Pedestrian Detection, Tracking and Re-Identification for Search in Visual Surveillance Data

Csaba Beleznai[1], Michael Rauter[1], Martin Hirzer[2], and Peter M. Roth[2]

[1] AIT Austrian Institute of Technology GmbH, Vienna, Austria
csaba.beleznai@ait.ac.at
[2] Inst. f. Computer Graphics and Vision, Graz University of Technology, Graz, Austria

Abstract. Visual surveillance data might encompass vast data amounts. Given the amount of data the need for search and data exploration arises naturally. Various authorities such as infrastructure operators and law enforcement agencies are confronted with search needs based on a visual description and/or behavioral patterns (motion path, activity) in order to find a "needle in a haystack of digital data". In this paper we present a framework which allows for an efficient search in visual surveillance archives. The paper describes following core algorithmic components of the search framework: Human detection employing pedestrian-specific shape and motion cues along with occlusion modelling; Tracking of multiple interacting pedestrians using a hierarchical spatio-temporal association scheme. Finally, pedestrian re-identification is demonstrated based on appearance matching in order to recognize a given person across a network of spatially disjoint cameras. We present results[1] for the detection, tracking and re-identification subtasks on various challenging datasets and describe the overall framework in detail.

1 Introduction

Exciting perspectives are emerging in the field of visual surveillance. Due to the rapidly growing amount of cameras and video data there is a need for quickly pinpointing relevant data within the "sea" of irrelevant. Manual search or browsing in such large archives is typically not feasible, since it is extremely time consuming, exhausting, and most likely unsuccessful because relevant data represents only a small fraction of the entire dataset. Consequently security-critical events often go undetected or cannot be prevented. This reduces the effectiveness of video surveillance systems. To render a video searchable, intermediate representations of the visual content are required. The set of these representations is often termed *meta-data*. Core algorithmic functionalities such as human detection and segmentation, tracking and appearance modelling are needed to derive

[1] The presented work is a compilation of our results originating from three (ICIP'12 [6], ECVW'11 [4], ICIP'10 [5]) papers, complemented by recent, mostly implementation and engineering results and achievements.

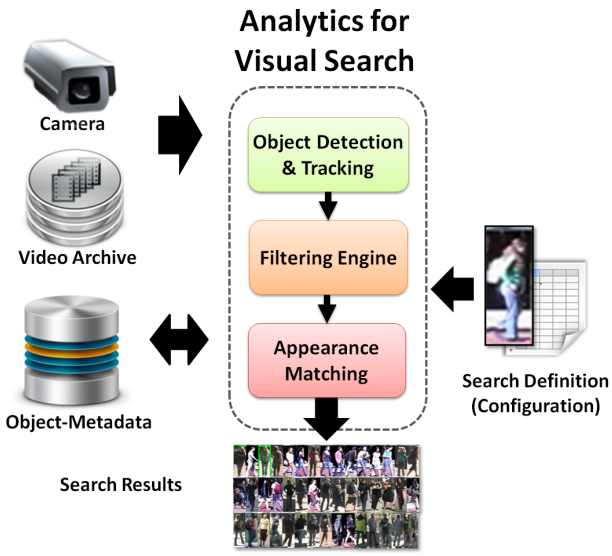

Fig. 1. The overall algorithmic framework performing the core analysis tasks (also termed as visual analytics). Visual data from surveillance archives or from camera nodes are used as an input. Visual analysis (as a service) generates intermediate representations (meta-data) and the final search result in form of a ranked list of hypothetical matches.

reliable representations which generate the meta-data. In recent years there has been an increased interest in visual surveillance search, also called as *forensic visual search* within the domain of visual analytics; nevertheless only few systems address the search task in the surveillance context. A relevant example is the IBM Smart Surveillance System [1], which is able to index a video according to multiple search criteria, thus allowing for various query types such as dominant object colors, object size and type and visual features of the human face. Berriss et al. [2] employ the MPEG-7 dominant color descriptor to efficiently associate and retrieve the same person across camera views in a retail environment.

We demonstrate a visual search framework (see Figure 1) and its main algorithmic components in great detail. The visual search task is accomplished by detecting pedestrians in videos, tracking them over time and generating a discriminative representation (meta-data) for all detected humans. These representations can be then used in a query-by-example manner to compare a query image to the pool of detected objects and retrieve a list of similarity-ranked potential matches. The main algorithmic components of the search framework are: First, real-time human detection accomplishing promising results in challenging scenes is presented, using the fast integral image based contour integration concepts presented in our previous paper [3]. Next, a computationally efficient multiple object tracking is described based on a simple spatio-temporal grouping scheme [5] and hierarchical partitioning of observations. Finally, a highly accurate pedestrian re-identification method is demonstrated adopting 4D spatial-color histogram

representations and the Large Margin Nearest Neighbor (LMNN) metric learning step [6] to estimate the transition between camera pairs.

The paper is structured as follows. The presented work includes several algorithmic topics, therefore we provide a brief state-of-the art description for all related subtasks in Section 2. Section 3 provides a concise overview on the overall visual search framework. Section 4, 5, and 6 describe the algorithmic solutions for pedestrian detection, tracking and re-identification, also including characteristic results for these core algorithmic units. Finally, Section 7 summarizes and concludes the paper.

2 Related work

In this section we describe the most relevant work related to the individual algorithmic topics.

Pedestrian detection: The need for automated detection of humans in digital images is substantial. The human detection task is a core issue in many applied fields of computer vision such as video surveillance, automotive safety and human-computer interaction. During the last two decades the pedestrian detection problem has received a great amount of interest and various representations and detection schemes have been proposed. The conventional blob-based object representation scheme is being gradually complemented or even replaced by representations and detection schemes primarily originating from the domain of visual object recognition. A typical example for such a novel scheme is the use of part-based representations encoding structure coupled with discriminative classification [7]. Variability of the human shape is treated in most cases by a set of local part models. Zhao and Nevatia [8] use a parametric human body model composed of elliptic shapes and probabilistically infer the most likely human configuration in the image using a computed motion segmentation. Blob-based motion segmentation is also used by Rodriguez and Shah [9] to implicitly capture local shape variations by learning a codebook of local descriptors. Lin et al. [10] decompose the human body into parametric parallelogram-shaped parts and generate a compact shape tree for efficient model evaluation. Recent reviews include [11–13] which benchmark state-of-the art schemes with respect to multiple criteria of practical relevance (most importantly scale and occlusion).

Multiple human tracking: Several tracking approaches exist relying on global data association techniques, which relate to our employed method. Markov Chain Monte Carlo techniques [14, 15], iteratively sample the hypothesis space and typically require a large number of iterations to find the close-to-optimum solution. Several recent techniques employ hierarchical association concepts to reduce the space of possible associations. Fei et al. [16] propose a graph-based optimization to solve the association problem and perform tracklet linking in several stages. Zhang et al. [17] also use a graph-theoretic approach (min-cost flow) to iteratively expand and associate the set of observations to form trajectories. Xing et al. [18] and Huang et al. [19] present a similar concept of first constructing tracklets using a conventional tracking method which is followed by

a one-step or multiple-step association stage. Problems with noisy and missing data still persist in all of these approaches, and the computional cost of association techniques is significant when tracking in dense or moderately dense (> 20 targets) scenes.

Pedestrian re-identification: Recognizing an individual person across a network of spatially disjoint cameras or distinct video streams by an informative description for human appearance is challenging. Each step of the representation process is associated with ambiguities. Many of the proposed person re-identification methods try to find a very distinctive and at the same time robust feature representation for describing a person's appearance. For instance, Wang et al. in [20] divide the image of a person into regions and capture their color spatial structure in a co-occurrence matrix. However, their approach is limited to people seen from similar viewpoints, an assumption that can not be made in most realistic setups. In [21], Farenzena et al. segment the silhouette of a person in order to find symmetry and asymmetry axes, which are then used for accumulating color and texture features. Cheng et al. [22] apply Pictorial Structures to tackle the person re-identification task. A body configuration composed of chest, head, thighs and legs is fit onto pedestrian images and used to extract per-part color information. Other methods build on learning to obtain a more discriminative feature model. For instance, Lin et al. [23] proposes to learn pairwise dissimilarities applicable for nearest neighbor classification. Prosser et al. [24] regard the person re-identification problem as a ranking problem and learn a subspace where the potential true match gets the highest rank. However, all of these methods ignore a simple given information: the transition from one camera to the other. Modeling the brightness transfer function between cameras can learn photometric changes. Metric learning [25] can also enable a more effective classification.

In our visual surveillance search framework we combine several state-of-the art algorithmic components which allow for an accurate visual search, while maintaining fast computation and small meta-data memory footprint for indexing and searching large scale data.

3 The overall search framework

The visual search framework employs two stages: *indexing* (generating meta-data) and *similarity-based search*. The *indexing* step is based on the visual analysis of the raw visual input: objects (pedestrians) are detected and tracked in order to reliably segment them in space and time, and ultimately to derive specific appearance representations for these image regions. In the second *search* step, the input of the visual search is an image exemplar (such as a snapshot of a person). A discriminative description (meta-data descriptor) is derived from this query image and it is compared to the previously computed descriptions. Based on these comparisons a ranked list of hypothetical matches is returned. Coupled with the visual appearance constraints, conventional spatio-temporal constraints or rules limiting space and time can also be used to guide or com-

plement the search process. Based on the returned list of potential matches the user has also the possibility to interactively refine the query by specifying details (e.g. which visual detail should the system focus on) and thus guiding the search framework towards the sought object.

In this paper we focus on the indexing step since it is the primary component containing relevant computer vision algorithms, often also termed visual analytics. All search system components are embedded into a framework following a service-oriented architecture strategy, implying that each component can be independently invoked as a networked service. Individual algorithmic components can thus be executed in a distributed manner within a network. For the sake of completeness we briefly describe the individual system services, while the visual analytics part is described in much more detail. The individual core services are the following functionalities:

(i) multiple video archives and camera inputs - our framework is capable to access a wide range of networked devices (archive systems, cameras) by using standard communication protocols.

(ii) visual analytics - visual analysis contains three relevant steps in form of detection, multiple object tracking and appearance-based modeling.

(iii) configuration and meta-data database services - parameters and generated meta-data such as the spatial and temporal location of objects, spatial extents and motion path, corresponding bitmaps and derived appearance descriptors are stored in a database. This database represents much less data than the original visual input and can be searched efficiently.

(iv) user interface and visualization - Visualization and interaction in form of services allow the user to carry out search on a remote client, such as a personal computer or mobile device.

In the following sections we describe the invidual vision algorithms (detection, tracking and appearance modeling) - carried out during the indexing phase as part of the visual analytics services - and their outcomes in more detail. Within the framework human detection aims at the generation and segmentation of image regions depicting humans; tracking plays a significant role to derive time-aggregated visual representations for each individual; finally, re-identification attempts to retrieve corresponding pedestrian image pairs within the large unstructured dataset of all pedestrian images.

4 Pedestrian detection

Given a digital image we would like to estimate the spatial configuration of humans (c^*) such that the hypothesized configuration best describes the observed image features I. Hence the detection task is postulated as a *maximum a posterior* (MAP) estimation problem:

$$c^* = \arg\max_{\mathbf{c}} \; P(\mathbf{c}|I), \tag{1}$$

A *configuration* encompasses a set of human hypotheses $c = \{h_1, h_2, \ldots, h_n\}$, where n denotes the number of humans forming the configuration. A given hu-

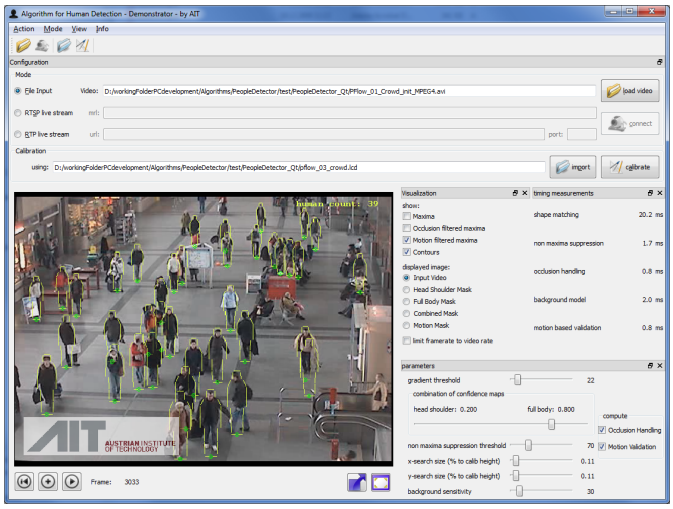

Fig. 2. Screenshot displaying sample pedestrian detection results.

man hypothesis h_i is characterized by a foot position $\mathbf{x}_i$ in the image and a corresponding shape C_i represented by a contour template: $h_i = \{x_i, C_i\}$. According to Bayes theorem the posterior probability is proportional to:

$$P(\mathbf{c}|I) \propto P(I|\mathbf{c})P(\mathbf{c}), \tag{2}$$

where $P(I|\mathbf{c})$ is the joint image-based likelihood and $P(\mathbf{c})$ denotes the prior probability of a configuration. Next, given this equation we describe the prior term and image-based measurement term in more detail and outline their role within the overall detection framework.

Employed priors: All spatial arrangements of individual human models are considered equally probable, therefore the prior depends on individual human model parameters (C) only. We assume that pedestrians stand upright on a common ground plane. We perform an off-line calibration step estimating a model $H(\mathbf{x})$ of the projected 2D human height in the scene. The estimated human height at a given image location is governed by the function $H(\mathbf{x}) = H_i(\mathbf{x})\, N(\mu_h, \sigma_h^2)$, where N is a Gaussian distribution (μ_h=1.0, σ_h= 0.08).

The variable human shape is modeled by a set of contour models. The variation of model parameters is learned in the following manner: 120 pedestrian images of the INRIA dataset [26] were annotated manually by adjusting a prototype contour set consisting of 13 oriented line segments to the human shapes seen in the training images. Annotated shapes - obtained for frontal and side views - were registered with respect to each other using foot and head locations on a common vertical human axis. A *Point Distribution Model* [27] representing the characteristic variation of segment end point coordinates is learned and k_T (k_T=30) shape samples $\{T_i\}_{i=1..k_T}$ are generated by considering only the principal modes of variation.

Joint image-based likelihood: The employed cues are shape and motion. Shape models are matched to edge-based image observations and motion proba-

Fig. 3. Top row: Sample detection results and mean occlusion rates (latter computed from annotations) obtained for the PETS2009 dataset [28]. Bottom row: Obtained detection and false alarm rates computed from all frames when using our framework.

bilities are computed from a binary map of moving foreground and static background generated by a conventional adaptive background modeling approach. Assuming independence between the two cues the image-based likelihood can be written as :

$$P(I|\mathbf{c}) = P(I_c|\mathbf{c})\, P(I_m|\mathbf{c}), \tag{3}$$

where $P(I_c|\mathbf{c})$ and $P(I_m|\mathbf{c})$ denote the shape-based and motion-based likelihoods, respectively.

A final joint optimization step estimates the configuration maximizing the posterior given the computed shape and motion-based probabilities. The optimization employs a greedy pruning strategy starting out from an initial pedestrian configuration (obtained by local maximum search and subsequent non-maxima suppression), where the computed image-based likelihoods and a local occlusion analysis are used to retain valid pedestrian hypotheses. Figure 2 shows a screen capture with a sample output of our real-time human detector demonstrator.

Results: Experiments carried out using the PETS2009 dataset [28] target the evaluation of detection performance as a function of varying human density. Manual annotation of the three sequences was used to generate ground truth for pedestrian location and visibility (occlusion rate between individuals). As it can be seen from Figure 3 detection and false alarm rates are significantly affected by the increasing human density from Scene 1 to Scene 3: heavily occluded persons remain often undetected and high human density produces clutter, thus leading to an increased rate of false alarms.

The framework was implemented for the CPU, the GPU and as a hybrid implementation. Figure 4 displays timing measurements using different template settings and implementations for the template matching step, which is the computationally most costly part of the framework. The total run-time of the complete algorithm on *PAL* resolution (720×576 *pixels*) images is 31 *ms*, 27 *ms*, and 23 *ms* for the CPU, GPU, and hybrid version, respectively. The employed

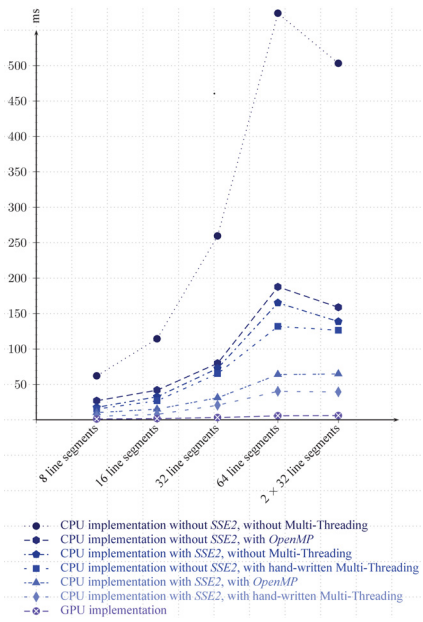

Fig. 4. Timing measurements for the template matching step: different implementations with different combinations of optimization measures.

hardware was an Intel Xeon CPU with 4 physical and 4 virtual cores @ $2.93\,GHz$, 12 GB RAM and a NVIDIA GeForce GTX 460. As it can be seen, the thoroughly optimized CPU implementation is not far behind the GPU implementation. The above experimental results show that the detection framework is capable of reliably detecting humans in moderately crowded scenarios at a high speed and it exhibits a slow degradation of detection performance at higher human densities.

5 Multiple pedestrian tracking

Our proposed approach is a data-oriented tracking method which relies on (i) two sets of observations (X and X_{weak}) provided by the human detector [3] and (ii) a prior height model $H(y)$.

The first set of observations $X = \{\mathbf{x}_i\}$ is generated by the detector using its optimum detection threshold T_{opt}. The second disjoint set of observations $X_{weak} = \{\mathbf{z}_i\}$ is created by collecting detection responses between T_{opt} and a much lower detection threshold T_{weak}. This second set of weak evidence is used only at the final association stage to select optimum trajectory hypotheses. The threshold values T_{opt} and T_{weak} are learned from a set of training videos by locating the optimum working point in their respective ROC curves.

The primary set of detection responses consists at the frame t_i of the attributes $\mathbf{x}_i = (x_i, y_i, a_i, o_i, t_i)$. x_i and y_i denote the image coordinates, a_i is a set of appearance-based descriptors and o_i is a binary flag of occlusion status

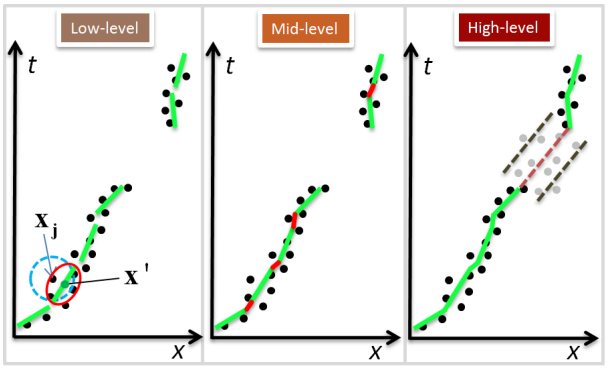

Fig. 5. Simplified illustration of the individual association steps. Left: Estimated trajectory segments (green) by local PCA. Center: Conservatively constrained linking between segments. Right: Final association step taking also weak evidence (sub-threshold detection responses (gray dots)) into account.

(set to 1 when more than half of the pedestrian's area is dynamically occluded, as quantified by the detector). The appearance $a_i = \{a_i^{up}, a_i^{low}\}$ is captured by computing Sigma Set [29] descriptors using the *Lab* and first derivative channels for the upper and lower halves of each detected object. The secondary set of weak evidence consists of the spatio-temporal coordinates only: $\mathbf{z}_i = (x_i, y_i, t_i)$.

The height model $H(y)$ of the projected 2D human height in the scene is obtained by an off-line calibration step.

Low-level association stage: The observations X are aggregated over time in the space-time volume. Aggregated observations exhibit a clear large-scale structure. In order to reveal the correlated structure of data we apply Principal Component Analysis (equivalent to the analysis of the local structure tensor) locally (further on denoted as local PCA or LPCA) to a subset of data points. The first grouping step is performed by the LPCA analysis and it generates a set of trajectory segments $S = \{S_k\}$ where $\{S_k = (x_k^1, y_k^1, t_k^1, x_k^2, y_k^2, t_k^2, \bar{a}_k, q_k)\}$. The first six coordinates denote the tail and head points of the segment, $\bar{a}_k$ represents an aggregated appearance descriptor and q_k is a quality measure. The individual steps of segment and descriptor generation are described as follows:

1. An initial observation $\mathbf{x}_j$ is chosen. (see Figure 5 left).

2. An analysis window (blue circle in Figure 5) with a radius of $\gamma H(y_j)$ ($\gamma = 0.5$) is used to locate the density maximum of local data distribution. Mean shift iterations using a uniform kernel are performed starting from (x_j, y_j, t_j) until convergence, which center the analysis window onto the data.

3. At the located density maximum $\mathbf{x}$' an eigenvalue decomposition of the local covariance matrix estimate is performed yielding a sorted set of eigenvalues $(\lambda_k^1, \lambda_k^2, \lambda_k^3)$ in descending order and corresponding eigenvectors (red ellipse in Figure 5).

4. Given the obtained approximation by a linear subspace we use the principal eigenvector to generate an oriented line segment (further on denoted as trajectory segment) to represent the local trend in the data distribution. The segment

Fig. 6. (top row): Example tracking results obtained for the PETS2009 S2.L1 sequence. (bottom row): Example tracking results obtained for the the railway station sequence.

is represented by the tail (x_k^1, y_k^1, t_k^1) and head (x_k^2, y_k^2, t_k^2) end points centered around $\mathbf{x}$'.

5. Based on the obtained eigenvalues the extent of anisotropy is estimated. If the underlying data structure is strongly correlated, then the ratio between the principal axis and both of the minor axes should be large. Otherwise, the data is isotropically scattered, which is an indication that no clear trend in the distribution can be estimated. Therefore, we formulate the quality measure $q_k \in [0, 1]$ of the estimated principal component as:

$$q_k = 1 - \exp\left(-\frac{\lambda_k^1}{\lambda_k^m + \delta}\right), \tag{4}$$

where $\lambda_k^m = max(\lambda_k^2, \lambda_k^3)$ and δ is small value to prevent division by zero.

6. An aggregated set of descriptors is computed for the segment from observations contributing to the LPCA estimate. The Sigma Set descriptors are compact and they can be easily aggregated by computing the arithmetic mean over a set of descriptors. Occluded observations ($o_j = 1$) contribute to the upper-body aggregated descriptor only.

7. A next observation is chosen to start from step 1 again. In our case observations having contributed to an LPCA estimate are excluded from the set of possible starting points.

Mid- and high-level association stages: The goal of this association stage is to a perform 'safe' linking between nearby consistent trajectory segments. Association is performed only for segments with $q_k > T_q$ ($T_q = 0.95$). We express the link probability between two trajectory segments as the product of motion-based, appearance-based and time-based affinities, derived from measured kinematic smoothness, appearance similarity and temporal ordering between individual trajectory segments. Using the estimated pairwise affinities

between segments and the standard Hungarian algorithm [30] we determine the optimal assignment within a spatio-temporal window around sampled segment locations. In the final high-level association stage the kinematic constraints are relaxed and weak evidence in form of sub-threshold detection responses is used to infer possible links between trajectory fragments (see Figure 5 right). More details on the data association scheme can be found in our paper [5].

Results: Example tracking results are shown for two video sequences. Video sequence 1 is the PETS 2009 S2.L1 dataset [28] depicting some walking people. The second video sequence shows a varying (from sparse to dense) density of crowd at a railway station. Results of the proposed tracking approach yield significantly better results than a conventional frame-to-frame association scheme (not shown), while exhibiting only slightly higher computational overhead.

6 Pedestrian re-identification

In this section we describe the last step of the visual analysis (indexing) part of the search framework where discriminative appearance features are extracted from the pedestrian image patches segmented by the previous detection and tracking steps. The pedestrian appearance modeling and re-identification step is illustrated in Figure 7. In particular, in order to better cope with photometric variations across different cameras, we build on two appearance modeling stages. First, we introduce a compact structure-encoding descriptor, which is mainly based on color information. Second, based on this description we learn a metric from a training set containing annotated matching image pairs, which yields a considerably better representation for the final nearest neighbor classification based matching step.

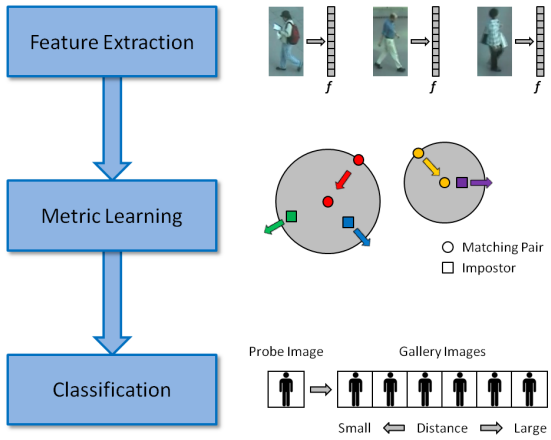

Fig. 7. Person re-identification system consisting of three stages: (a) feature extraction, (b) metric learning, (c) nearest neighbor classification.

Appearance modeling: A common approach to describe human visual appearance is via color histograms. Conventional color histograms lack spatial information therefore much effort has been untertaken to incorporate spatial features in order to enhance structural specificity. Joint feature space representations are appealing since they can be easily constructed, nevertheless, with increasing dimensionality they become sparsely populated, generate a large memory footprint and comparison between features becomes difficult. We employ a simple concept to approximate a high-dimensional distribution within a 4D feature space by a set of its projections: normalized height and *Lab* color coordinates are quantized to 40 bins and features of each pixel are mapped into three 2D histograms spanned by the *height-L*, *height-a* and *height-b* channels. Note that the current implementation still uses different appearance features for tracking and re-identification, a deficiency which will be removed in the near future.

Histogram-based features are known to benefit from computing the χ^2 distance in favor of the Euclidean distance. Thus, to bridge the gap between our histogram-based features and the proposed learning algorithm we first perform a homogeneous kernel mapping as proposed by [31]. In this way, the mapping enables to approximate the χ^2 distance without implications on the learner. Further, after obtaining the kernel mapping we perform a PCA to reduce the dimensionality of the feature space.

Metric learning for person re-identification: Metric learning allows to optimize ranking or classification results by exploiting the intrinsic structure of the feature space. One appealing class of metric learning algorithms is Mahalanobis distance learning. Given two data points $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{x}_j \in \mathbb{R}^d$, the squared Mahalanobis distance is estimated by

$$d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j), \tag{5}$$

where $\mathbf{M} \succeq 0$ is a positive semidefinite matrix.

In this work we build on Large Margin Nearest Neighbor (LMNN) [32] metric learning, which aims at improving k-NN classification. It has shown to yield robust results over a wide range of applications. The main idea of LMNN is to establish a local perimeter plus margin around each instance. Samples with different labels that invade the perimeter (impostors) are penalized, yielding the following objective function:

$$\epsilon(\mathbf{M}) = \sum_{j \rightsquigarrow i} \left[d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) + \mu \sum_l (1 - y_{il}) \xi_{ijl}(\mathbf{M}) \right] . \tag{6}$$

The first term minimizes the distance between target neighbors $\mathbf{x}_i$, $\mathbf{x}_j$, indicated by $j \rightsquigarrow i$. The second term denotes the amount by which impostors $\mathbf{x}_l$ invade the perimeter of i and j, where the slack variable $\xi_{ijl}(\mathbf{M})$ is given by

$$\xi_{ijl}(\mathbf{M}) = 1 + d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) - d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_l). \tag{7}$$

More details on the metric learning step can be found in our paper [6].

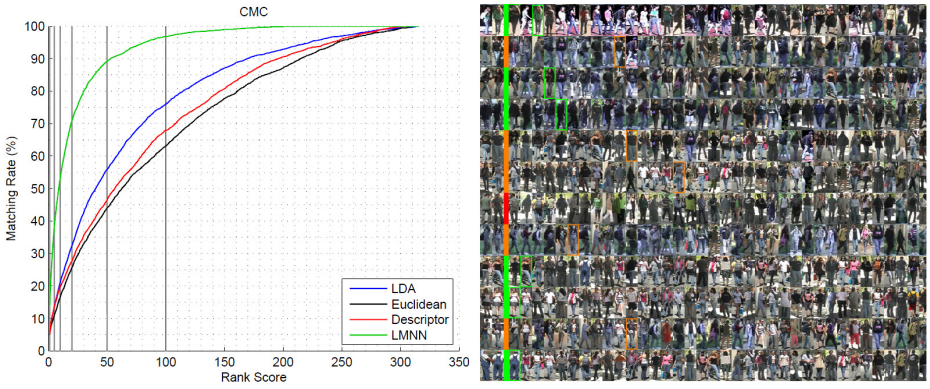

Fig. 8. (Left): CMC plots for the VIPeR dataset for different applied metrics. (Right): Sample outputs for person re-identification. The leftmost column shows individual query images; the second column displays corresponding true matches, while the image rows further to the right show the corresponding ranked lists of best matches. Color coding shows the quality of obtained match (green = good, red = poor).

Matching: During the learning stage, the thus obtained features are used as input for learning the Mahalanobis matrix $\mathbf{M}$. During matching using Eq. (5) the distances between the query image sample and the set of stored images in a database (so-called probe set) are estimated, and a ranking is provided.

Results: We evaluated our approach on the VIPeR dataset [33]. The VIPeR dataset contains 632 person image pairs. The main challenges are viewpoint, pose and illumination changes between the two images of an individual. For evaluation on this dataset, we followed the procedure described in [33]: the 632 image pairs are randomly split into a training and a test set of equal size, and images of pairs in the test set are randomly assigned to the probe and the gallery set. Each image from the probe set is then matched with all images from the gallery set. The whole procedure is repeated 10 times and the average performance is depicted in form of Cumulative Matching Characteristic (CMC) curves [20], representing the expectation of finding the true match within the first n ranks.

The corresponding results are shown in Figure 8, where we compare the original descriptor to the proposed metric-based evaluation. For the latter one PCA was used to reduce the number of dimensions to 45. It can be seen that due to the dimension reduction no performance is lost, and it is revealed that estimating the camera transition by a learned metric leads to superior results. In addition, as a simple baseline, we also show results obtained via Linear Discriminant Analysis (LDA) (carried out within the original feature space), which can be considered as a simple metric learner.

The above described algorithmic components have been used to build a final visual search framework. Within this framework human detection targets the generation and segmentation of image regions depicting humans; tracking plays a significant role to derive aggregated visual representations for each individual; finally, re-identification focuses on the discriminative matching of corresponding

pedestrian image pairs within the large unstructured dataset of all pedestrian images. The current implementation does not take spatial or temporal relations between pedestrian pairs into account when performing re-identification; this is a complementary information which can be easily integrated into the matching step.

7 Summary and conclusions

In this paper we presented a powerful visual search tool and its algorithmic components which allow for appearance-based similarity search of pedestrians in large surveillance archives or between spatially disjoint camera views. Most of the system's time-critical components have been implemented in a parallel fashion on general purpose graphics hardware or as optimized code for the CPU which allow for a significant acceleration of computations. The high computational speed and the highly specific visual representation with small memory footprint allow for time and memory efficient indexing (meta-data generation) and search in large datasets.

Acknowledgements

This work was supported by the Embedded Computer Vision (ECV) project under the COMET program and the SHARE project in the IV2Splus program, both projects of the Austrian Research Promotion Agency (FFG). Furthermore, support from the SECRET Interactive project of the KIRAS Security Research Promotion Programme of the Austrian Federal Ministry of Transport, Innovation and Technology is acknowledged.

References

1. Hampapur A., Brown L., Feris R., Senior A., Shu C. F., Tian Y., Zhai Y., Lu M.: Searching surveillance video. EEE Conference on Advanced Video and Signal Based Surveillance (AVSS'07), (2007)
2. Berriss W.P., Price W.G., Bober M.Z.: Real-Time Visual Analysis and Search Algorithms for Intelligent Video Surveillance. International Conference on Visual Information Engineering, (2003)
3. Beleznai, C., Bischof, H.: Fast human detection in crowded scenes by contour integration and local shape estimation. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (2009) 2246–2253
4. Beleznai, C., Schreiber, D., Rauter, M.: Pedestrian Detection using GPU-accelerated Multiple Cue Computation. Proc. Computer Vision and Pattern Recognition Workshops (CVPRW) (2011) 58–65
5. Beleznai, C., Schreiber, D.: Multiple object tracking by hierarchical association of spatio-temporal data. International Conference on Image Processing, (2010) 41–49
6. Hirzer, M., Beleznai, C., Köstinger, M., Roth, P. M., Bischof, H.: Dense appearance modeling and efficient learning of image transitions for person re-identification. ICIP 2012, (2012) 41–44

7. Felzenszwalb, P. F., Girshick, R. B., McAllester, D. A.: Cascade object detection with deformable part models. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (2010) 2241–2248
8. Zhao, T., Nevatia, R.: Bayesian Human Segmentation in Crowded Situations. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (2003) 459–466
9. Rodriguez, M. D., Shah, M.: Detecting and segmenting humans in crowded scenes. Proc. of the 15th international conference on Multimedia (2007) 353–356
10. Lin, Z., Davis, L. S., Doermann, D.: Hierarchical Part-Template Matching for Human Detection and Segmentation. Proc. IEEE Int'l Conf. on Computer Vision (2007) 1–8
11. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian Detection: An Evaluation of the State of the Art- IEEE Trans. on Pattern Analysis and Machine Intelligence **99** (2011)
12. Hussein, M., Porikli, F., Davis, L.: A comprehensive evaluation framework and a comparative study for human detectors. Trans. Intell. Transport. Sys., **10** (2009) 417–427
13. Enzweiler, M. and Gavrila, D. M.: Monocular Pedestrian Detection: Survey and Experiments. IEEE Trans. on Pattern Analysis and Machine Intelligence **31** (2008) 2179–2195
14. Oh, S., Russell, S. J., Sastry S. S.: Markov chain Monte Carlo data association for general multiple-target tracking problems. Proc. 43rd IEEE Conf. on Decision and Control (2004) 735–742
15. Yu, Q., Medioni, G.: Multiple-Target Tracking by Spatiotemporal Monte Carlo Markov Chain Data Association. IEEE Trans. on Pattern Analysis and Machine Intelligence **31** (2008) 2196–2210
16. Fei, Y., Christmas, W., Kittler, J.: Layered Data Association Using Graph-Theoretic Formulation with Application to Tennis Ball Tracking in Monocular Sequences. IEEE Trans. on Pattern Analysis and Machine Intelligence **30** (2008) 1814–1830
17. Zhang, L.,Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (2008) 1–8
18. Xing, J., Ai, H., Lao, S.: Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (2009) 1200–1207
19. Huang, C, Wu, B., Nevatia, R.: Robust Object Tracking by Hierarchical Association of Detection Responses. Proc. European Conf. on Computer Vision, (2008) 788–801
20. Wang, X., Doretto, G., Sebastian, T. B., Rittscher, J., Tu, P. H.: Shape and appearance context modeling. Proc. IEEE Int'l Conf. on Computer Vision (2007) 1–8
21. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (2010) 2360–2367
22. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom Pictorial Structures for Re-identification. Proc. British Machine Vision Conf., (2011)
23. Lin, Z., Davis, L. S.:, Learning Pairwise Dissimilarity Profiles for Appearance Recognition in Visual Surveillance. Advances Int'l Visual Computing Symposium, **I** (2008) 23–34
24. Prosser, B., Zheng, W.-S., Gong, S., Xiang, T.: Person Re-Identification by Support Vector Ranking. Proc. British Machine Vision Conf., (2010) 21.1–21.11

25. Dikmen, M., Akbas, E., Huang, T. S., Narendra, A.: Pedestrian recognition with a learned metric. Proc. Asian Conf. on Computer Vision, (2010)
26. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (2005) 886–893
27. Cootes, T. F., Taylor, C. J., Cooper, D. H., Graham J.: Active shape models—their training and application. Computer Vision and Image Understanding, **61** (1995) 38–59
28. http://www.cvg.rdg.ac.uk/PETS2009/
29. Hong, X., Chang, H., Shan, S., Chen, X., Gao, W.: Sigma Set: A small second order statistical region descriptor. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (2009) 1802–1809
30. Kuhn H. W.: The Hungarian Method for the assignment problem. Naval Research Logistics Quarterly, **2** (1955) 83–97
31. Vedaldi, A., Zisserman, A.: Efficient Additive Kernels via Explicit Feature Maps. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (2010)
32. Weinberger, K. Q., Saul, L. K.: Fast solvers and efficient implementations for distance metric learning. Int. Conference on Maching Learning, (2008)
33. Gray, D., Brennan, S., Tao, H.: Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. Workshop on Performance Evaluation of Tracking and Surveillance, (2007)