaying two groups of methods: *learning-based* and *direct* approaches. In the former, a dataset is split into training and test sets, with the training individuals used to learn features and/or strategies for combining features to achieve high re-id accuracy, and the test ones used as validation. Direct methods are instead pure feature extractors. An orthogonal classification separates the *single-shot* and the *multi-shot* techniques. As learning-based methods, an ensemble of discriminant localized features and classifiers is selected by boosting in [13]. In [16], pairwise dissimilarity profiles between individuals are learned and adapted for nearest-neighbor classification. Similarly, in [22], a high-dimensional signature formed by multiple

[1] http://www.re-identification.net/

features is projected onto a low-dimensional discriminant space by Partial Least Squares reduction. Contextual visual information is exploited in [26], enriching a bag-of-word-based descriptor by features derived from neighboring people, assuming that people stay together across different cameras. [2] casts re-id as a binary classification problem (one vs. all), while [19] as a relative ranking problem in a higher dimensional feature space where true and wrong matches become more separable.

As direct methods, a spatio-temporal local feature grouping and matching is proposed in [12]: a decomposable triangulated graph is built that captures the spatial distribution of the local descriptions over time. In [24], images are segmented into regions and their color spatial relationship acquired with co-occurrence matrices. In [15], interests points (SURF) are collected in subsequent frames and matched. Symmetry and asymmetry perceptual attributes are exploited in [8], based on the idea that features closer to the bodies' axes of symmetry are more robust against scene clutter. Covariance features, originally employed for pedestrian detection, are tailored in [3] for re-id, extracted from coarsely located body parts. In [4], epitomic analysis is used to collapse a set of images into a small collage of overlapped patches containing the essence of textural, shape and appearance properties. To be brief, in addition to color, a large number of features types is employed for re-id: textures [8, 13, 19, 22], edges [22], Haar-like features [2], interest points [12] and image regions [8, 13, 24]. The features, when not collected densely, can be extracted from horizontal stripes, triangulated graphs, concentric rings [26], symmetry-driven structures [8], and localized patches [3].

Our method lies in the class of the direct approaches, and can work in both single- and multi-shot modes. As Table 2 shows, our single-shot approach inhabits the poorest information corner of this taxonomy, but is still able to perform as well as his richer competitors.

3 The proposed approach

In general, a simple procedure implements our method: we localize body parts using PS; we extract visual information from them to create an ID signature; finally, signatures of probes and gallery images are matched and evaluated. In the following, we summarize the basic PS technique of [1] and describe the single- and multi-shot working modalities of our method.

3.1 Pictorial structures basics

In PS, the body model is decomposed into a set of parts whose configuration is denoted as $L = \{\mathbf{l}_p\}_{p=1}^{N}$, where $\mathbf{l}_p = (x_p, y_p, \vartheta_p, s_p)$ encodes position, orientation and scale of part p, respectively. Given the image evidence D, the posterior of L is modeled as $p(L|D) \propto p(D|L)p(L)$, where $p(D|L)$ is the image likelihood and $p(L)$ is a prior modeling the parts connectivity. The kinematic dependencies between body parts are mapped onto a directed acyclic graph (DAG) with edges E, giving $p(L) = p(\mathbf{l}_1)\prod_{(i,j)\in E} p(\mathbf{l}_i|\mathbf{l}_j)$, where $\mathbf{l}_1$ is the root node (the torso), and $p(\mathbf{l}_i|\mathbf{l}_j)$ models the joint between two connected parts. Meanwhile, the image evidence D is obtained with discriminatively trained part models, each providing an evidence map $\mathbf{d}_p$. Assuming that all $\{\mathbf{d}_p\}_{p=1}^{N}$ are conditionally independent given the configuration L, and each part depends only on its own configuration, we factorize the likelihood in $p(D|L) = \prod_{p=1}^{N} p(\mathbf{d}_p|\mathbf{l}_p)$. Thus, the posterior over the configuration L becomes:

$$p(L|D) \propto p(\mathbf{l}_1)\prod_{p=1}^{N} p(\mathbf{d}_p|\mathbf{l}_p)\prod_{(i,j)\in E} p(\mathbf{l}_i|\mathbf{l}_j). \quad (1)$$

The model is trained on a dataset of annotated images, independently for the part detectors and the kinematic prior. In all the following experiments, the learned PS is exclusively used to efficiently infer the posterior on new images. For further details, see [1].

3.2 The single-shot modality

In this modality, every image is treated individually. Fitting a PS on a pedestrian image follows the original framework of [1], where general detectors are used to locate the body parts. In the following, we choose a body configuration composed by $N=6$ parts (chest, head, thigs and legs, as in Fig. 3(a)), which we found out detailed enough to represent upright pedestrians in frontal/back or side views. See more details in Sec. 4. After fitting the PS, we consider 1) the per-part chromatic content, and 2) the color displacement within the body mask. We treat the former aspect by calculating color histograms of each part independently, and then concatenating and normalizing them into a single feature vector for each image. In particular, we use a modified HSV characterization [18], where hue and saturation are jointly taken by a 2D histogram to retain much of the chromatic specificity, and brightness is counted separately. Also, there is a distinct count for the full black color, to eliminate ill-defined hue and saturation values at low brightness. Moreover, because parts have different sizes (*e.g.*, the torso is about three times larger than the head), we multiply the part histograms with a set of weights $\{w_p\}_{p=1}^{N}$ before the concatenation and normalization. In this way, we ensure the ability to tune the algorithm to both different size of parts and their importance for the re-id task. Fine-tuning these parameters can be easily performed through cross-validation, and this substantially improves the performance since it allows to adapt to different visual conditions (color saturation, image brightness, body sizes) in the different datasets considered.

Concerning the per-region color displacement, we use the Maximally Stable Color Region (MSCR)[2] operator [11], which detects a set of blob regions by looking at successive steps of an agglomerative clustering of image pixels. Each step groups neighboring pixels with similar color within a threshold that represents the maximal chromatic distance between colors. Those maximal regions that are stable over a range of steps become MSCRs. As experimentally validated in [8], height and color of these regions are features particularly suited for re-id. In our approach, we extract the MSCR blobs from within the PS body mask.

The color histograms and the MSCRs ultimately form our desired ID signature. Matching two signatures $I_A = (\mathrm{H}_A, \mathrm{MSCR}_A)$ and $I_B = (\mathrm{H}_B, \mathrm{MSCR}_B)$ is carried out by calculating the distance d as:

$$d(I_A, I_B) = \beta_{\mathrm{H}} \cdot d_{\mathrm{H}}(\mathrm{H}_A, \mathrm{H}_B) + \tag{2}$$

$$(1-\beta_{\mathrm{H}}) \cdot d_{\mathrm{MSCR}}(\mathrm{MSCR}_A, \mathrm{MSCR}_B), \tag{3}$$

where β_{H} is a calibration parameter, d_{H} is the Bhattacharyya distance, and for d_{MSCR} we employed the same distance between MSCRs as in [8].

3.3 The multi-shot modality

With multiple images for each subject, each part occurs more times, and we can exploit this variety to improve the PS fitting. In this case, our goal is to learn a local model that can capture the appearance of the subjects' parts, and alter the MAP body configuration estimates so that the subject is consistently localized across all images, increasing the re-id accuracy. These ad-hoc part detectors will act jointly with the general-person (GP) detectors of [1]. This Custom PS (CPS) is a two-step iterative process that alternates between estimating the body configuration and updating the appearance model. Let us assume we are processing T images of a given subject (Fig. 3(b)): at the first iteration, since we provide an initial uninformative appearance model, the MAP body estimates will coincide with those of PS, which is driven by the GP detectors and the kinematic prior alone. We can now collect all

[2] http://www2.cvl.isy.liu.se/~perfo/software/

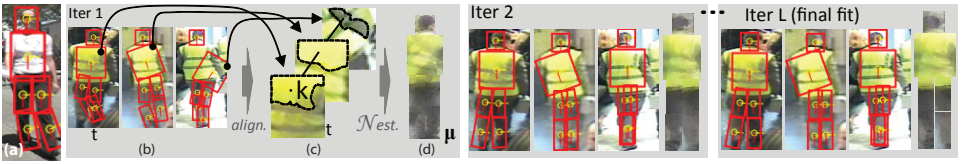

Figure 3: Our approach. Single-shot PS in (a). Multi-shot CPS at iteration 1: (b) initial PS fitting; (c) the parts are aligned and per-pixel statistics is collected employing spatio-temporal reasoning; (d) the ad-hoc part detectors are estimated, whose means μ are shown. At every iteration until L, the fitting becomes more accurate due to the improving part detectors.

the parts, remove the transformations, and estimate a Gaussian distribution $\mathcal{N}(\mu_k, \sigma_k)$ for all pixels k. Whenever T is small, the resulting Gaussian estimates would be poor, very sensitive to noise. In order to reinforce the statistics, for each pixel location k, we increase the samples by including spatial neighbors of similar color (Fig. 3(c)) by performing Mean Shift segmentation [5] on each subimage t and including the neighbors of k that belong to the same segment. The resulting Gaussian distribution is more robust to noise (Fig. 3(d)).

We now use such robust models to provide evidence maps for each part. By filtering the image with the Gaussian distributions (efficiently with FFTs), we get likelihood scores for each possible combination of scale, rotation and position of the parts in the image. At the next PS step, we fuse the GP and Gaussian scores to provide parts localizations. Among the many rules for fusion, we found out that selecting the max (per-pixel) gave us good results.

We are aware that employing a single Gaussian works only for rigid parts: it is not a problem for arms and legs, because clothes often have uniformly colored pants and sleeves. For the torso and the head, this assumption certainly does not hold as front and back of the head are very different, like front and back of the torso in the presence of a backpack. Nevertheless, our approach improves the PS fitting and re-id accuracy rates. Upgrading to mixture of Gaussians designed through model selection strategies will be surely a future improvement of CPS. Experimentally, CPS converges after 4-5 iterations, and we can finally extract ID signatures like in the single-shot case. As for the matching, when we compare M probe signatures of a given subject against N gallery signatures of another one, we simply calculate all the possible $M \times N$ single-shot distances, and keep the smallest one.

4 Results

We tested our approach on seven datasets: iLIDS, ETHZ1,2, and 3, VIPeR, VIPeR for humans (VIPeRHuman), and CAVIAR for re-id (CAVIAR4REID). The first five are all the publicly available benchmarks for re-id, and we additionally promote two new ones: the VIPeRHuman provides a standard against humans' ability for re-id, and the CAVIAR4REID shows several peculiar conditions found separately in the other datasets. This section is organized in three parts. The first illustrates the common algorithm setup we used in all experiments, the second gives deep insight into our approach on iLIDS, and the third shows our results on the remaining datasets, Re-id performance is reported in terms of the recognition rate, via the cumulative matching characteristic (CMC) curve which represents the expectation of finding the correct match in the top n matches. Also, a quantitative scalar appraisal of our curves is given by the normalized area under curve (nAUC) value.

4.1 Algorithm setup

The setup of [1] for pedestrian detection does not work well in our scenes. The immediate differences are in framing and pose: we have images tightly bound around mostly centered

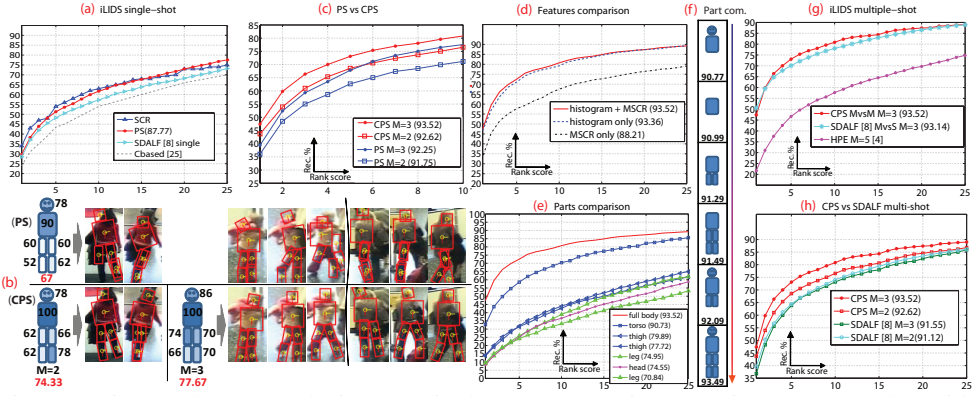

Figure 4: iLIDS dataset analysis. (a) Single-shot comparison (PS is our approach), with nAUC values within parentheses. (b) Multi-shot case qualitative and quantitative comparison between PS and CPS fittings, with the puppet showing the accuracy rate of each part and the global detection accuracy (below the puppet). Some images with the PS or CPS superimposed (with M=2 and M=3, the row below). (c) PS in multi-shot vs. CPS. (d) CMC curves for the separate features in CPS. (e) CMC curves for separate parts in CPS. (f) nAUC scores obtained by using ensembles of parts attached to the torso. (g) Multi-shot comparison vs other approaches. (h) CPS vs SDALF in a multi-shot perspective. See text.

human figures, viewed either in frontal, back or side views. After having tested the full body configuration of 10 parts, we settled on a reduced body setting (see Fig.3(a)), dropping upper and lower arms. In fact, we found out that these parts were often misplaced, due to either the framing, the small size images (as low as 17×39), the severe noise, the illumination changes or the (self-)occlusions.

4.2 Exploratory analysis on iLIDS

The iLIDS MCTS videos have been captured at a busy airport arrival hall [26]: the dataset consists of 119 pedestrians with 479 images normalized to 64×128 pixels. The images come from non-overlapping cameras, subject to quite large illumination changes and occlusions. On average, each individual has 4 images. The best single-shot performance is obtained by a covariance-based technique (SCR)[3], while the best multi-shot modality by SDALF [8].

We first consider the single-shot case, reproducing the same experimental settings of [8, 26]. We randomly select one image for each pedestrian to build the gallery set, while the rest forms the probe set; then, the matching between each probe and the gallery set is estimated. This procedure is repeated 10 times, and the average CMC is displayed. Fig. 4(a) shows that PS performs slightly worse than SCR, but better than SDALF and the Context-based strategy (Cbased) of [26], which is learning-based.

As for the multi-shot case, in order to genuinely evaluate the proposed CPS, we refer to a multi-vs-multi matching policy introduced in [8], where both probe and gallery sets have groups of M images per individual and the distance between two given groups can be taken to be the shortest among all pairs of images. For the sake of clarity, we want to point out that the single-shot modality introduced in [26] and the multi-vs-multi case presented in [8] are not directly comparable. In fact, let us assume we have a total of N images for an individual. In the single-shot case, one image is randomly chosen as gallery sample and the other $N-1$ are chosen as probes. In the multi-shot case, M images are taken as gallery and other M as probe. Now, the single-shot evaluation of a native multi-image dataset uses all probes

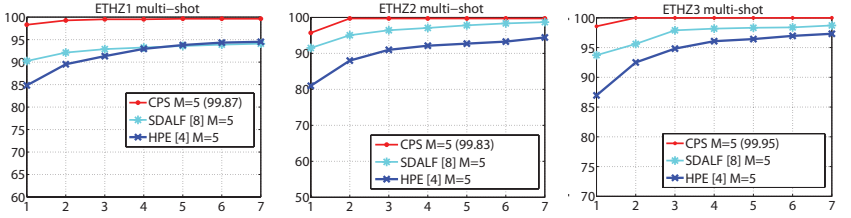

Figure 5: Multi-shot results on the ETHZ 1,2,3 datasets.

to retrieve the CMC curve, focusing on their collective behavior, while multi-shot will deal with a smaller subset of images in a setting that is more realistic for a live re-id system.

Having clarified the terms of comparisons, we first evaluated the improvement in the localization of the parts (Fig.4(b)). To give quantitative figures, we randomly extracted 50 images from the dataset and manually annotated the parts, calculating the fitting accuracy with the protocol of [10]: PS has 67% of accuracy, while CPS raises to 74.33% (M=2) and 77.67% (M=3). Qualitatively, we show a few examples of MAP body estimates.

Second, we compared the difference between multi-shot PS and CPS. With PS, we collected M individual signatures and used the same matching policy as CPS. In Fig. 4(c), we show CMC curves (first ten ranks) for M=2 and M=3.

We repeat this procedure 10 times, with different random gallery/probe partitions. CPS performance is 7.8% higher than PS in average for the first rank. Note that CPS with M=2 is superior to PS with M=3.

Next, we analyze details of our best performance: multi-shot CPS M=3. Regarding the features employed, from Fig. 4(d) it is apparent that the color histograms do the main job, while MSCRs help refine the result. As for the relative importance of the body parts, Fig. 4(e) clearly evidences the major role of the torso, being the larger and more central area of the body. If we next look at how ensembles of parts together with the torso behave, Fig. 4(f) confirms that head and lower legs sometimes bring misleading information, being the smaller and more peripheral areas, but this is highly dependent on the dataset (iLIDS has a lot of lower image occlusions). In Fig. 4(g), we compare multi-shot results. As competitors, we consider the best performances of SDALF (obtained in the Multi-vs-Single modality N=3, where galleries had three signatures and probes had a single one), and HPE [4] multi-vs-multi with M=5. We get the highest nAUC score, even if at the first rank SDALF does slightly better (50.25% vs 47.39%). In order to clearly compare CPS vs SDALF efficacy, we carried out a final experiment with the multi-vs-multi modality which shows the superiority of CPS in exploiting multiple instances per person, as depicted in Fig. 4(h).

4.3 Other comparative results

ETZH Dataset [7]. Three video sequences have been captured with moving cameras at head height, originally intended for pedestrian detection. In [22], samples have been taken for re-id[3], generating three variable size image sets with 83 (4.857 images), 35 (1.936 images) and 28 (1.762 images) pedestrians, respectively. All images have been resized to 32×64 pixels. The challenging aspects of ETHZ are illumination changes and occlusions, and while the moving camera provides a good range of variations in people's appearances, the poses are rather few. SDALF multi-vs-multi M=5 is the best in the literature. Our multi-shot CPS M=5 retrieves a near 100% nAUC (Fig. 5). It is worth noting that this dataset does not replicate a genuine videosurveillance setup, because of the head-height moving camera.

[3] http://www.umiacs.umd.edu/~schwartz/datasets.html

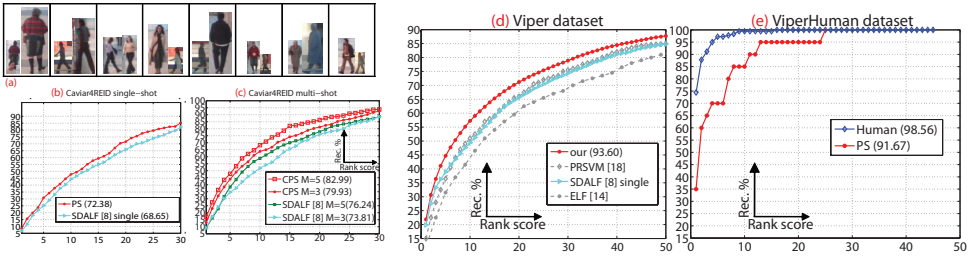

Figure 6: CAVIAR4REID results on the left: (a) some images of the dataset, at their natural relative proportions, each pair portraying the same individual; (b) single-shot results and (c) multi-shot results with probes and galleries from different cameras. VIPeR and VIPeRHuman on the right: (d) comparative results on VIPeR (gray-dotted lines are methods based on a learning stage), and (e) human capabilities against our single-shot PS approach.

CAVIAR for re-id Dataset. CAVIAR4REID[4] has been extracted from the CAVIAR database, in particular the recordings from two different cameras in an indoor shopping center in Lisbon. The pedestrians images have been cropped using the provided ground truth. Of the 72 different individuals identified (with images varying from 17×39 to 72×144), 50 are captured by both views and 22 from only one camera. For each pedestrian, we selected 10 images from 2 different camera views maximizing the variance with respect to resolution changes, light conditions, occlusions, and pose changes (see samples in Fig. 6).

We set up this dataset to merge together video surveillance challenges like the wide range of poses and real surveillance footage in iLIDS, and the multiple images and wide range of resolutions of ETHZ. To take full advantage of these conditions, we decided to take probe and gallery images from different cameras, one image each for single-shot, M for multi-shot.

Our approach outperforms SDALF in both modes (Fig. 6): in (b), single-shot comparison on 72 individuals, and, in (c), multi-shot on 50 individuals (M=3,5).

VIPeR Dataset [14]. This dataset[5] contains two views of 632 pedestrians, each pair of views made up of 48×128 images taken from different cameras, under different viewpoint, pose and light conditions. It is the most challenging dataset currently available for single-shot re-id. In the literature, results on VIPeR are typically produced by mediating over ten runs, each consisting in a partition of 316 pairs, with all probes matched against all galleries. The best performance so far on this dataset is obtained by PRSVM [19] with very similar results by SDALF. In Fig. 6 (d), we show the comparative CMC curves. Our approach outperforms the other methods, setting the rank-1 matching rate at 21.84%, and rank 10 at 57.21%.

VIPeR for humans Dataset. Intuitively, the best visual recognizers are people. In order to compare human and algorithmic capabilities, we extracted 45 image pairs from VIPeR, focusing on "difficult" samples with strong pose changes and similar clothes. Then, we built a web application where a test subject is presented a random probe image and asked to perform re-id by choosing the most plausible correspondence among the 45 gallery candidates. In case the subject succeeds, the trial presents the next samples until the end is reached, otherwise the system keeps on asking for the true correspondence. In this way, with multiple trials a "human" CMC is built. The same trials are given to our algorithm, and a "machine" CMC is built. 180 trials were performed, by 18 people. In Fig. 6 (e), we confirm that human capabilities are far beyond the current best techniques. In particular, we ask the subjects

[4] Available at http://www.re-identification.net/
[5] http://vision.soe.ucsc.edu/?q=node/178

through questionnaires about the cues they employed. The color of the parts is the primary cue. The second is the "type" of clothing worn (jacket or shirt, t-shirt and so on). If the situation is ambiguous, the gender may help, as may the presence of discriminant particulars (logos on the shirt). Part of the experiments were performed with an eye tracker, producing the results discussed in the introduction, and visualized in Fig. 1.

5 Conclusions

Exploring the use of PS for re-id is a promising brand new research direction, because it allows us to finely localize human body parts that provide distinctive features for composing and matching ID signatures for individuals. Experiments suggest that improving the localization enhances the re-id performance. Therefore, we proposed a strategy to improve the PS fitting by modeling the common appearance of a given individual within multiple images, forging the Custom Pictorial Structure (CPS). Overall, our approach improves on previous state-of-the-art results in multi-shot re-id, as well as doing very well in single-shot. The possible enhancements are many, from a better (multi-modal) appearance modeling in CPS, to the adoption of learning strategies to further improve the discriminative power.

6 Acknowledgments

Research funded by the EU-Project FP7 SAMURAI, grant FP7-SEC-2007-01 No. 217899.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1014–1021, Miami, USA, June 2009. URL http://www.d2.mpi-inf.mpg.de/node/381.

[2] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person Re-identification Using Haar-based and DCD-based Signature. In *2nd Workshop on Activity Monitoring by Multi-Camera Surveillance Systems (AMMCSS), in conjunction with 7th IEEE Intl. Conf. on Advanced Video and Signal-Based Surveillance (AVSS)*, August 2010.

[3] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person Re-identification Using Spatial Covariance Regions of Human Body Parts. In *7th IEEE Intl. Conf. on Advanced Video and Signal-Based Surveillance (AVSS)*, August 2010.

[4] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-Shot Person Re-identification by HPE Signature. In *Intl. Conf. on Pattern Recognition (ICPR)*, pages 1413–1416, August 2010.

[5] D. Comaniciu and P. Meer. Mean Shift: A Robust Approach Toward Feature Space Analysis. *IEEE TPAMI*, 24(5):603–619, 2002.

[6] M. Eichner and V. Ferrari. Better Appearance Models for Pictorial Structures. In *British Machine Vision Conference (BMVC)*, September 2009.

[7] A. Ess, B. Leibe, and L. V. Gool. Depth and Appearance for mobile scene analysis. In *IEEE Intl. Conf. on Computer Vision (ICCV)*, 2007.

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person Re-Identification by Symmetry-Driven Accumulation of Local Features. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[9] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial Structures for Object Recognition. *Intl. J. on Computer Vision*, 61(1):55–79, 2005.

[10] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *IEEE Conf. on CVPR*, 2008.

[11] Per-Erik Forssén. Maximally Stable Colour Regions for Recognition and Matching. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2007.

[12] N. Gheissari, T. B. Sebastian, P. H. Tu, , J. Rittscher, and R. Hartley. Person Reidentification Using SpatioTemporal Appearance. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1528–1535, 2006.

[13] D. Gray and H. Tao. Viewpoint Invariant Pedestrian Recognition with an Ensamble of Localized Features. In *Euro. Conf. on Computer Vision (ECCV)*, pages 262–275, 2008.

[14] D. Gray, S. Brennan, and H. Tao. Evaluating Appearance Models for Recongnition, Reacquisition and Tracking. In *IEEE Intl. Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, October 2007.

[15] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *ACM/IEEE Intl. Conf. on Distributed Smart Cameras (ICDSC)*, pages 1–6, September 2008.

[16] Z. Lin and L.S. Davis. Learning Pairwise Dissimilarity Profiles for Appearance Recognition in Visual Surveillance. In *4th Intl. Symp. on Adv. in Visual Computing*, 2008.

[17] UK Home Office. i-LIDS multiple camera tracking scenario definition, 2008.

[18] K. Okuma, A. Taleghani, N. De Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In *Euro. Conf. on Computer Vision*, 2004.

[19] B. Prosser, W. Zheng, S. Gong, and T. Xiang. Person Re-Identification by Support Vector Ranking. In *British Machine Vision Conference (BMVC)*, 2010.

[20] D. Ramanan. Learning to parse images of articulated bodies. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1129–1136. 2007.

[21] D. Ramanan, D. A. Forsyth, and A. Zisserman. Tracking People by Learning Their Appearance. *IEEE TPAMI*, 29(1):65–81, 2007.

[22] W.R. Schwartz and L.S. Davis. Learning discriminative appearance-based models using partial least squares. In *XXII SIBGRAPI 2009*, 2009.

[23] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding People in Repeated Shots of the Same Scene. In *British Machine Vision Conference (BMVC)*, 2006.

[24] X. Wang, G. Doretto, T. B. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In *IEEE Intl. Conf. on Computer Vision (ICCV)*, pages 1–8, 2007.

[25] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 2006.

[26] W. Zheng, S. Gong, and T. Xiang. Associating Groups of People. In *British Machine Vision Conference (BMVC)*, 2009.