Person Re-identification by Articulated Appearance Matching

Dong Seon Cheng and Marco Cristani

Abstract Re-identification of pedestrians in video-surveillance settings can be effectively approached by treating each human figure as an articulated body, whose pose is estimated through the framework of Pictorial Structures (PS). In this way, we can focus selectively on similarities between the appearance of body parts to recognize a previously seen individual. In fact, this strategy resembles what humans employ to solve the same task in the absence of facial details or other reliable biometric information. Based on these insights, we show how to perform single image re-identification by matching signatures coming from articulated appearances, and how to strengthen this process in multi-shot re-identification by using Custom Pictorial Structures (CPS) to produce improved body localizations and appearance signatures. Moreover, we provide a complete and detailed breakdown of the system that surrounds these core procedures, with several novel arrangements devised for efficiency and flexibility. Finally, we test our approach on several public benchmarks, obtaining convincing results.

1 Introduction

Human re-identification (re-id) consists in recognizing a person in different locations over various non-overlapping camera views. We adopt the common assumption that individuals do not change their clothing within the observation period, and that finer biometric cues (face, fingerprint, gait, etc..) are unavailable: we consider, that is, only *appearance-based* re-id.

In this paper, we present an extensive methodology for person re-id through articulated appearance matching, based on Pictorial Structures (PS) [17], and its variant

Dong Seon Cheng
Dept. of Computer Science & Engineering, HUFS, Korea, e-mail: cheng_ds@hufs.ac.kr

Marco Cristani
Dip. di Informatica, University of Verona, Italy, e-mail: marco.cristani@univr.it

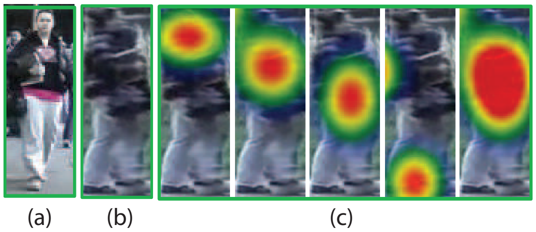


Fig. 1: Re-id performed by a human subject: (a) the test probe, (b) the correct match in the gallery, and (c) the fixation heat maps from eye-tracking over consecutive 1s intervals - the hotter the color, the longer the time spent looking at that area.

Custom Pictorial Structures (CPS) [9], to decompose the human appearance into body parts for pose estimation and signature matching. In the PS framework of [1], the parts are initially located by general part detectors, and then a full body pose is inferred by solving their kinematic constraints. In this work, we propose a novel type of part detector, fast to train and to use, based on the *histogram of oriented gradients* (HOG) [10] features and a *linear discriminant analysis* (LDA) [25] classifier. Moreover, we use the *belief propagation* algorithm to infer MAP body configurations from the kinematic constraints, represented as a tree-shaped factor graph.

More in general, our proposal takes inspiration from how humans approach appearance-based re-id. As we showed in [9], monitoring subjects performing re-id confirmed a tendency to scan for salient (structurally known) parts of the body, looking for part-to-part correspondences (we reproduce a sample of the study in Fig. 1). We think that encoding and exploiting the human appearance per parts is a convenient strategy for re-id, and PS is particularly well suited to this task. In particular, we exploit the conventional PS fitting on separate individual images for *single-shot* re-id, which consists in matching pairs probe/gallery of images for each subject. Our approach aims at obtaining robust signatures from features extracted from the segmented parts.

Secondly, for *multi-shot* re-id, where each subject has multiple images distributed between probe set and gallery set, we can use the extra information to improve the re-id process in two ways: by improving the PS fitting using the CPS algorithm [9] that iteratively performs appearance modeling and pose estimation, and by using *set-matching* to compute distances between probe set and gallery set. The rationale of CPS is that the local appearance of each part should be relatively consistent among images of the same subject, and hence it is possible to build an appearance model. Thus, localizing parts can be enhanced by evaluating the similarity to the model.

Our goal in this work is to crystallize the use of PS for re-id with a complete and detailed breakdown of the stages in our process. We intend to introduce several novel arrangements devised for efficiency and flexibility, with an eye towards future extensions. In particular, we introduce a new class of part detectors based on HOG features and linear discriminant analysis to feed the PS pose estimation algorithm, and a new color histogram technique to extract feature vectors. Experiments have been carried out on many publicly available datasets (iLIDS, ETHZ1,2,3, VIPeR, CAVIAR4REID) with convincing results in all modalities,

The chapter is organized as follows: we analyze related work in Sec. 2; we provide an overview of our approach in Sec. 3 and all the detail in Sec. 4; we provide details about the training of our part detectors in Sec. 5; and we discuss the experiments in Sec. 6. Finally, Sec. 7 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 [30], a two-step image parsing procedure is proposed, that enriches an edge-based model by adding chromatic information. In [12], 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 [31], 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 [36], modeling the parts with Gaussian distributions in the color space. ADDITIONAL REQUESTED SOTA, DPM[15].

Re-identification: Appearance-based techniques for re-identification can be organized in 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 [23]. In [26], pairwise dissimilarity profiles between individuals are learned and adapted for nearest-neighbor classification. Similarly, in [34], a high-dimensional signature formed by multiple features is projected onto a low-dimensional discriminant space by Partial Least Squares reduction. Contextual visual information is exploited in [40], enriching a bag-of-word-based descriptor by features derived from neighboring people, assuming that people stay together across different cameras. [3] casts re-id as a binary classification problem (one vs. all), while [29, 41] as a relative ranking problem in a higher dimensional feature space where true and wrong matches become more separable. In [18], re-identification is cast as a semi-supervised single-shot recognition prob-

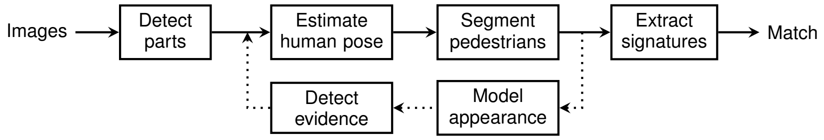


Fig. 2: Diagram of the stages in our approach. In single-shot mode, the estimated pose of the articulated human figure is used to segment the image and extract the features, joined into a signature. In multi-shot mode, with multiple images for each pedestrian, we model the common appearance of its parts, and thus refine the part detections with additional evidence, to be able to improve the pose estimation.

lem where multiple features are fused at the classification output level, using the multi-view learning approach of [27]. Finally, re-identification is cast as a Multiple Instance Learning in [33], where in addition a method for synthetically augmenting the training dataset is presented.

As direct methods, a spatio-temporal local feature grouping and matching is proposed in [21]: a decomposable triangulated graph is built that captures the spatial distribution of the local descriptions over time. In [37], images are segmented into regions and their color spatial relationship acquired with co-occurrence matrices. In [24], interests points (SURF) are collected in subsequent frames and matched. Symmetry and asymmetry perceptual attributes are exploited in [14, 7], 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 [4] for re-id, extracted from coarsely located body parts; later on, such descriptors are embedded into a learning framework in [2] In [8], 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 [23, 34, 14, 29], edges [34], Haar-like features [3], interest points [21] and image regions [23, 37, 14]. The features, when not collected densely, can be extracted from horizontal stripes, triangulated graphs, concentric rings [40], symmetry-driven structures [14, 7], and localized patches [4]. Very recently, depth-based methods include into the analysis other modalities and sensors (such as RGB-D cameras) to extract 3D soft-biometric cues from depth images: this will avoid the constraint that people must be dressed in the same way during a re-identification session [6]. Another unconventional application of re-identification considers Pan-Tilt-Zoom cameras, where distances between signatures are also computed across different scales [32].

For an extensive review on the re/identification methods, please see [11].

Our method lies in the class of the direct approaches, and can work in both single- and multi-shot modes.

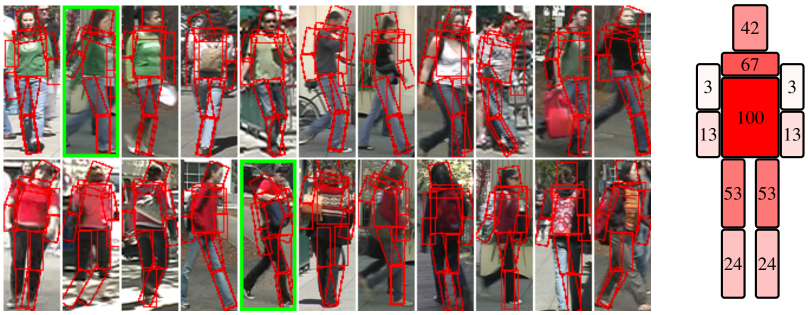


Fig. 3: (Left) Two illustrative lineups in single-shot re-identification from the VIPeR experiments: the leftmost image is the probe and the rest are gallery images sorted by increasing distance from the probe. The correct match is shown with a green outline. (Right) Model of the articulated human figure, with percentages and color intensities proportional to the importance of a part in the VIPeR experiment.

3 Overview of our approach

This section gives an overview of our re-identification process, which is summarized in Fig. 2. Implementation details of each stage can be found later, in Section 4. The method is based on obtaining good pedestrian segmentations from which effective re-identification signatures can be extracted. The basic idea is that we can segment accurately after we estimate the pose of the human figure within each image, and this pose estimation can be performed with Pictorial Structures.

The single-shot modality

Every image is processed individually to retrieve a feature vector that acts as its signature. By calculating distances between signatures, we can match a given probe image against a set of gallery images, ranking them from lowest to highest distance, and declaring the rank-1 gallery to be our guess for the identity of the probe.

Our proposed approach tries to increase the effectiveness of the signatures by filtering out as much of the background scene as possible, and by decomposing a full pedestrian figure into semantically reasonable body parts (like head, torso, arms and legs) in such a way that we can compose a full signature by joining part signatures. This increases the robustness of the method to partial (self)occlusions and changes in local appearance, like the presence of bags, different looks between frontal, back and side views, and imperfections in the pose estimation. Fig. 3 (left) shows two cases from the VIPeR experiment, illustrating several aspects of the problems just mentioned. It is clear that the segmentations provide a good filtering of the background scene, even when they do not perfectly isolate the pedestrian figure.

However, the decomposition into parts is not sufficient to overcome persistent dataset-wise occlusions or poor image resolution. For example, the iLIDS dataset is made up of images taken from airport cameras, and an overwhelming number of pedestrians are captured with several bags, backpacks, trolleys and other occluding objects (including other different pedestrians). In this challenging situation, legs and arms are often hidden and their discriminating power is greatly reduced. Therefore, our approach is to balance the contributions of each part through a weight that indicates, percentage wise, its importance with respect to the torso, which remains the main predictor. Fig. 3 (right) shows the weights for the VIPeR experiment.

The multi-shot modality

Multi-shot re-identification is performed when probe and gallery sets are made of multiple images for each subject. We can exploit this situation in two ways: firstly, by using *set matching* (the minimal distance across all pairs) when comparing signatures, so that the most unlike matches are discarded; secondly, by improving the pose estimations based on the appearance evidence. We create this evidence by building an appearance model of each pedestrian and using it to localize his parts with greater accuracy than just by using the generalized part detectors. Then, we feed this information back into the PS algorithm to compute new pose estimations, and hence segmentations. This process can be repeated until we reach a satisfactory situation.

In the end, our goal is to reinforce a coherent image of pedestrians, such that we can compute more robust signatures. Then, with multiple signatures available, the most natural way to match a probe set to the gallery sets is to find the closest pairs: this potentially matches frontal views with frontal views, side views with side views, occluding bags with occluding bags, and so on.

4 Details of our approach

We now give a detailed description of the stages in our re-identification approach, with a critical review of our previous method [9], where we adapted Andriluka's publicly available Pictorial Structures code to perform articulated pose estimation. Here instead, we developed a new and completely independent system with a novel part detector and our own implementation of the Pictorial Structures algorithm.

4.1 Part Detection

In [1], the authors use discriminatively trained part detectors to feed their articulated pose estimation process. In particular, their part detectors densely sample a shape context descriptor that captures the distribution of locally normalized gradient orientations in a log-polar histogram. With 12 bins for the location and 8 bins for the gradient orientation, they obtain 96 dimensional descriptors. Then, they concatenate the histograms of all shape context descriptors falling inside the bounding box of a part. During detection, many positions, scales, and orientations of parts are scanned in a sliding window fashion. All color images are converted to gray-scale before feature extraction.

To classify the feature vectors, they train an AdaBoost classifier [20] using as weak learners simple decision stumps that test histogram bins against a threshold. More formally, given a feature vector $\mathbf{x}$, there are $t = 1, \ldots, T$ stump functions $h_t(\mathbf{x}) = \mathrm{sign}(\xi_t(x_{n(t)} - \varphi_t))$, where φ_t is a threshold, ξ_t is a label equal to ±1, and $n(t)$ is the index of the bin chosen by the stump. Training the AdaBoost classifier

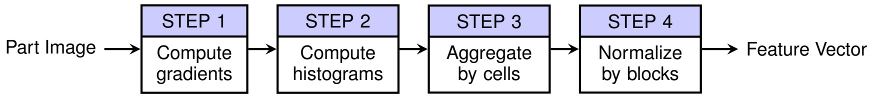


Fig. 4: Overview of the HOG feature extraction.

results in a strong classifier $H_i(\mathbf{x}) = \text{sign}(\sum_t \alpha_{i,t} h_t(\mathbf{x}))$ for each part i, where $\alpha_{i,t}$ are the learned weights of the weak classifiers.

During training, each annotated part is scaled and rotated to a canonical pose prior to learning, and the same process is applied during testing of candidate parts. The negative feature vectors come from sampling the image regions outside the objects, and the classifiers are then re-trained with a new training set augmented with false positives from the initial round. The classifier outputs are then converted into pseudo-probabilities by interpreting the normalized classifier margin as follows:

$$f_i(\mathbf{x}) = \frac{\sum_t \alpha_{i,t} h_t(\mathbf{x})}{\sum_t \alpha_{i,t}} \tag{1}$$

$$\tilde{p}(\mathbf{d}_i|\mathbf{l}_i) = \max(f_i(\mathbf{x}(\mathbf{l}_i)), \varepsilon_0), \tag{2}$$

where $\mathbf{x}(\mathbf{l}_i)$ is the feature vector for part configuration $\mathbf{l}_i$, and $\varepsilon_0 = 10^{-4}$ is a cutoff threshold. Even if the authors claim it works well, this simple conversion formula in fact produces poorly calibrated probabilities, as it is known that AdaBoost with decision stumps sacrifices the margin of the easier cases to obtain larger margins on cases close to the decision surface [35]. Our experience suggests that it produces weak and sparse candidate part configurations, because the decision boundary is assigned probability zero (not 0.5 as you would expect) and the weak margins (none of which approach 1) are linearly mapped to probabilities. A better choice would be to calibrate the predictions using Platt scaling [28].

4.1.1 The HOG-LDA detector

Histograms of oriented gradients (HOG) features for pedestrian detection were first introduced by Dalal and Triggs in [10]. They proved to be efficient and effective for object detection, not only pedestrians, both as wholes and as collection of parts [39]. The HOG features are usually combined with a linear SVM classifier, but [25] shows that an opportunely trained *linear discriminant analysis* (LDA) classifier can be competitive while being faster, and easier, to train and test

Calculating the HOG features requires a series of steps, shown summarized in Fig. 4. At each step, Dalal and Triggs experimentally show that certain choices produce better results than others, and they call the resultant procedure *the default detector* (HOG-dd). Like other recent implementations [16], we largely operate the same choices, but also introduce some tweaks.

STEP 1. Here, we assume the input is an image window of canonical size for the body part we are considering. Like in HOG-dd, we directly compute the gradients with the masks $[-1, 0, 1]$. For color images, each RGB color channel is processed separately, and pixels assume the gradient vector with the largest norm. While it does not take full advantage of the color information, it is better than discarding it like in the Andriluka's detector.

STEP 2. Next, we turn each pixel gradient vector into an histogram by quantizing its orientation into 18 bins. The orientation bins are evenly spaced over the range $0^\circ - 180^\circ$ so each bin spans 10°. For pedestrians there is no a-priori light/dark scheme between foreground and background (due to clothes and scenes) that justifies the use of the "signed" gradients with range $0^\circ - 360^\circ$: in other words, we use the contrast insensitive version [16]. To reduce aliasing, when an angle does not fall squarely in the middle of a bin, its gradient magnitude is split linearly between the neighboring bin centers. The outcome can be seen as a sparse image with 18 channels, which is further processed by applying a spatial convolution, to spread the votes to 4 neighboring pixels [38].

STEP 3. We then spatially aggregate the histograms into cells made by 7×7 pixel regions, by defining the feature vector at a cell to be the sum of its pixel-level histograms.

STEP 4. As in the HOG-dd, we group cells into larger blocks and contrast normalize each block separately. In particular, we concatenate features from 2×2 contiguous cells into a vector $\mathbf{v}$, then normalize it as $\tilde{\mathbf{v}} = \min(\mathbf{v}/||\mathbf{v}||, 0.2)$, L2 norm followed by clipping. This produces 36-dimensional feature vectors for each block. The final feature vector for the whole part image is obtained by concatenating the vectors of all the blocks.

When the initial part image is rotated such that its orientation is not aligned with the image grid, the default approach is to normalize this situation by counter-rotating the entire image (or the bounding box of the part) before processing it as a canonical window. This can be computationally expensive during training, where image parts have all sorts of orientations, and during testing, even if we limit the number of detectable angles. Furthermore, dealing with changes in the scaling factor of the human figures and the foreshortening of limbs introduces additional computational burdens. In the following, we introduce a novel approximation method that manages to speed up the detection process.

Rotation and Scaling Approximation

Let p be a body part defined by a matrix of $M_p \times N_p$ cells (see Fig. 5). Rotating this part by θ degrees away from the vertical orientation creates two problems: how to compute the histograms in STEP 2, and how to aggregate them by cells in STEP 3. STEP 1 can compute gradients regardless of the rotation and STEP 4 does not care after we have the cell aggregates.

The first problem arises because we need to collect a histogram of the gradient angles with respect to the axis of the rotated part, and they are instead expressed with respect to the image grid. We propose our first approximation: with a fine enough binning of the histograms (our resolution of 10° is double the HOG-dd), we can approximate the "rotated" histograms by circularly shifting the bin counts of the neutral histograms of $-r_\theta$ places, where $r_\theta = \text{round}(\theta/10^\circ)$. This operation is

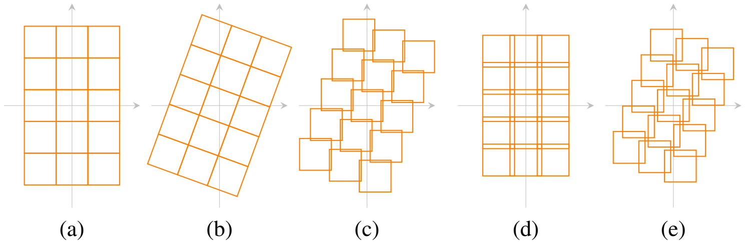


Fig. 5: Rotation approximation for a part defined by a matrix of 5 × 3 cells. From left to right: (a) default configuration with disjointed cells, (b) clockwise rotation by 20°, (c) approximation by non-rotated cells, (d) tighter configuration with cells overlapping by 1 pixel on each side, (e) rotation approximation of the tighter configuration. This approximation allows us to use the integral image technique.

much more efficient than re-computing the features after counter-rotating the source image, and can be performed fast for all the rotation angles we are interested in.

We solve the second problem by approximating the rotated cells with no rotation at all. As can be seen in Fig. 5), this leaves quite large holes in the covering of the part image, which is only partially mitigated by the spatial convolution in STEP 2 that spreads bin votes around. Our solution is to use a tighter packing of the cells, overlapped by 1 pixel on each side, so that they leave much smaller holes even at the worst angle for this approximation. The main purpose of avoiding rotated cells is that we can now use the integral image trick to efficiently aggregate histograms by cells for detection.

Scaling and foreshortening can be approached similarly, just by scaling the cells size (smaller or bigger than 7 × 7 pixels) and positioning them appropriately. As a partial motivation, [39] show that conveniently placed parts (cells in our approach) can effectively cope with perspective warps like foreshortening. As before, if we want to obtain HOG feature vectors for a different scaling factor, we can directly start with STEP 3 without going back to the start of the algorithm.

Efficient Detection

Detection of a given part from a new image is usually performed with a sliding window approach: a coarse or fine grid of detection points is selected, and the image is tested at each point by the detector, once for every orientation angle and scale allowed for the part (we usually are not interested in all angles or scales for pedestrians). This means extracting HOG feature vectors for many configurations of position, orientation, scale, and all the approximations introduced so far make this task very efficient, especially when we use the integral image technique.

In fact, at the end of STEP 2, instead of providing the gradient histograms, we compute their integral image, so that all sums in STEP 3 can be performed in constant time for each cell, in every configuration we wish for. If the resolution of the orientation angles matches the one in the histograms binning, we expect the least amount of information loss to happen in the approximations.

The last component of our fast detection algorithm is the LDA classifier. As shown in [25], LDA models can be trained almost trivially, and with little or no loss in performance compared to SVM classifiers. An LDA model classifies a given feature vector $\mathbf{x}_i$ as a part p instead of background if

$$\mathbf{w}_p^t \mathbf{x}_i - c_p > 0 \tag{3}$$

where

$$\mathbf{w}_p = \mathbf{S}^{-1}(\mathbf{m}_p - \mathbf{m}_{bg}) \tag{4}$$

$$c_p = \mathbf{w}_p^t(\mathbf{m}_p + \mathbf{m}_{bg})/2. \tag{5}$$

The background mean $\mathbf{m}_{bg}$ and the common covariance $\mathbf{S}$ are trained from many images including different objects and scenes, and $\mathbf{m}_p$ is trained from feature vectors extracted from annotated images (see left Fig. 6).

Furthermore, given the scores $f_i = \mathbf{w}_p^t \mathbf{x}_i - c$, we retrieve well calibrated probability values $p(\mathbf{x}_i)$ using the Platt scaling method [28], where

$$p(\mathbf{x}_i) = \frac{1}{1 + \exp(Af_i + B)} \tag{6}$$

and the parameters A and B are found using maximum likelihood estimation as

$$\arg\min_{A,B} \{ -\sum_i y_i \log p(\mathbf{x}_i) + (1 - y_i) \log(1 - p(\mathbf{x}_i)) \} \tag{7}$$

using the calibration set (f_i, y_i) with labels $y_i \in \{0, 1\}$.

4.2 Pose estimation

After the part detectors independently scan an input image, giving us image evidence $D = \{\mathbf{d}_p\}$, it is time to detect full body configurations, denoted as $L = \{\mathbf{l}_p\}$, where $\mathbf{l}_p = (x_p, y_p, \vartheta_p, s_p)$ encodes position, orientation and scale of part p, respectively. In Pictorial Structures (PS), 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 links between parts. The latter is also called the *kinematic prior* because it can be seen as a system of masses (parts) and springs (joints) that rule the body's motions.

In fact, we can represent the prior as a factor graph (see Fig. 6), where we have two types of factors: the detection maps $p(\mathbf{d}_p|\mathbf{l}_p)$ (gray boxes) and the joints $p(\mathbf{l}_i|\mathbf{l}_j)$ (black boxes). This graph is actually a tree with the torso $p = 1$ as root, which means that we can use standard (non loopy) belief propagation to get the MAP estimates.

In particular, the joints are modeled as Gaussian distributions around the mean location of the joint, and messages passing from part i to part j can be quickly computed by using Gaussian convolution in the coordinate system of the joint, reachable by applying a transformation $\mathbf{l}_{ij} = T_{ij}(\mathbf{l}_i)$ from part i and $T_{ji}^{-1}(\mathbf{l}_{ij})$ towards part j. After training, a learned prior is made up of these transformations together with the joint covariances (see Fig. 6).

Furthermore, if we only require a single body detection (the default situation with one pedestrian per image), only the messages from the leaves to the root must be accurately computed. At that point, the MAP estimate for the torso is

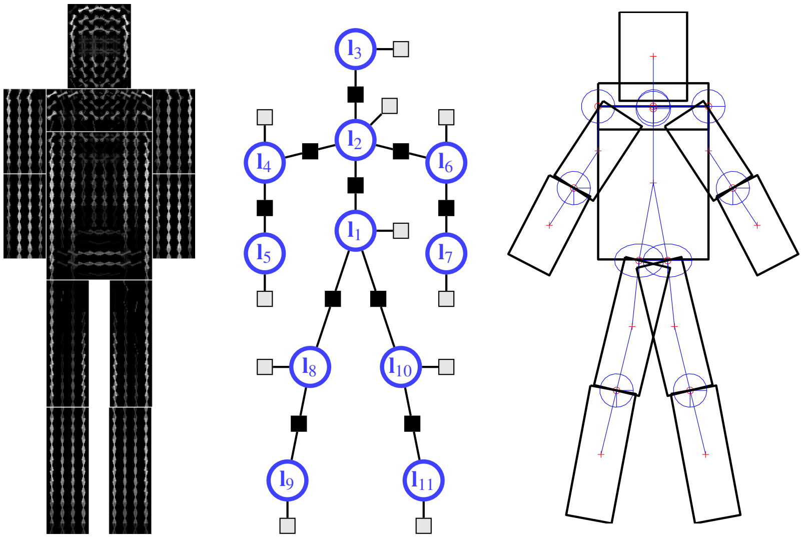


Fig. 6: (left) Composite image showing the positive weights in all the model weights $\mathbf{w}_p$ after training: each block shows the gradients that vote positively towards a part identification, with brighter colors in proportion to the vote strength. (center) Factor graph of the kinematic prior model for pose estimation. (right) Learned model of the relative position and rotation of the parts, including spatial localization covariances of the joints.

$\hat{\mathbf{l}}_1 = \arg\max_{\mathbf{l}_1} p(\mathbf{l}_1)$, and single delta impulses at $\hat{\mathbf{l}}_p$ can be messaged back to the leaves to find the MAP configurations for the other body parts.

Differently from other PS implementations for human figures, we decided to create configurations of 11 parts, adding a *shoulders* part, following the intuition of Dalal [10] that the head-shoulders combination seems to be critical for a good pedestrian detection.

4.3 Pedestrian Segmentation

To obtain well discriminating signatures, it is crucial to filter out as much of the background scene as possible, which is a potential source of spurious matches. After computing the pose estimation, we retrieve a segmentation of the image into separate body part regions, depending on the position and orientiation within the full body configuration. We encode such information in the form of image masks: thus, we get 11 body part masks and a combined set-union full body mask. We experimented early on several methods to further refine the masks to remove the residual background, but all such attempts resulted in worse performances. In part, this is due to the limited size of the images, usually cropped close to the pedestrian, that makes figure/background inference difficult.

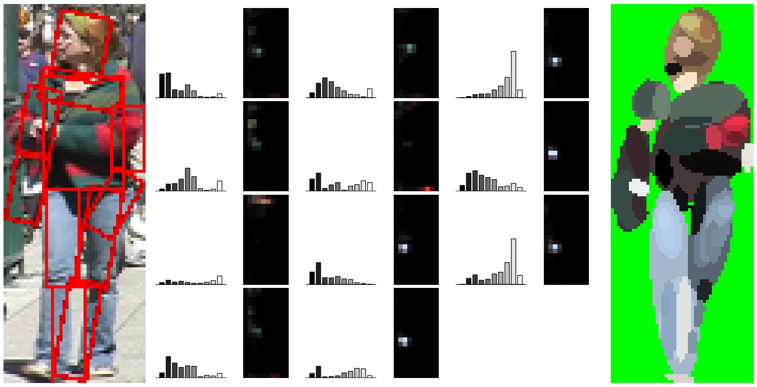


Fig. 7: (Left) A sample image from VIPeR with parts segmentation. (Center) Color histogram features, shown here separately for the 11 parts, each comprising of a histogram of grays and a histogram of colors. (Right) Blobs from the MSCR operator.

4.4 Feature Extraction

Having the masks, the task is to identify feature extraction methods that provide discriminating and robust signatures. As in our previous work [9], we rely on two proven techniques: color histograms and *maximally stable color regions* (MSCR) [19]. We experimented on several different variants of color histograms, both in our previous work and in this one: it is our experience that each dataset is suited to certain methods rather than others, with no method clearly outperforming the rest.

However, we reached a good compromise with a variant that separates shades of gray from colored pixels. We first convert all pixel values (r,g,b) to the HSV color space (h,s,v), and then we perform the following selections: all pixels with value $v < \tau_{black}$ are counted in the bin of blacks, all remaining pixels with saturation $s < \tau_{gray}$ are counted in the gray bins according to their value v, all remaining pixels are counted in the color bins according to their hue-saturation coordinates (h,s).

We basically count the dark and unsaturated pixels separately from the others, and we ignore the brightness of the colored pixels, counting only their chromaticity in a 2D histogram (see Fig. 7). This procedure is also tweaked in several ways to improve speed and accuracy: the HSV channels are quantized into $[20,10,10]$ levels, the votes are (bi)linearly interpolated into the bins to avoid aliasing, the residual chromaticity of the gray pixels is counted into the color histograms with a weight proportional to their saturation s. The image regions of each part are processed separately and provide a combined grays-colors histogram (GC histogram in short) which is vectorized and normalized. We then multiply each of these histograms by the part relevance weights λ_p (shown for example in Fig. 3 (right)), and then concatenate and normalize to form a single feature vector. Moreover, we allow the algorithm to adapt to particular camera settings by varying the importance of grays vs colors with a weight w_G, which can be tuned for each dataset.

Independently, the full body masks are used to constrain the extraction of the MSCR blobs. The MSCR operator 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 MSCR blobs. As in [14], we create a signature $\mathrm{MSCR} = \{(y_i, \mathbf{c}_i) | i = 1, \dots, N\}$ containing the height and color of the N blobs. The algorithm is setup in a way that provides many small blobs and avoids creating ones too big (see Fig. 7). The rationale is that we want to localize details of the pedestrians appearance, which is more accurate for small blobs.

4.5 Signatures Matching

The color histograms and the MSCR blobs ultimately form our desired image signatures. Matching two signatures $I_a = (\mathbf{h}_a, \mathrm{MSCR}_a)$ and $I_b = (\mathbf{h}_b, \mathrm{MSCR}_b)$ is carried out by calculating the distance

$$d(I_a, I_b) = \beta \cdot d_h(\mathbf{h}_a, \mathbf{h}_b) + (1-\beta) \cdot d_{MSCR}(\mathrm{MSCR}_a, \mathrm{MSCR}_b), \tag{8}$$

where β balances the Bhattacharyya distance $d_h(\mathbf{h}_a, \mathbf{h}_b) = -\log(\sqrt{\mathbf{h}_a}^t \sqrt{\mathbf{h}_b})$ and the MSCR distance d_{MSCR}. The latter is obtained by first computing the set of distances between all blobs $(y_i, \mathbf{c}_i) \in \mathrm{MSCR}_a$ and $(y_j, \mathbf{c}_j) \in \mathrm{MSCR}_b$:

$$v_{ij} = \gamma \cdot d_y(y_i, y_i) + (1-\gamma) \cdot d_{lab}(\mathbf{c}_i, \mathbf{c}_j) \tag{9}$$

where γ balances the height distance $d_y = |y_i - y_j|/H$ and the color distance $d_{lab} = \|labcie(\mathbf{c}_i) - labcie(\mathbf{c}_j)\|/200$, which is the Euclidean distance in the LABCIE color space. Then, we compute the sets $M_a = \{(i, j) | v_{ij} \leq v_{ik}\}$ and $M_b = \{(i, j) | v_{ij} \leq v_{kj}\}$ of minimum distances from the two point of views, and finally obtain their average:

$$d_{MSCR}(\mathrm{MSCR}_a, \mathrm{MSCR}_b) = \frac{1}{|M_a \cup M_b|} \sum_{(i,j) \in M_a \cup M_b} v_{ij}. \tag{10}$$

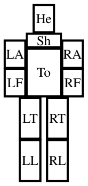
The normalization factor H for the height distance is set to the height of the images in the dataset, while the parameters β and γ are tuned through cross-validation.

Additionally, we have experimented with different distances than the Bhattacharyya, like Hellinger, L1, L2, Mahalanobis, χ^2, but performances were inferior.

4.6 Multi-shot Iteration

In multi-shot mode, we use CPS to improve the segmentations before extracting the features. This is a two-step iterative process that alternates between setting/updating the appearance model for the parts and updating the pose estimations. At the first iteration, we start with the conventional PS fittings, fed by the general part detectors. We thus collect all the part regions in the given images, normalize the different orientations, and stack them to estimate their common appearance. In particular, CPS employs a Gaussian model $\mathcal{N}(\mu_k, \sigma_k)$ in RGB space for all pixels k. In order to reinforce the statistics, the samples are extended by including spatial neighbors of similar color by performing k-means segmentation on each subimage t and in-

Table 1 Setup of the HOG-LDA detectors: configuration of the body parts used in our approach, with the canonical size in pixels and in number of cells. Detected orientations angles are −30°, −20°, −10°, 0°, 10°, 20°, 30°.

Parts	Size (pixels)	Size (cells)	Codenames
Torso	43×31	7×5	 He Sh LA RA To LF RF LT RT LL RL
Shoulders	13×31	2×5	
Head	25×19	4×3	
2×Arms	25×13	4×2	
2×Forearms	25×13	4×2	
2×Thighs	37×13	6×2	
2×Legs	27×13	6×2	

cluding the neighbors of k that belong to the same segment. The resulting Gaussian distribution is thus more robust to noise.

In the lead up to the second step of the iteration, these Gaussian models are used to evaluate the original images, scoring each location for similarity, providing thus *evidence maps* $p(\mathbf{e}_p|\mathbf{l}_p)$. This process can be efficiently performed using FFT-based Gaussian convolutions. Then, these maps must be combined with the part detections to feed the PS algorithm. Differently from [9], we experimented with different ways to combine them. It is our experience that maps that are too sparse and poorly populated generate pose estimations that rely on the default configuration in the kinematic prior. A fusion rule based on multiplication of probabilities (the default approach in a Bayesian update setting) tends to reduce the maps to isolated peaks. We thus propose a fusion rule based on the probability rule for union, which provides richer, but still selective, maps:

$$p(\mathbf{f}_p|\mathbf{l}_p) = p(\mathbf{d}_p|\mathbf{l}_p) + p(\mathbf{e}_p|\mathbf{l}_p) - p(\mathbf{d}_p|\mathbf{l}_p)p(\mathbf{e}_p|\mathbf{l}_p), \tag{11}$$

where the resulting $p(\mathbf{f}_p|\mathbf{l}_p)$ is then used in place of $p(\mathbf{d}_p|\mathbf{l}_p)$ in the pose estimation algorithm of Subsec. 4.2. Experimentally, CPS converges after 4-5 iterations, and we can finally extract 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×N single-shot distances, and keep the smallest one.

5 Training

Training was performed on the PARSE[1], the PASCAL VOC2010[13], and the INRIA Person[2] databases. PARSE consists of 305 images of people in various poses that can be mirrored to generate 610 training images. The database also provides labels for each image, in the form of locating 14 body points of interest. From these points it is possible to retrieve configurations of body parts to train the PS models, and our setup is described in Table 1. PASCAL and INRIA are used to generate negative cases: PASCAL has 17125 images containing all sorts of objects, including human figures of different sizes; INRIA Person has a negative training set of 1218 non-person images. In particular, as in [25], all the images in PASCAL were used to extract the background model for the HOG-LDA detectors, while the first

[1] `http://phoenix.ics.uci.edu/software/pose/`

[2] `http://pascal.inrialpes.fr/data/human/`

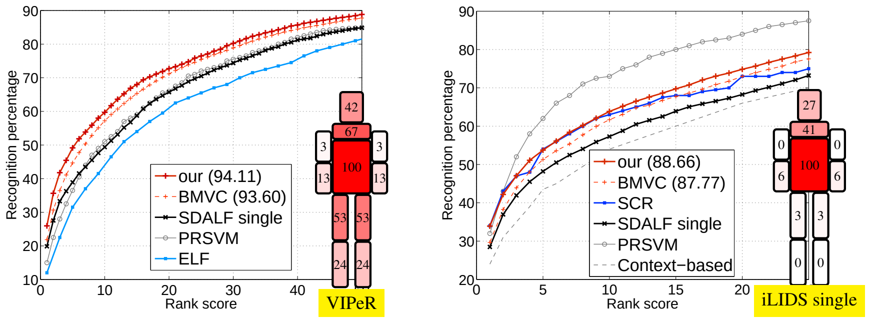


Fig. 8: Results of single-shot experiments on VIPeR (left) and iLIDS (right). Also shown on the puppets are the corresponding part weights: note how the legs in iLIDS are utterly useless because of too many occlusions by bags and trolleys.

200 annotated images in PARSE (mirrored to 400) were used to compute the foreground models for the parts. The remaining 105 images (mirrored to 210) and parts randomly drawn from INRIA Person's negative set were used to train the Platt calibration parameters. The PS kinematic model was trained on PARSE.

6 Experimental Results

In this section we present the experimental evaluation of our approach and we compare our results to those at the state of the art. The main performance report tool for re-identification is the Cumulative Matching Characteristic (CMC) curve, which plots the cumulative expectation of finding the correct match in the first *n* matches. Higher curves represent better performances, and hence it is also possible to compare results at-a-glance by computing the normalized area under curve (nAUC) value, indicated on the graphs within parentheses after the name when available. What follows is a detailed explanation of the experiments we performed on these datasets: VIPeR, iLIDS, ETHZ, CAVIAR for re-id.

Experimental Setup: The HOG-LDA detectors scan images once every 4 pixels and interpolate the results in between. The PS algorithm discards torso, head, shoulders detections below 50, 40, 30 percent of the image height, respectively. Only one scale is evaluated in each dataset since the images are normalized. The calibration parameters γ, β, w_G, and the part weights $\{\lambda_p\}$ are tuned by cross-validation on a portion of each dataset, before performing the test runs.

VIPeR Dataset [22]: This dataset contains 632 pedestrian image pairs taken from arbitrary viewpoints under varying illumination conditions. Each image is 128×48 pixels and presents a centered unoccluded human figure, although cropped short at the feet in some side views. In the literature, results on VIPeR are typically produced by mediating over ten runs, each consisting in a partition of randomly selected 316 image pairs. Our approach handily outperforms our previous result (BMVC in the figures), as well as SDALF [14], PRSVM [29], and ELF [22], setting the rank-1

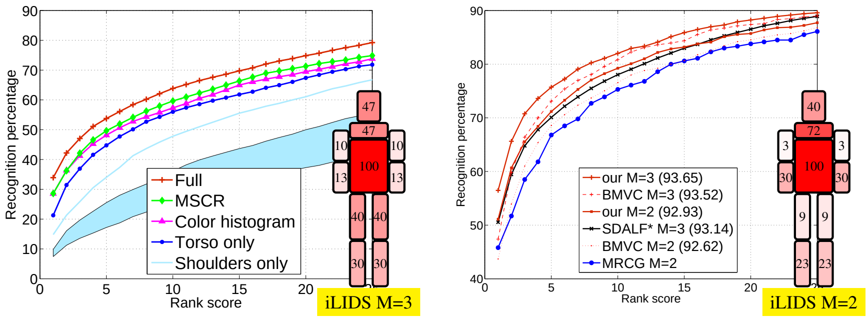


Fig. 9: (Left) Breakdown of our iLIDS multi-shot experiment showing the performance of the full distance, only the MSCR, only the color histograms, separately for the torso and shoulders parts (the shaded region contains the other parts curves). (Right) Comparison with the state of the art for multi-shot on iLIDS.

matching rate at 26%, and exceeding 61% at rank-10 (see Fig. 8 (left)). We note that weights for arms are very low, due to the fact that pose estimation is unable to correctly account for self-occlusions in side views, which abound in this dataset.

iLIDS Dataset: The iLIDS MCTS videos have been captured at a busy airport arrival hall [40]: the dataset consists of 119 pedestrians with 479 images that we normalize to 64×192 pixels. The images come from non-overlapping cameras, subject to quite large illumination changes and occlusions. On average, each individual has 4 images, with some ones having only 2. In the single-shot case, we reproduce the same experimental settings of [40, 14]: we randomly select one image for each pedestrian to build the gallery set, while all the remaining images (360) are used as probes. This is repeated 10 times, and the average CMC is displayed in Fig. 8 (right): we outperform all methods except for PRSVM [29], where the comparison is slightly unfair due to a completely different validation setup (learning-based). We do well compared to a covariance-based technique (SCR) [4] and the Context-based strategy of [40], which is also learning-based.

As for the multi-shot case, we follow a multi-vs-multi matching policy introduced in [14], where both probe and gallery sets have groups of M images per individual. We obtain our best result with $M = 3$, shown in Fig. 9 (left): the full distance combines individually good performances of the MSCR and color histogram distances detailed in Subsec. 4.5; also of note is that torso and shoulders are far more reliable than the other parts, even though, the high importance given to thighs and legs (see puppet) indicates a good support role in difficult matches.

In Fig. 9 (right), we compare our multi-shot results with SDALF* (obtained in the multi-vs-single modality $M = 3$, where galleries had three signatures and probes had a single one), mean Riemannian covariance grids (MRCG) [5]. We outperform all other results when we use $M = 3$ images, and we do resonably well even with $M = 2$. Although the parts weights do not give a definitive picture, it is suggestive to see worthless extremities in the single-shot experiment getting higher in the multi-shot $M = 2$ case, and finally becoming quite helpful in the $M = 3$ case.

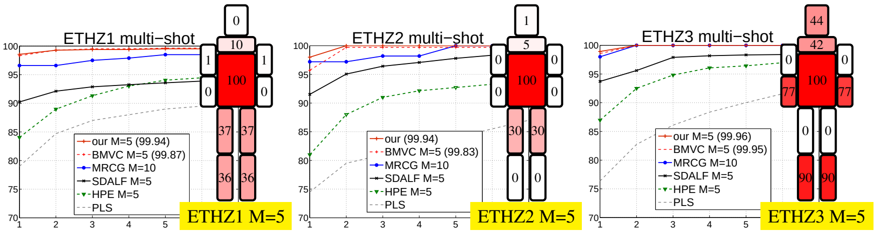


Fig. 10: Results of multi-shot experiments on the ETHZ sequnces.

ETHZ Dataset: Three video sequences have been captured with moving cameras at head height, originally intended for pedestrian detection. In [34], 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 × 96 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. Nevertheless, our approach is very close to obtaining perfect scores with $M = 5$. See Fig. 10 for a comparison with MCRG, SDALF and HPE. Note how the part weights behave rather strangely in ETHZ3: since the part weights are tuned on a particular tuning subset of the dataset, if this happens to give perfect re-id on a wide range of values for the parameters, then it is highly likely that they turn up some unreasonable values. In fact, checking the breakdown of the performances, it is apparent that the torso alone is able to re-id at 99.85%.

CAVIAR for re-id Dataset: CAVIAR4REID[4] has been introduced in [9] to provide a challenging real-world setup. The images have been cropped from CAVIAR video frames recorded by two different cameras in an indoor shopping center in Lisbon. 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. In our experiments, we reproduce the original setup: focusing only on the 50 double-camera subjects, we select M images from the first camera for the probe set and M images from the second camera as the gallery set, and then perform multi-shot re-id (called Camera-based multi-vs-multi, or CMvsM in short). All images are resized to 32 × 96 pixels. Both in single-shot and multi-shot, we outperform our previous results, SDALF (see Fig. 11) and AHPE [8]. The part weights also suggest relatively poor conditions, mainly due to the low resolution.

Computation Time: All experiments were run on a machine with one CPU (2.93 Ghz, 8 cores) and 8GB of RAM. The implentation was done in MATLAB (except for the MSCR algorithm), using the facilities of the Parallel Computing toolbox to take advantage of the multicore architecture. To establish a baseline, experiments on

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

[4] Available at `http://www.re-identification.net/`

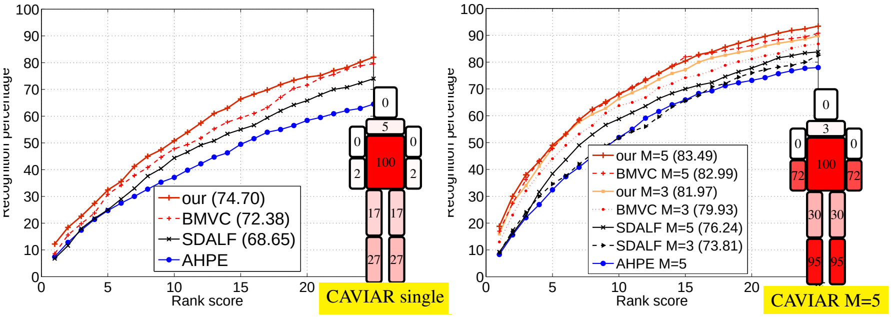


Fig. 11: Results of multi-shot experiments on CAVIAR4REID.

the VIPeR dataset with our approach initially require for each of the 1264 images: part detection to extract probability maps of size $128 \times 48 \times N_r \times N_p$ ($N_r = 7$ number of orientation angles, $N_p = 11$ number of parts), pose estimation, and feature extraction. Then, we calculate distances between all probes and galleries to produce a 632×632 matrix, and compute the matching and associated CMC curves for 10 trial runs of randomly chosen 316 subjects. The time taken by the last step is negligible since it is simply a matter of selecting and sorting distances, and can be safely ignored in this report.

We took the publicly available C++ source code of [1] and compiled it under Windows (after suitable adjustments), to compare against our approach: its part detection with SHAPE descriptors and AdaBoost is faster than our pure MATLAB code, while the pose estimation is slower because it provides full marginal posteriors (useful in other contexts than re-id) against our MAP estimates. We also report the speed of our approach when activating 8 parallel workers in MATLAB, noting that the C++ implementation can also run parallel processes. The time taken by distance calculations heavily depends on the distance being used: Bhattacharyya, Hellinger and L2 can be fully vectorized and take less than 1 s, χ^2 and L1 are slower, and distances like Earth Mover Distance are basically inpractical.

Running the full experiment on VIPeR takes approximately 30 minutes in single-thread mode, and 12 minutes using 8 parallel workers (see Table 2). Training the background model for the HOG-LDA detectors takes approximately 3 hours but it is done once for all detectors (even future ones for different parts or objects, as detailed in [25]), and negligible time for the foreground models. The kinematic prior estimation is also practically istantaneous.

7 Conclusions

When we approach the problem of person re-identification from a human point of view, it is reasonable to exploit our prior knowledge about person appearances: that they are decomposable into articulated parts, and that matching can be carried out per parts as well as on wholes. Thus, we proposed a framework for estimating the

Procedure	Input	Output	Time taken
Part detection [1]	VIPeR images (128×48 *pixels*)	1264 maps (128×48×7×11 *mats*)	12.5 min
Part detection *(single)*			20.5 min
Part detection *(8 parallel)*			4.4 min
Pose estimation [1]	VIPeR maps	1264 masks (128×48×11 *bin images*)	6.8 min
Pose estimation *(single)*			4 min
Pose estimation *(8 parallel)*			2 min
GC extraction	VIPeR images+masks	1264 hists (210×11 *mats*)	11-13 sec
MSCR extraction	VIPeR images+masks	1264 blobs lists	30 sec
MSCR dist. calculation	VIPeR blobs	632×632 mat	3.5-5 min

Table 2: Comparison of computation times for several procedures.

local configuration of body parts using Pictorial Structures, introducing novel part detectors that are easy and fast to train and to apply.

In our methodology and experimentation, we strived to devise discriminating and robust signatures for re-id. We currently settled on color histograms and MSCR features because of their speed and accuracy, but the overall framework is not dependant on them, and could be further enhanced. In fact, we plan to publicly release the source code of our system as an incentive for more comparative discussions.

Acknowledgements This work was supported by Hankuk University of Foreign Studies Research Fund of 2013.

References

1. Andriluka, M., Roth, S., Schiele, B.: Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1014–1021. Miami, USA (2009)
2. Bak, S., Charpiat, G., Corvee, E., Bremond, F., Thonnat, M.: Learning to match appearances by correlations in a covariance metric space. In: Computer Vision–ECCV 2012, pp. 806–820. Springer (2012)
3. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Person Re-identification Using Haar-based and DCD-based Signature. In: 2nd Workshop on Activity Monitoring by Multi-Camera Surveillance Systems (AMMCSS) (2010)
4. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Person Re-identification Using Spatial Covariance Regions of Human Body Parts. In: 7th IEEE Intl. Conf. on Advanced Video and Signal-Based Surveillance (AVSS) (2010)
5. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Multiple-shot human re-identification by mean riemannian covariance grid. In: Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on, pp. 179–184. IEEE (2011)
6. Barbosa, I.B., Cristani, M., Bue, A., Bazzani, L., Murino, V.: Re-identification with rgb-d sensors. In: A. Fusiello, V. Murino, R. Cucchiara (eds.) Computer Vision ECCV 2012. Workshops and Demonstrations, *Lecture Notes in Computer Science*, vol. 7583, pp. 433–442. Springer Berlin Heidelberg (2012)
7. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. Computer Vision and Image Understanding **117**(2), 130–144 (2013)

8. Bazzani, L., Cristani, M., Perina, A., Murino, V.: Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recognition Letters **33**(7), 898–903 (2012). Special Issue on Awards from ICPR 2010
9. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: British Machine Vision Conference (BMVC), pp. 1–11 (2011). Http://dx.doi.org/10.5244/C.25.68
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893 (2005)
11. Doretto, G., Sebastian, T., Tu, P., Rittscher, J.: Appearance-based person reidentification in camera networks: problem overview and current approaches. Journal of Ambient Intelligence and Humanized Computing **2**(2), 127–151 (2011)
12. Eichner, M., Ferrari, V.: Better Appearance Models for Pictorial Structures. In: British Machine Vision Conference (BMVC) (2009)
13. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html
14. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person Re-Identification by Symmetry-Driven Accumulation of Local Features. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2010)
15. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9), 1627–1645 (2010)
16. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
17. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial Structures for Object Recognition. Intl. J. on Computer Vision **61**(1), 55–79 (2005)
18. Figueira, D., Bazzani, L., Minh, H., Cristani, M., Bernardino, A., Murino, V.: Semi-supervised multi-feature learning for person re-identification. In: International Conference on Advanced Video and Signal-based Surveillance (AVSS) (2013)
19. Forssén, P.E.: Maximally Stable Colour Regions for Recognition and Matching. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2007)
20. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences **55**(1), 119–139 (1997)
21. Gheissari, N., Sebastian, T.B., Tu, P.H., , Rittscher, J., Hartley, R.: Person Reidentification Using SpatioTemporal Appearance. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 1528–1535 (2006)
22. Gray, D., Brennan, S., Tao, H.: Evaluating Appearance Models for Recognition, Reacquisition and Tracking. In: IEEE Intl. Workshop on Performance Evaluation for Tracking and Surveillance (PETS) (2007)
23. Gray, D., Tao, H.: Viewpoint Invariant Pedestrian Recognition with an Ensamble of Localized Features. In: Euro. Conf. on Computer Vision (ECCV), pp. 262–275 (2008)
24. Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B.: 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), pp. 1–6 (2008)
25. Hariharan, B., Malik, J., Ramanan, D.: Discriminative decorrelation for clustering and classification. In: ECCV, pp. 459–472 (2012)
26. Lin, Z., Davis, L.: Learning Pairwise Dissimilarity Profiles for Appearance Recognition in Visual Surveillance. In: 4th Intl. Symp. on Adv. in Visual Computing (2008)
27. Minh, H.Q., Bazzani, L., Murino, V.: A unifying framework for vector-valued manifold regularization and multi-view learning. In: Proceedings of the 30th International Conference on Machine Learning (ICML-13), ICML '13 (2013)
28. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)

29. Prosser, B., Zheng, W., Gong, S., Xiang, T.: Person Re-Identification by Support Vector Ranking. In: British Machine Vision Conference (BMVC) (2010)
30. Ramanan, D.: Learning to parse images of articulated bodies. In: Advances in Neural Information Processing Systems (NIPS), pp. 1129–1136 (2007)
31. Ramanan, D., Forsyth, D.A., Zisserman, A.: Tracking People by Learning Their Appearance. IEEE TPAMI **29**(1), 65–81 (2007)
32. Salvagnini, P., Bazzani, L., Cristani, M., Murino, V.: Person re-identification with a ptz camera: an introductory study. In: IEEE International Conference on Image Processing (ICIP 2013) (2013)
33. Satta, R., Fumera, G., Roli, F., Cristani, M., Murino, V.: A multiple component matching framework for person re-identification. In: 16th international conference on Image analysis and processing (ICIAP), ICIAP'11, pp. 140–149. Springer-Verlag, Berlin, Heidelberg (2011)
34. Schwartz, W., Davis, L.: Learning discriminative appearance-based models using partial least squares. In: XXII SIBGRAPI 2009 (2009)
35. Shapire, R., Freund, Y., Bartlett, P., Lee, W.: Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics **26**(5), 1651–1686 (1998)
36. Sivic, J., Zitnick, C.L., Szeliski, R.: Finding People in Repeated Shots of the Same Scene. In: British Machine Vision Conference (BMVC) (2006)
37. Wang, X., Doretto, G., Sebastian, T.B., Rittscher, J., Tu, P.H.: Shape and appearance context modeling. In: IEEE Intl. Conf. on Computer Vision (ICCV), pp. 1–8 (2007)
38. Wang, X., Han, T.X., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: ICCV, pp. 32–39 (2009)
39. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures-of-parts. Pattern Analysis and Machine Intelligence, IEEE Transactions on (2012)
40. Zheng, W., Gong, S., Xiang, T.: Associating Groups of People. In: British Machine Vision Conference (BMVC) (2009)
41. Zheng, W.S., Gong, S., Xiang, T.: Reidentification by relative distance comparison. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(3), 0653–668 (2013)

Index