

Appearance-based person reidentification in camera networks: problem overview and current approaches

Gianfranco Doretto · Thomas Sebastian · Peter Tu · Jens Rittscher

Received: 30 January 2010 / Accepted: 4 October 2010 / Published online: 14 January 2011
© Springer-Verlag 2011

Abstract Recent advances in visual tracking methods allow following a given object or individual in presence of significant clutter or partial occlusions in a single or a set of overlapping camera views. The question of when person detections in different views or at different time instants can be linked to the same individual is of fundamental importance to the video analysis in large-scale network of cameras. This is the person reidentification problem. The paper focuses on algorithms that use the overall appearance of an individual as opposed to passive biometrics such as face and gait. Methods that effectively address the challenges associated with changes in illumination, pose, and clothing appearance variation are discussed. More specifically, the development of a set of models that capture the overall appearance of an individual and can effectively be used for information retrieval are reviewed. Some of them provide a holistic description of a person, and some others require an intermediate step where specific body parts need to be identified. Some are designed to extract appearance features over time, and some others can operate reliably also on single images. The paper discusses algorithms for speeding up the computation of signatures. In particular it describes very fast procedures for computing co-occurrence matrices

by leveraging a generalization of the integral representation of images. The algorithms are deployed and tested in a camera network comprising of three cameras with non-overlapping field of views, where a multi-camera multi-target tracker links the tracks in different cameras by reidentifying the same people appearing in different views.

Keywords Re-identification · Surveillance · Tracking · Appearance matching · Integral image · Co-occurrence · Integral representation

1 Introduction

Installations of camera networks nowadays are widespread in various domains that range from home surveillance applications, to small business and large retail applications, to facility access, sports venue, mass-transit, and environment monitoring, to open borders surveillance, just to mention a few. This raises the need for automated methods able to extract, and access high-level semantic information carried by the extremely high volume of recorded video data. In many surveillance tasks knowing whether in a given scene, at a given position and time, a given person was present is of paramount importance, and justifies the efforts devoted to the development of systems that can perform detection and tracking of people (Tu et al. 2007).

Intrinsic to the idea of tracking a person is the concept of maintaining his / her identity. In fact, tracking from one video frame to the next means being able to tell that the person that is pointed to is the same that was pointed in the previous frame. When only one video feed is processed, these identity management issues are addressed by so-called data association techniques, such as generalized nearest neighbor (Blackman and Popoli 1999), joint probabilistic

G. Doretto (✉)
West Virginia University, P.O. Box 6901,
Morgantown, WV 26506, USA
e-mail: gianfranco.doretto@mail.wvu.edu

T. Sebastian (✉) · P. Tu · J. Rittscher
GE Global Research, Niskayuna, NY 12309, USA
e-mail: sebastia@research.ge.com

P. Tu
e-mail: tu@research.ge.com

J. Rittscher
e-mail: rittsche@research.ge.com

data association filtering (Rasmussen and Hager 1998), multiple hypothesis tracking (Cox and Hingorani 1994), or Bayesian multi-target tracking (Isard and MacCormick 2001). Growing from one video feed to multiple video feeds, recorded simultaneously by multiple cameras (i.e. a camera network) with overlapping field of views, farther complicates the tracking problem. Here, besides pointing to the same person from one frame to the next, tracking means being able to point to the same person from one camera to the next, when he/she disappears from the former and appears in the latter. Most of current approaches leverage camera calibration, together with the spatiotemporal information of the target to maintain the identity during camera hand-off (Krahnstoeber et al. 2006; Khan and Shah 2006).

As the dimensions of a site grow, it quickly becomes unsustainable to be able to deploy a camera network where there are enough overlapping field of views to not leave uncovered any area of interest. In these conditions tracking across such “blind gaps” cannot take direct advantage of the space-time proximity of a person between two consecutive frames, or of the joint camera calibration and kinematic history of the person. Indeed, there is uncertainty in the behavior of a person in a blind gap, which is very hard to predict, not to mention that knowing the intrinsic and extrinsic calibration (Ma et al. 2004) would be a costly tedious process.

Passive biometric cues such as face (Senior et al. 2002), or gait (Wang et al. 2003) immediately stand out as information that might be valuable to address the identity management problem across blind gaps. *Signatures* describing the face or gait of an individual could be acquired “on-the-fly,” while the person is being reliably tracked. When the same person reappears in a field of view (from a blind gap), the same type of signature is extracted and matched against the original one. This *reidentification* process would allow to reassign to that individual the same identity, and tracking history, that was previously associated to him. In other words, *reidentification extends tracking beyond blind gaps*.

Despite the research efforts in face and gait recognition (Phillips et al. 2005; Bissacco and Soatto 2009), due to the low resolution and pose variation of individuals in typical visual surveillance footage, none of the two techniques were ever used for person reidentification in camera network settings, until recently. The interested reader is pointed to (Bäumel et al. 2010) for an approach that exploits face. Conversely, when suitable assumptions on people behavior in blind gaps can be made, there are approaches that constraint the reidentification problem, and combine spatial layout models with kinematic models to pruning the candidate set to be matched. The interested reader can consult (Makris et al. 2004; Rahimi et al. 2004; Javed et al. 2007) for such approaches that rely on kinematics and geometry.

Besides biometric cues, and geometric and kinematic priors, if the assumption that a person will not change his clothes during a blind gap is valid, the *whole body appearance* is an alternative cue that can be exploited for identity management in camera networks. The use of that information in such context is referred to as *appearance-based person reidentification*, which is the main focus of this paper. Note that, in general, a blind gap does not need to be generated by a person disappearing and reappearing in two disjoint field of views, instead, it could be due to a long occlusion of the person being tracked within the same field of view. Therefore, even a single camera tracker can take advantage of a form of reidentification capability to link the fragmented tracks of the same individual.

An appearance-based person reacquisition approach relies on estimating a signature (or model) representing the identity of the person being tracked. This is accomplished by either exploiting the appearance information of a person in one image only, or by integrating such information over multiple images. The former methods are referred to as *single-shots*, whereas the latter ones are referred to as *multiple-shots*. Such appearance information is extracted by cropping each video frame in correspondence of a tight bounding box, which is typically computed by the tracking algorithm that is following the person. Since the cropped image contains background image clutter, most of the reacquisition approaches take advantage of a foreground-background separation algorithm to better extract the image pixels that belong to the person only.

Once the whole-body appearance information to be used is identified, the goal is to compute a *distinctive*, and *invariant* signature for the person being tracked. It has to be distinctive because it should maximize the discrepancy computed against the signature of another individual. It has to be invariant because it should not change, regardless of the particular *pose* of the individual with respect to the camera (*viewpoint*), and regardless of the particular *illumination* conditions. Building distinctive and invariant signatures is key to performing reliable matching and therefore reidentification. Note that this is an extremely difficult task, especially given by the typical unconstrained surveillance scenario where people are imaged. In fact, small variations in pose or illumination rapidly change the appearance of loose or wrinkled *clothing*. Moreover, changes in pose and / or viewpoint result in different *occlusions* of body parts¹. Finally, image *resolution*, typically relatively low, has also an important impact on the invariance of signatures.

¹ In real scenarios occlusions may arise by other people, or objects, in the scene that are covering the imaged person. This paper does not consider such cases, and assumes that a person appears unobstructed in front of the camera.

1.1 Overview of the approach

A typical way of building appearance models (or signatures) is by first computing *local descriptors* of an image, and then by aggregating them using different strategies. Different descriptors and strategies lead to different approaches for building signatures.

A local descriptor can be as simple as the image intensity, or a sophisticated chain of operations. Two different local descriptors are predominantly used in this work (see Sect. 3) The first one computes *histograms of oriented gradients*² (HOG) in the *Log-RGB* color space (Funt and Finlayson 1995), and can be used for generating either single-shot or multiple-shots signatures. The second one aims at combining color and structural information while being robust against dynamic appearance changes of clothing. Such descriptor is computed over multiple consecutive frames, and is suitable for generating multiple-shots signatures. Color information is represented by hue and saturation, where robustness to illumination variations is achieved via normalization. Structural information is computed with an *edgel* extraction procedure outlined as follows.

Unlike most rigid objects, the structural appearance of loose fitting or wrinkled clothing on perambulating individuals is highly dynamic. Hence, the application of a traditional edge operator (Canny 1986) will produce many spurious edges corresponding to wrinkles and folds in clothing. To address this issue, a spatiotemporal segmentation algorithm that generates salient edgel information is applied to the imagery (see Sect. 4) The watershed algorithm is used to generate an over-segmentation of each frame. A spatiotemporal graph is then generated by treating each region as a node, and placing edges between spatially and temporally adjacent regions. A graph partitioning algorithm that models each cluster as a minimum spanning tree is then used to generate salient edgels corresponding to the boundaries of each type of clothing.

The simplest way to aggregating local descriptors leads to models that are *holistic* representations of the whole-body appearance; as opposed to the *parts-based* methods that divide the body into regions, and therefore a parts-matching mechanism is required when appearance models are compared (see Fig. 1). Given the bounding box delimiting a person, an example of holistic representation of the appearance is the concatenation of the color and the structural histograms, provided by the hue and saturation, and the salient edgels histograms. In Sect. 5 this is referred to as the *bounding-box* model.

Bag-of-features approaches (Schiele and Crowley 2000; Lowe 2004; Mikolajczyk and Schmid 2005; Varma and Zisserman 2005; Winn et al. 2005; Fei-Fei and Perona 2005), are holistic representations of images given by a distribution / collection of, possibly densely computed, local descriptors, vector-quantized according to a predefined *appearance dictionary* (made of *appearance labels*). These rather simple models perform remarkably well in both specific object (intra-category) and object category (inter-category) recognition tasks, and are robust to occlusions, illumination, and viewpoint variations. Here the local descriptor based on the HOG in the Log-RGB color space is densely computed³ and vector-quantized to build a signature that is the histogram of appearance labels (see Sect. 5).

The major criticism of histogram-based models is their failure to capture higher-order information, such as the spatial distribution of the local descriptors (i.e. the appearance labels). Some approaches directly address this issue (Lazebnik et al. 2003; Wolf and Bileschi 2006; Shotton et al. 2006; Savarese et al. 2006), but they mainly focus on inter-category discrimination, as opposed to recognizing specific objects. Section 6 introduces the *appearance context* model, a holistic representation with the goal of capturing the spatial relations between local descriptors by describing the co-occurrence of appearance labels.

In parts-based methods it is the description of corresponding parts that is used for matching appearance models. Part identification and correspondence can be carried out in different ways. One is to use *interest point operators* (Lowe 2004; Mikolajczyk and Schmid 2005; Bay et al. 2008). However, their responses do not persist over extended periods of time due to the dynamic nature of the appearance of clothing. This is addressed by choosing an operator that generates a large number of responses in regions with high information content, thus increasing the probability of establishing true correspondences between images of the same individual. In Sect. 8 the Hessian affine invariant operator is used for this purpose (Mikolajczyk and Schmid 2005). Signature matching is used to establish correspondences between two sets of interest points, and associated parts. Reidentification is established by computing a match score based on the cardinality of the final set of correspondences.

Another way to identify body parts is by fitting a model to establish a mapping from one individual to another. In Sect. 9 a decomposable triangulated graph (Amit and Kong 1996; Felzenszwalb 2005) is used to model the articulated shape of a person. A dynamic programming algorithm is used to fit the model to the image of the person. *Model fitting*

² Different flavors of HOG's have proven to be successful in several settings (Lowe 2004; Dalal and Triggs 2005; Kumar and Hebert 2006).

³ In challenging situations a dense representation has been found outperforming the sparse one also by other authors (Gheissari et al. 2006; Vedaldi and Soatto 2006).

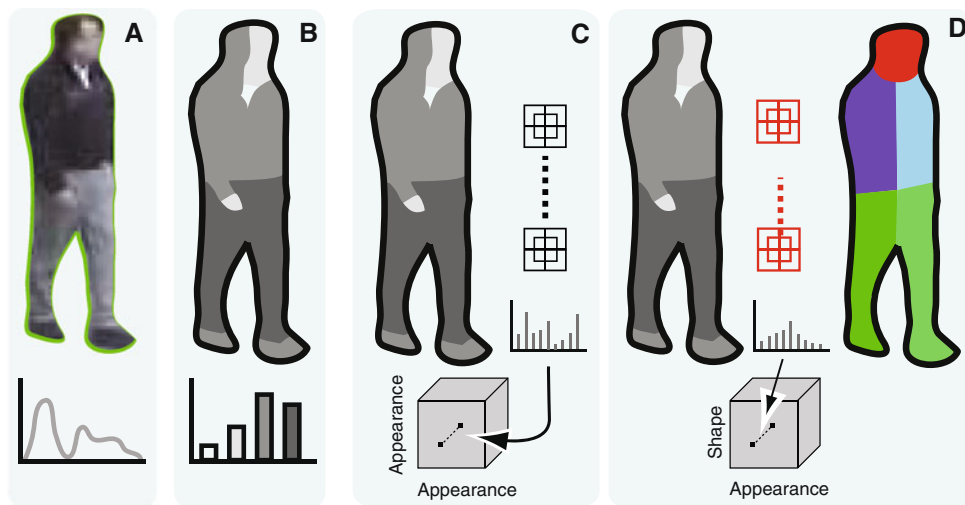


Fig. 1 Appearance models. Appearance models are typically built by first computing local descriptors of an image, and then by aggregating them using different strategies. One of the simplest is by computing their histogram. A local descriptor can be as simple as the image intensity (a), or a sophisticated chain of operations. Local descriptors are typically vector-quantized according to an appearance dictionary (b). More sophisticated aggregation strategies capture the spatial

localizes different body parts such as arms, torso, legs and head, thus facilitating the comparison of the corresponding appearance and structure.

Finally, an alternative method to identify parts is based on a modified *shape context* (Belongie et al. 2002) algorithm, which uses a *shape dictionary* learnt a priori. This effectively segments the image into regions that are loosely associated to specific body parts. The approach is used in conjunction with the *shape and appearance context* model (see Sect. 10), which is a parts-based method that extends the appearance context by using body parts explicitly to improve distinctiveness.

The proposed models entail computing several statistics over image subregions, which can be sped up very effectively by taking advantage of the *integral representation* (see Sect. 7.1). This generalizes the popular integral image and integral histogram (Viola and Jones 2004; Porikli 2005) to perform fast computations of statistics (e.g. mean and covariance) of multidimensional vector-valued functions over (discrete) domains of arbitrary shape. Based on this framework, it is possible to perform fast computations of co-occurrences (see Sect. 7.2), which leads to the real-time computation of the reidentification appearance models (Wang et al. 2007).

2 Related work

Several approaches have been proposed where signatures based on the whole body appearance of an individual are

relationships between appearance labels. Appearance label co-occurrence is one option (c). Part-based methods divide the body into parts, and the description of corresponding parts determines the matching, like for the interest operator and the model fitting approaches. Conversely, for corresponding parts one might attempt to match the appearance description of the surrounding parts, like it is done for the shape and appearance context model (d)

compared. In (Jaffré and Joly 2004) a color histogram of the region below the face (found by a face detector) serves as the signature for comparison. Seigneur et al. (2004) designs a related approach using clothing color descriptors. The early work of Nakajima et al. (2003) uses a holistic representation to simultaneously estimating pose and identity of four individuals by training a multi-class SVM. The works of Park et al. (2006) and Bird et al. (2005) are early examples of parts-based approaches for retrieval from video surveillance archives, and for detecting people loitering by reidentifying the same person being present for a long time.

Perhaps the first work that is mostly related to the application space of this paper is by Javed et al. (2003). They propose a Bayesian framework that attempts to fuse motion models and a simple color histogram representation to perform person reidentification between cameras with unknown spatial configuration. Followup work (Javed et al. 2005; Porikli 2003) shows that the inter-camera brightness transfer function lies in a low-dimensional subspace, and can be learnt using a set of corresponding calibration objects. Reidentification is then achieved by comparing the color histograms adjusted accordingly⁴.

Since the works of Javed et al. (2003) and Javed et al. (2005) the holistic representation with histograms has been

⁴ With a focus on vehicles, related work on appearance models has been done for reidentification (Guo et al., 2005), tracking (Zhao and Tao 2005), and category recognition (Ozcanli et al. 2006; Ma and Grimson 2005).

very popular for reidentification. In (Damen and Hogg 2007) it is used to associate people dropping off objects to the same people that pick them up. In (Pham et al. 2007), it is exploited to retrieve recurrent people from a surveillance video archive. Gandhi and Trivedi (2007) use histograms to characterize people with the so-called panoramic appearance map for reidentification in camera networks with overlapping field of views. In (Madden et al. 2007) histograms are made more robust to illumination variations with a form of equalization updated with online k -means clustering. Prosser et al. (2008) exploits histograms and builds on (Javed et al. 2005) to improve the estimation of the inter-camera brightness transfer function. Finally, Lin and Davis (2008) rely on a joint color and height histogram as the low-level feature to train a multi-class classifier for recognizing people.

Going beyond the representation based on histograms, there are a number of approaches that use training data to learn a holistic representation based on different low level features. This is usually done to make the final model robust to viewpoint variations. Gray and Tao (2008) use AdaBoost to learn a strong classifier made of weak learners that are functions of image position and intensity. Lo Presti et al. (2009) estimates and maintains in each node of a distributed camera network a Latent Dirichlet allocation model based on the bag-of-features representation. In (Truong Cong et al. 2009, 2010) a concatenation of features is used in combination with an SVM and manifold learning to perform reidentification in a lower dimensional space to handle multiple shots. Teixeira and Corte-Real (2009) update a bag-of-features model based on SIFT descriptors with online learning to improve matching. Finally, Bak et al. (2010b) uses AdaBoost in a cascaded approach for learning models based on Haar features and color.

Besides the early parts-based approaches (Park et al. 2006; Bird et al. 2005) that identify each part deterministically, and describe them with an histogram, more recently parts have been identified and described with interest point operators. In (Hamdoun et al. 2008) reidentification is based on matching SURF (Bay et al. 2008) interest points, collected during short video sequences, with a pyramid matching scheme. Cai et al. (2008) collects signatures of patches along edges that are matched to corresponding patches with the help of geometric constraints. Finally, (Oliveira de Oliveira and de Souza Pio 2009) augments the SURF descriptor with explicit color information to improve matching.

Other than interest operator parts-based approaches, recent works identify body parts in different ways, and also describe them with a combination of descriptors. Hu et al. (2008) adopts a fairly elaborate model composed by a generative and a discriminative model updated with online learning. The body is decomposed in several parts, each of

which is described with features such as color histograms, autocorrelograms, and a bag-of-features model based on SIFT descriptors. Similarly, Schwartz and Davis (2009) decompose the bounding box of a person in parts, each of which is described with appearance co-occurrence information, color histograms, and gradient histograms. Such high-dimensional signature is then projected onto a low-dimensional discriminant latent space by partial least squares reduction. A learning phase is required to update the models each time a new person is added to the pool of signatures used for reidentification. Bak et al. (2010a) trains specialized HOG detectors (Dalal and Triggs 2005) to identify body parts, each of which is then represented by a region covariance descriptor (Tuzel et al. 2006). A spatial pyramid matching paradigm is adopted for reidentification. Finally, (Farenzena et al. 2010) divides the body in three parts by computing body symmetry axis, thus obtaining the head, torso, and legs. Each part is then represented by color histograms, maximally stable color region descriptors (Forssen 2007) and a new patch matching analysis. No discriminative learning is necessary for this model.

3 Local descriptors

This section introduces the idea of local descriptor, which is the result of applying basic operations to the pixels of an image in order to build an appearance model. Computing a local descriptor allows either to highlight certain characteristics of the raw data, e.g. enhancing the edges, or to gain robustness / independence with respect to nuisance variables, e.g. illumination variations.

Let I be an image defined for every pixel \mathbf{x} belonging to a discrete domain Λ of dimensions $M \times N$ pixels. In this exposition I would represent the image inside a tight bounding box surrounding a person for which an appearance model has to be computed. The symbol Φ indicates a generic operation which maps I , and the pixel \mathbf{x} , to an r -dimensional *local descriptor* vector

$$\varphi(\mathbf{x}) = [\varphi_1(\mathbf{x}), \dots, \varphi_r(\mathbf{x})]^T \doteq \Phi(I, \mathbf{x}). \quad (1)$$

The operator Φ may represent either a bank of linear filters, or a complex non-linear procedure. A couple of operators that were found to be effective for reidentification are now introduced. The first one is the *histogram of oriented gradients in the Log-RGB color space* (or HOG Log-RGB operator, in short), the second one is the *hue, saturation, and edgel* (or HSV-edgel operator, in short).

3.1 HOG Log-RGB local descriptor

The experimental Sect. 11 reports a comparison between several bag-of-features models (introduced in Sect. 5)

using different operators Φ , such as different color spaces and filter banks, and tests their matching performance with the signature (14). What was found to perform very effectively is the operator Φ that computes the following local descriptor

$$\varphi(\mathbf{x}) \doteq \begin{bmatrix} \text{HOG}(\nabla \log(I_R), \mathbf{x}) \\ \text{HOG}(\nabla \log(I_G), \mathbf{x}) \\ \text{HOG}(\nabla \log(I_B), \mathbf{x}) \end{bmatrix}, \quad (2)$$

where I_R, I_G, I_B , are the R, G, and B channels of the image I , respectively. The operator $\text{HOG}(\cdot, \mathbf{x})$ computes the ℓ bins *histogram of oriented gradients* of the argument, on a region of $w \times w$ pixels around \mathbf{x} . In particular, the value of the gradient at pixel \mathbf{x} , given by $\nabla \log(I(\mathbf{x}))$, is counted, in the bin identified by the gradient orientation $\angle \nabla \log(I(\mathbf{x}))$, not by 1 but by the magnitude of the gradient $|\nabla \log(I(\mathbf{x}))|$. Note that the gradient of the Log-RGB space has an effect similar to the homomorphic filtering in that it makes the descriptor robust to illumination changes.

3.2 HSV-edgel local descriptor

The hue, saturation, and edgel operator describes the color and the structural information at every location. The color information is captured by the hue and saturation of a pixel value. The structural information gives a description of the edge, provided that there is one at that position. This description is referred to as *edgel*. Therefore, the operator Φ is such that

$$\varphi(\mathbf{x}) \doteq [I_H(\mathbf{x}), I_S(\mathbf{x}), \angle e(\mathbf{x}), e_R(\mathbf{x}), e_G(\mathbf{x}), e_B(\mathbf{x})]^T, \quad (3)$$

where I_H , and I_S represent the hue and saturation, and e indicates the edgel. In particular, to overcome the sensitivity of the hue to changes in illumination and shadows in outdoor scenes, here it is used the hue definition that is invariant to brightness and Gamma (Swain and Ballard 1991). For a given RGB color value the hue is given by

$$I_H = \arccos \frac{\log(I_R) - \log(I_G)}{\log(I_R) + \log(I_G) - 2 \log(I_B)}, \quad (4)$$

as opposed to the traditional definition which is

$$I_H = \arctan \frac{0.5[(I_R - I_G) + (I_R - I_B)]}{\sqrt{(I_R - I_G)(I_R - I_G) + (I_R - I_B)(I_G - I_B)}}. \quad (5)$$

As for the edgel e , it encodes the dominant local boundary orientation $\angle e$ (vertical or horizontal), as well as the ratios between the RGB color components of the two regions on either side of the edgel. The ratios of the RGB color components e_R, e_G, e_B , are each quantized to 4 possible values. This means that the edgel description $(\angle e, e_R, e_G, e_B)$ can be represented with 7 bits.

Note that the direct application of traditional edge detection algorithms such as (Canny 1986) to images of clothing produces many spurious responses. Hence, the HSV-edgel operator is used in conjunction with a spatio-temporal segmentation algorithm that is particularly suited to work with cloth appearance, and that generates stable salient edgels by rejecting edge information that is temporally unstable. Such procedure is described in Sect. 4. Because of its intrinsic nature, the HSV-edgel operator is typically used to produce multiple-shots signatures, where consecutive video frames are available.

4 Salient edgel extraction

This section introduces an algorithm for extracting stable structural information in the form of edgels which are used for defining appearance signatures. Even though many articles of clothing are derived from materials with uniform reflectance properties, a given type of material may appear quite different across an image and over time. This is because the surface normals of loosely-fitting clothing under articulated motion are highly dynamic. The proposed spatiotemporal segmentation method groups pixels that belong to the same type of fabric. Stable salient edgels are located at those pixels that are on the boundaries between two such groupings.

Observe that intra-fabric boundaries are not stable over time due to folds and wrinkles. This idea is exploited in the spatiotemporal segmentation. For a given time window an over-segmentation is performed on each image. This results in a set of contiguous regions $\{R_i^t\}$, where R_i^t is the i -th region of image I_t . A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined for a set of vertices $\mathcal{V} = \{v_i^t\}$ and edges $\mathcal{E} = \{e_{i,i'}^{t,t'}\}$ where v_i^t corresponds to region R_i^t and $e_{i,i'}^{t,t'}$ is an edge connecting vertices v_i^t and $v_{i'}^{t'}$. Region grouping is performed by partitioning \mathcal{G} into a set of clusters. A number of authors (Moscheni et al. 1998; Patras et al. 2001) have used region grouping over time for the purpose of foreground-background separation. However, the objective here differs from these applications in that the goal is to achieve a stable segmentation of the foreground. Another example of spatiotemporal segmentation can be found in (Zitnick and Kang 2005), which is meant to be consistent across neighboring frames and is used for video editing.

The over-segmentation used to define \mathcal{V} is performed in two stages. First, a Sobel operator is applied to the foreground of each grey level image and this is followed by Gaussian filtering. Second, a watershed segmentation algorithm (Vincent and Soille 1991) is applied. This results in regions of uniform intensity value for the foreground of the image, and the method is appropriate for articles of clothing that are not overly textured.

Given \mathcal{V} , the edge structure can then be defined. Two types of graph edges are constructed: spatial and temporal. If two regions R_i^t and $R_{i'}^t$ share a common boundary, then a spatial edge $e_{i,i'}^t$ is formed. For each region R_i^t , the region R_i^{t+1} is determined such that R_i^{t+1} has the highest likelihood of corresponding to the same material as R_i^t . This establishes the temporal edge $e_{i,i}^{t,t+1}$. The selection of R_i^{t+1} is determined based on estimates of the motion field, through the use of the *frequency image* $H_t(\mathbf{x}, \xi)$, defined as

$$H_t(\mathbf{x}, \xi) = \sum_{k=0}^{\xi} \eta(I_t(\mathbf{x}) - I_{t+k}(\mathbf{x})), \tag{6}$$

where for a threshold δ , η is such that

$$\eta(z) \doteq \begin{cases} 1, & \text{if } |z| < \delta; \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

For a given region with uniform intensity and uniform motion, the values of $H_t(\mathbf{x}, \xi)$ will be higher on the side of the region that corresponds to the direction of forward motion. For each overlapping region R_i^t and $R_{i'}^{t+1}$, the integral of $H_t(\mathbf{x}, \xi)$ over the intersection of R_i^t and $R_{i'}^{t+1}$ is computed. The overlapping region with the highest frequency integral is selected for a temporal edge. See Fig. 2 for an example of the frequency image.

If two adjacent regions correspond to the same piece of fabric they will periodically have a similar appearance. This is because intra-fabric boundaries are inherently unstable. Based on this, follows the definition of the edge weight $w_{i,i'}^{t,t'}$, which is the cost of grouping two regions together, given by

$$w_{i,i'}^{t,t'} = |K(i, t) - K(i', t')|; \tag{8}$$

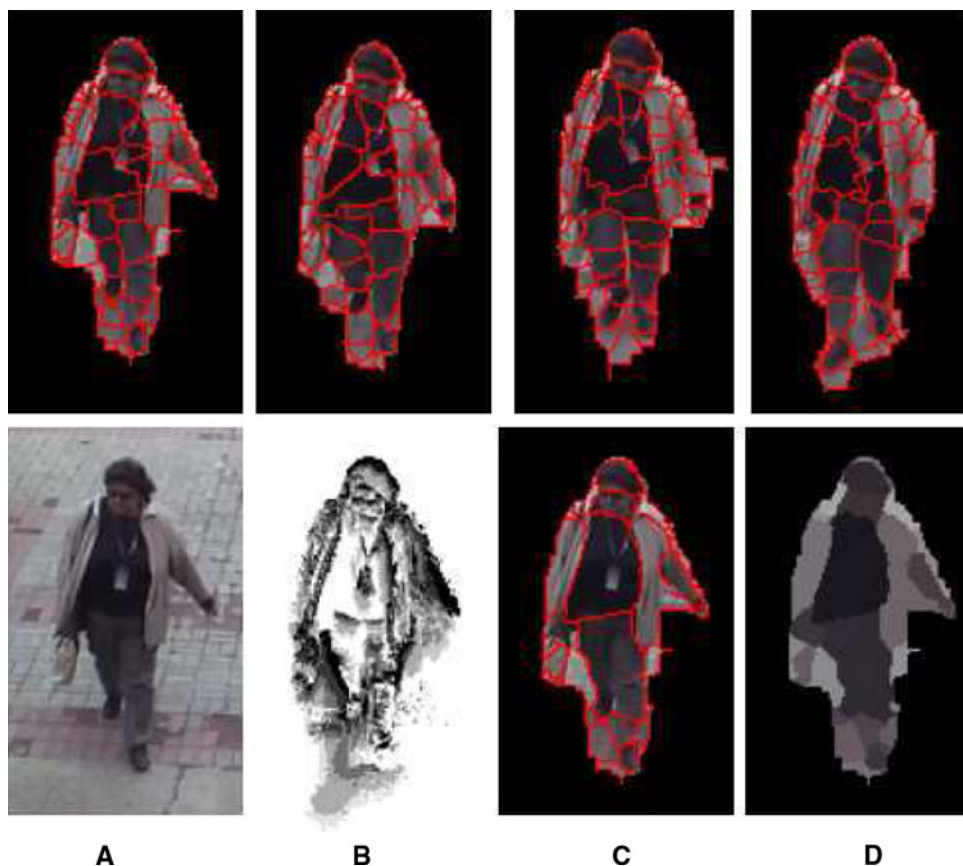
$$w_{i,i'}^{t,t+1} = \frac{1}{3} |K(i, t) - K(i', t+1)|, \tag{9}$$

where $K(i, t)$ is the median intensity value for region R_i^t . Note that a temporal edge allows for greater variation in appearance. In order to partition the graph the following principle is adopted: two regions should be grouped if there is a low cost path connecting them through space, time or a combination of both. This leads to an algorithm based on a search for clusters that have low-cost spatiotemporal minimal spanning trees, which is described in the following Sect. 4.1

4.1 Graph partitioning

Once the spatiotemporal graph \mathcal{G} has been generated for ten consecutive frames, the graph partitioning algorithm proposed by Felzenszwalb and Huttenlocher (2004) is used for

Fig. 2 Salient edgels. *Upper row* over segmentation of the foreground whole body appearance information for frames 0, 3, 6, 9. *Lower row* **a** original image, **b** frequency image for $\xi = 10$, **c** final segmentation after graph partitioning, **d** median image for final segmentation



grouping spatiotemporally similar regions. The idea is to merge connected clusters whenever the distance between them is less than the internal variation of each of the individual clusters. To efficiently implement this approach each cluster C is represented by the minimum spanning tree \mathcal{E}^C passing through all its vertices \mathcal{V}^C . The maximum edge weight of the minimum spanning tree is used to define the internal variation ΔC of the cluster C , i.e.

$$\Delta C \doteq \max\{w_{i,i'}^{t,t'} : e_{i,i'}^{t,t'} \in \mathcal{E}^C\}. \tag{10}$$

Given two clusters C_h and C_k , the inter-cluster distance $\Delta(C_h, C_k)$ is defined as the lowest edge weight between them, i.e.

$$\Delta(C_h, C_k) \doteq \min\{w_{i,i'}^{t,t'} : v_i^t \in \mathcal{V}^{C_h}, v_{i'}^{t'} \in \mathcal{V}^{C_k}, e_{i,i'}^{t,t'} \in \mathcal{E}\}. \tag{11}$$

Two clusters are merged if the inter-cluster distance is small when compared to the internal variation of the individual clusters. Specifically, clusters C_h and C_k are merged if

$$\Delta(C_h, C_k) \leq \min\left\{\Delta C_h + \frac{\kappa}{|C_h|}, \Delta C_k + \frac{\kappa}{|C_k|}\right\}, \tag{12}$$

where $|C|$ is the size of cluster C . The term $\kappa/|C|$ is based on the size of the cluster and encourages the formation of larger clusters. Since at the beginning of the merging process the internal variation of very small clusters tends to be too small, without that term the merging process would stop prematurely.

A greedy algorithm is proposed in (Felzenszwalb and Huttenlocher 2004) to obtain the graph segmentation that satisfies the above conditions. All edges in the spatiotemporal graph are sorted according to non-decreasing edge weights, and are then processed in that order. Let an edge $e_{i,i'}^{t,t'}$ between two separate clusters C_h and C_k be the one under consideration. If $w_{i,i'}^{t,t'} \leq \min\{\Delta C_h + \kappa/|C_h|, \Delta C_k + \kappa/|C_k|\}$, then C_h and C_k are merged and the edge $e_{i,i'}^{t,t'}$ added to the minimum spanning tree of the combined cluster. This step is repeated until all edges have been processed. It has been shown that the segmentation produced by the above algorithm is optimal in that the maximum edge weight for the minimum spanning tree for each cluster is smaller than the weights of all edges to each of their neighboring clusters. See Fig. 2 for an example of the application of this algorithm.

5 Histogram-based appearance models

In this section the HSV-edgel descriptor and the HOG Log-RGB descriptor are used to build two holistic models of the whole body appearance of an individual. They are referred

to as the *bounding box model* and the *bag-of-features model*, and they are both based on a histogram representation of the foreground appearance of the individual.

Once an operator Φ has been applied to the image, the computation of an histogram is made through the following steps. The first one performs a *quantization*, indicated by the function q , of each of the vectors φ , for every \mathbf{x} that belongs to the foreground. q is such that the composition $q \circ \varphi$ assumes values in a cartesian product of finite sets, each of which represents quantization levels. The second step computes the *histogram* h by doing the frequency count of the levels either jointly or separately within each set. In the latter case h results in a concatenation of histograms, each of which describes the distribution of the levels of one set.

5.1 Bounding box model

This model computes the HSV-edgel descriptor for every foreground pixel and then computes a histogram h , which is the concatenation of two parts. The first one performs the frequency count of the joint quantization of the hue and saturation channels I_H , and I_S . Similarly to (Mikolajczyk et al. 2004), the second part of h represents the structural qualities of the region through the edgels, each of which is quantized according to the 7 bits representation introduced in Sect. 3.

During matching, the distance between two models (or signatures) h_i and h_j is defined by the intersection histogram (Swain and Ballard 1991), i.e.

$$d(h_i, h_j) = 1 - \frac{2 \sum_a \min\{h_i(a), h_j(a)\}}{\sum_a h_i(a) + h_j(a)}, \tag{13}$$

where a is the variable that indexes a particular histogram bin.

5.2 Bag-of-features model

This model computes the HOG Log-RGB descriptor at every foreground pixel and then performs a vector quantization according to q , with quantization levels represented by $\mathcal{A} = \{a_1, \dots, a_m\}$. This produces the *appearance labeled image* $A \doteq q \circ \varphi$ (see Fig. 3 for an example). The set \mathcal{A} is referred to as the *appearance dictionary*, made of *appearance labels* learnt off-line. The following step is to compute the histogram of the labels h , which can be interpreted as follows. If a pixel position \mathbf{x} is randomly selected, the probability⁵ of that pixel being labeled with a is $h(a)$, i.e.

$$h(a) \doteq P[A(\mathbf{x}) = a | \mathbf{x} \in \Lambda]. \tag{14}$$

⁵ $P[\cdot]$ indicates a probability measure.

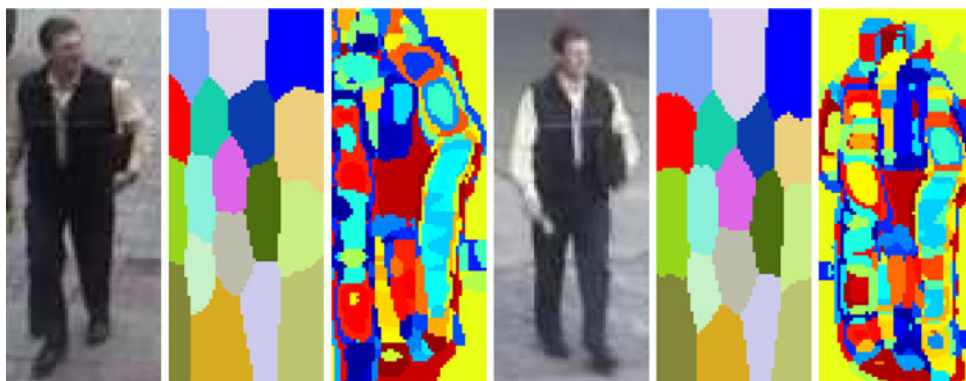


Fig. 3 Data set and shape and appearance labeled images. From left to right. Two samples from the data set of 143 different individuals, recorded from 3 different viewpoints, with corresponding shape labeled image S , and appearance labeled image A . Note that the decomposition into parts performed by S tries to compensate the misalignment induced by pose and viewpoint changes, as well as

person bounding box imprecisions. Also, neighboring pixels have similar shape context, which get quantized to the same label, thus producing the “piecewise-like” segmentation into body parts. As a welcome side effect, the occurrence computation takes great advantage of this segmentation, as C_S ends up being very small

At matching time, two models h_i and h_j are compared with the L_1 norm. Finally, the appearance dictionary \mathcal{A} is learnt off-line by applying k -means clustering to the training data.

6 Appearance context modeling

When two people are dressed up differently but with roughly the same amount of body surface covered with the same colors, they will likely have similar histogram-based signatures, regardless of how the colors are distributed in space. This is a major limitation of all the holistic models based on histograms because it significantly reduces their distinctive power. This issue is addressed here by describing a model that besides computing the histogram of the appearance labels, it describes how they are distributed by evaluating their spatial co-occurrence. Before defining the model, Sect. 6.1 introduces some notation and the general notion of occurrence.

6.1 Occurrence

Let $S : \Lambda \rightarrow \mathcal{S}$, and $A : \Lambda \rightarrow \mathcal{A}$ be two functions defined on a discrete domain Λ of dimensions $M \times N$, and assuming values in the label sets $\mathcal{S} = \{s_1, \dots, s_n\}$ and $\mathcal{A} = \{a_1, \dots, a_m\}$ respectively. Also, let $\mathcal{P} = \{p_1, \dots, p_l\}$ be a partition such that $\bigcup_i p_i$ represents the plane, and $p_i \cap p_j = \emptyset$, for $i \neq j$ (see Fig. 4 for an example). Given $p \in \mathcal{P}$ and a point on the plane \mathbf{x} , the quantity $p(\mathbf{x})$ indicates the partition element p , only translated by \mathbf{x} , i.e. $p(\mathbf{x}) \doteq \{\mathbf{x} + \mathbf{y} | \mathbf{y} \in p\}$. Also, the quantity $h(a, p(\mathbf{x}))$ indicates the histogram, or probability distribution, of the labels of A located in the region $p(\mathbf{x})$, i.e. $h(a, p(\mathbf{x})) \doteq P[A(\mathbf{z}) = a | \mathbf{z} \in p(\mathbf{x})]$. In other words, for a given A , after randomly selecting a point

$\mathbf{z} \in p(\mathbf{x})$, the probability that the label at that point will be a is given by $h(a, p(\mathbf{x}))$. Finally, let D_s indicate the set of points \mathbf{x} with label $s \in \mathcal{S}$, i.e. $D_s \doteq \{\mathbf{x} | S(\mathbf{x}) = s\}$. The definition of occurrence is given as follows (see Fig. 4).

Definition 1 The occurrence is a function $\Theta : \mathcal{A} \times \mathcal{S} \times \mathcal{P} \rightarrow \mathbb{R}_+$, such that the point (a, s, p) maps to

$$\Theta(a, s, p) \doteq E[h(a, p(\mathbf{x})) | D_s]. \tag{15}$$

The notation $E[\cdot | D]$ indicates the expectation with respect to a uniformly distributed random variable \mathbf{x} assuming values in D . The meaning of the occurrence function is the following: Given S and A , for a label $s = S(\mathbf{x})$ the histogram of the labels \mathcal{A} over the region $p(\mathbf{x})$ of A on average is given by $\Theta(\cdot, s, p)$. One special case is given when $\mathcal{S} = \mathcal{A}$, and $S(\mathbf{x}) = A(\mathbf{x})$, where Θ is typically referred to as co-occurrence. For a label $a = A(\mathbf{x})$ it says that the histogram of labels over the region $p(\mathbf{x})$ on average is $\Theta(\cdot, a, p)$.

6.2 Appearance context model

This model is obtained by computing the same local descriptor and quantization steps performed by the bag-of-features model. Subsequently, rather than forming the histogram of the labels, the model computes the co-occurrence Θ of the appearance labeled image A with the plane partition \mathcal{P} depicted in Fig. 9. Such appearance context signature is an $m \times m \times l$ matrix. The plane partition \mathcal{P} is made of p_1, \dots, p_l , L-shaped regions. Every quadrant of the plane is covered by $l/4$ regions (or partition elements). Each of the L-shape is $4N/l$ and $4M/l$ thick along the x_1 and x_2 directions, respectively. Therefore, the set of regions $\{p_i\}$ can be partitioned into groups of 4 elements forming concentric square rings.

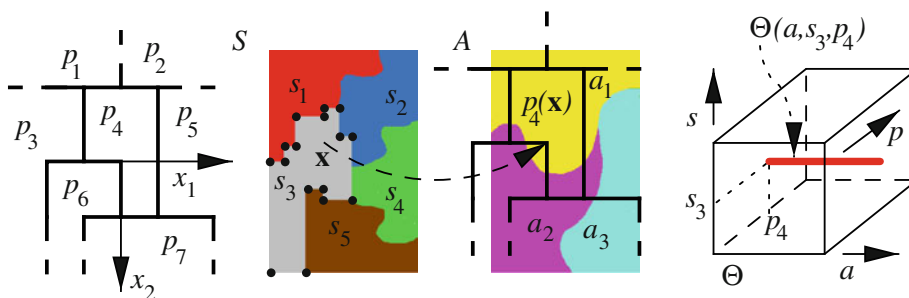


Fig. 4 Partition and occurrence definition. From *left to right* example of a generic partition of the plane \mathcal{P} ; example of a function S , and of a function A ; representation of Θ . If $h(a, p_4(\mathbf{x}))$ is the normalized count of the labels of A in $p_4(\mathbf{x})$ (the partition element p_4 translated at \mathbf{x}),

then by averaging $h(a, p_4(\mathbf{x}))$ over all $\mathbf{x} \in \{y|S(y) = s_3\} \doteq D_{s_3}$ one obtains $\Theta(a, s_3, p_4)$ (red line). The dots over S highlight the corner points $\nabla \cdot D_{s_3}$

The model captures the spatial arrangement of the appearance in the sense that for a label a , the co-occurrence $\Theta(\cdot, a, p)$ encodes the probability distribution of the labels in several spatial regions p , relative to the label a . Also, the information carried by the model (14) is included in the appearance context. In fact, by using Eq. 24 one can show that Θ reduces to Eq. 14. In particular, for every $b \in \mathcal{A}$ the following holds

$$h(a) = \frac{1}{|\Lambda|} \sum_{p \in \mathcal{P}} |p| \Theta(a, b, p), \tag{16}$$

where $|\cdot|$ identifies the area, or size of the argument. Section 7 derives a fast procedure, based on the integral representation of the image, for computing the co-occurrence.

Concerning invariance properties, the co-occurrence is known to be translation invariant. It can be made rotation invariant by choosing partition elements $\{p_i\}$ that are concentric circular rings. It is also robust to affine and pose changes (Huang et al. 1997; Savarese et al. 2006). The co-occurrence is not invariant with respect to the size of the image I . In order to have this property, before computing the appearance context the image needs to be normalized in size. Because of the morphology of the partition \mathcal{P} , the appearance context is not rotation invariant. For reidentification purposes this is a desirable property as it increases distinctiveness. In fact, lack of rotation invariance allows to have very large discrepancies between the models of a person wearing a white T-shirt and black pants versus a person wearing a black T-shirt and white pants.

7 Fast occurrence computation via integral representation

This section introduces a generalization of the popular integral representation of images known as integral image (Viola and Jones 2004). Then, it shows how this idea and formalism can be used to derive a computationally efficient algorithm for calculating the occurrence and the co-occurrence.

7.1 Integral representation

This section introduces a unified framework that generalizes the integral representation of images, and shows how such framework can be used for the fast computation of statistics not only over simple rectangular domains, but also over general non-simply connected rectangular domains.

Given a function $f(\mathbf{x}) : \mathbb{R}^k \rightarrow \mathbb{R}^m$, and a rectangular domain $D = [u_1, v_1] \times \dots \times [u_k, v_k] \subset \mathbb{R}^k$, if there exists an antiderivative⁶ $F(\mathbf{u}) : \mathbb{R}^k \rightarrow \mathbb{R}^m$, of $f(\mathbf{x})$, then

$$\int_D f(\mathbf{x}) \, d\mathbf{x} = \sum_{\mathbf{v} \in \mathbb{B}^k} (-1)^{\mathbf{v}^T \mathbf{1}} F(v_1 u_1 + \bar{v}_1 v_1, \dots, v_k u_k + \bar{v}_k v_k), \tag{17}$$

where $\mathbf{v} = (v_1, \dots, v_k)^T$, $\mathbf{v}^T \mathbf{1} = v_1 + \dots + v_k$, $\bar{v}_i = 1 - v_i$, and $\mathbb{B} = \{0, 1\}$. $k = 1$ and $k = 2$ correspond to the popular linear and planar cases, i.e.

$$\int_D f(x) \, dx = F(v_1) - F(u_1), \tag{18}$$

and

$$\int_D f(\mathbf{x}) \, d\mathbf{x} = F(v_1, v_2) - F(v_1, u_2) - F(u_1, v_2) + F(u_1, u_2), \tag{19}$$

respectively. Equation (17) defines a way for evaluating the integral of $f(\mathbf{x})$ that is very attractive when such quantity has to be repeatedly computed for different domains D . This is due to the following:

- In the discrete domain one specification of $F(\mathbf{u})$ can always be found, e.g.

$$F(\mathbf{u}) = \sum_{\mathbf{x} \leq \mathbf{u}} f(\mathbf{x}), \tag{20}$$

where $\mathbf{x} \leq \mathbf{u}$ is intended as $u_1 \leq x_1, \dots, u_k \leq x_k$. $F(\mathbf{u})$ here is referred to as the *integral representation* of $f(\mathbf{x})$.

⁶ If the Fubini's theorem for indefinite integrals holds, then $F(\mathbf{x})$ exists.

- $F(\mathbf{u})$ can be computed from a single pass inspection⁷ of $f(\mathbf{x})$, i.e. with computational cost $O(N^k)$, where N^k represents the dimension of the discrete domain where $f(\mathbf{x})$ is defined.
- Equation (17) enables the computation of statistics over the rectangular domain D in constant time $O(1)$, regardless of the size of D .

Some relevant specific cases of integral representation include the one where $k = 2, m = 1$, and $f(\mathbf{x})$ is a grayscale image, i.e. $f(\mathbf{x}) \doteq I(\mathbf{x})$, then $F(\mathbf{u})$ is referred to as the *integral image* of $I(\mathbf{x})$ (Viola and Jones 2004). If $f(\mathbf{x}) \doteq e \circ q \circ I(\mathbf{x})$, where $q : \mathbb{R} \rightarrow \mathcal{A}$ is a quantization (labeling) function, with quantization levels $\mathcal{A} = \{a_1, \dots, a_m\}$, and $e : \mathcal{A} \rightarrow \mathbb{N}^m$ is such that $a_i \mapsto e_i$, where e_i is the unit vector with only the i -th component different than 0, then $F(\mathbf{u})$ is the so called *integral histogram* of $I(\mathbf{x})$ with respect to q (Porikli 2005). In general, one has the freedom to design $f(\mathbf{x})$ in the best possible way in order to take advantage of the properties of the integral representation.

7.1.1 Computing statistics

By using the notation $E[\cdot|D]$ introduced in Sect. 6.1 one can write the expression of simple statistics, such as the mean⁸ of $f(\mathbf{x})$, where \mathbf{x} is intended as a uniform random variable assuming values in D , i.e.

$$E[f(\mathbf{x})|D] = \frac{1}{|D|} \int_D f(\mathbf{x}) d\mathbf{x}. \tag{21}$$

Similarly, the covariance of $f(\mathbf{x})$ over D can be expressed as

$$\begin{aligned} & E[(f(\mathbf{x}) - E[f(\mathbf{x})|D])(f(\mathbf{x}) - E[f(\mathbf{x})|D])^T | D] \\ &= \frac{1}{|D|} \int_D g(\mathbf{x}) d\mathbf{x} - \frac{1}{|D|^2} \int_D f(\mathbf{x}) d\mathbf{x} \int_D f(\mathbf{x})^T d\mathbf{x}, \end{aligned} \tag{22}$$

where $g(\mathbf{x}) : \mathbb{R}^k \rightarrow \mathbb{R}^{m \times m}$ is such that $\mathbf{x} \mapsto f(\mathbf{x})f(\mathbf{x})^T$. Higher-order moments could be written in this manner as well.

As the reader may notice, Eqs. 21 and 22 evaluate the left-hand side by computing integrals of either the function $f(\mathbf{x})$, or $g(\mathbf{x})$ over the domain D . Therefore, these operations can take advantage of the result expressed by Eq. 17. This means that one can first compute the integral representation of $f(\mathbf{x})$ and/or $g(\mathbf{x})$, and then evaluate the integral over D by performing a constant-time operation with computational complexity $O(1)$ thanks to (17). This is true regardless of the particular size of D . Note that, this way of computing statistics requires a previous evaluation of the integral

representation of each function. Therefore, this method becomes extremely efficient when the user is required to repeatedly compute similar statistics corresponding to different domains, given that the integral representation does not need to be recomputed.

The expressions (21) and (22) assume very different meanings according to the choice of $f(\mathbf{x})$. For instance, in the integral image they represent mean and covariance of the pixel intensities over the region D . On the other hand, in the integral histogram, (21) is the histogram of the pixels of the region D , according to a quantization q . In (Tuzel et al. 2006), (22) is used as an image region descriptor where $f(\mathbf{x})$ is the output of a bank of filters applied to the input image $I(\mathbf{x})$. In (Doretto and Yao 2010), (21) is used to perform the fast computation of region descriptors based on a generalization of image moments. In (Ke et al. 2005) (21) is used to extract volumetric features by maintaining an integral video representation of a sequence of images.

7.1.2 Domain generalization

The use of the integral representation has been mostly applied to regions D that are simple rectangles. Here the use of Eq. 17 is conveniently extended to the case of domains identified as follows (see Fig. 5 for a more intuitive representation).

Definition 2 $D \subset \mathbb{R}^k$ is a *generalized rectangular domain* if his boundary ∂D is made of a collection of portions of a finite number of hyperplanes perpendicular to one of the axes of \mathbb{R}^k .

For instance, the middle part of Fig. 5 shows in gray a generalized planar rectangular domain $D \subset \mathbb{R}^2$. As it can be seen, the boundary ∂D (represented by the bold lines) are made by a collections of segments, each of which is perpendicular to one of the two axes x_1 , or x_2 .

If $\nabla \cdot D$ indicates the set of corners of D , the following theorem describes how Eq. 17 extends to the case of generalized rectangular domains.

Theorem 1

$$\int_D f(\mathbf{x}) d\mathbf{x} = \sum_{\mathbf{u} \in \nabla \cdot D} \alpha_D(\mathbf{u}) F(\mathbf{u}), \tag{23}$$

where $\alpha_D : \mathbb{R}^k \rightarrow \mathbb{Z}$, is a map that depends on k . For $k = 2$ it is such that $\alpha_D(\mathbf{u}) \in \{0, \pm 1, \pm 2\}$, according to which of the 10 types of corners depicted in Fig. 5, \mathbf{u} belongs to.

Theorem 1, proved in Appendix, says that if D is a generalized rectangular domain, one can still compute the integral of $f(\mathbf{x})$ over D in constant time $O(1)$. This is done by summing up the values of $F(\mathbf{u})$, computed at the corners $\mathbf{u} \in \nabla \cdot D$, and multiplied by $\alpha_D(\mathbf{u})$, which depends on the

⁷ The single pass inspection of $f(\mathbf{x})$ is the k -dimensional extension of the 2-dimensional version described in (Viola and Jones 2004; Porikli 2005).

⁸ The operation $|\cdot|$ applied to a domain or a set indicates the area or the cardinality, respectively.

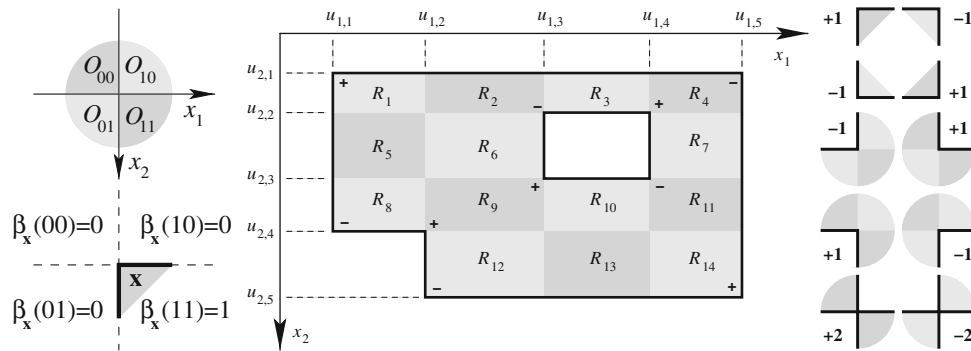


Fig. 5 Generalized rectangular domain and corner types. *Top-left* decomposition of \mathbb{R}^2 into his four quadrants $O_{11}, O_{01}, O_{00}, O_{10}$. *Bottom-left* example of a corner \mathbf{x} with the corresponding values assumed by the function $\beta_{\mathbf{x}}$. *Middle* Example of a generalized rectangular domain D , partitioned into simple rectangular domains

$\{R_i\}$. *Right* Function $\alpha_D(\mathbf{x})$. It assumes values different then zero only if \mathbf{x} is a corner of D . The specific value depends on the type of corner. For the planar case there are only ten types of corner, depicted here along with the corresponding values of α_D

type of the corner. For the planar case the types of corners are depicted in Fig. 5. Therefore, given any discrete domain D , by simply inspecting the corners to evaluate α_D , one can compute statistics over D in constant time $O(1)$. This simple and yet powerful result enables designing and computing fast and sophisticated region based image features, like the one in the following section.

7.2 Fast occurrence computation

This section introduces an algorithm for the fast computation of the occurrence and co-occurrence. The derivation is based on the fact that the occurrence is computed over a discrete domain Λ , where every possible sub-domain is a (discrete) generalized rectangular domain, and all the results of Sect. 7.1 can be applied, which lead to the following Theorem.

Theorem 2 The occurrence (15) is equal to

$$\Theta(a, s, p) = |D_s|^{-1} |p|^{-1} \sum_{\mathbf{x} \in \nabla \cdot D_s, \mathbf{y} \in \nabla \cdot p} \alpha_{D_s}(\mathbf{x}) \alpha_p(\mathbf{y}) G(a, \mathbf{x} + \mathbf{y}), \tag{24}$$

where

$$G(\cdot, \mathbf{x}) = \int_{-\infty}^{\mathbf{x}} \int_{-\infty}^{\mathbf{u}} e \circ A(\mathbf{v}) d\mathbf{v} d\mathbf{u}, \tag{25}$$

and⁹ $e : \mathcal{A} \rightarrow \mathbb{N}^m$ is such that the inner integral is the integral histogram of A .

Theorem 2, proved in the Appendix, leads to Algorithm 1. Note that even though the occurrence has been introduced for S and A defined on a two-dimensional domain,

the definition can be generalized to any dimension, and Theorem 2 still holds.

7.2.1 Complexity analysis

Given S and A , the naive approach to computing Θ has a time complexity of $O(N^4)$, where it is assumed that $M \sim N$, which is too slow for real-time applications, even if N is not very large. In (Huang et al. 1997) a dynamic programming approach reduces the cost to $O(N^3)$. In (Savarese et al. 2006) a particular partition \mathcal{P} where every $p \in \mathcal{P}$ is a square ring defined by $|\nabla \cdot p| = 8$ corners enables a computation cost¹⁰ of $O(N^2 l |\nabla \cdot p|) = O(N^2 C_{\mathcal{P}})$, where $C_{\mathcal{P}} \doteq l |\nabla \cdot p|$ represents the total number of corners of \mathcal{P} .

The computational cost of Algorithm 1 is computed as follows. Line 2 can be evaluated by a single pass inspection of A , and has the same computational cost of an integral histogram,¹¹ i.e. $O(N^2)$. Line 3–7 is another single pass inspection of S with cost $O(N^2)$. Line 12 costs $O(1)$. Line 11 is an average multiplying factor of $C_{\mathcal{P}}/l$, where $C_{\mathcal{P}} \doteq \sum_i |\nabla \cdot p_i|$. Line 10 is an average multiplying factor of C_S/n , where $C_S \doteq \sum_i |\nabla \cdot D_{s_i}|$. Line 8 and 9 are multiplying factors of n and l respectively. Therefore, the cost of 8–14 is $O(C_S C_{\mathcal{P}})$. Finally, the total cost of Algorithm 1 is $O(N^2 + C_S C_{\mathcal{P}})$, where in practice $C_S C_{\mathcal{P}} \sim N^2$. Therefore, Algorithm 1 has an effective cost of $O(N^2)$, which is $C_{\mathcal{P}}$ (the number of corner points of the partition \mathcal{P}) times faster then the one in (Savarese et al. 2006). It is interesting to note that Algorithm 1 is only marginally sensitive to the choice of the partition \mathcal{P} , which is allowed to be arbitrary, as opposed to (Savarese et al. 2006) where it is fixed.

⁹ Note that a $a \in \mathcal{A}$ is intended to index one of the elements of the m -dimensional vector $G(\cdot, \mathbf{x})$.

¹⁰ In (Savarese et al. 2006) $|\nabla \cdot p|$ is part of the hidden constants. Here the dependency is made explicit to better compare that approach with this one.

¹¹ Note that the analysis conducted here is independent of the cardinality of the label sets S and A .

Algorithm 1: Fast occurrence computation

```

Data: Functions  $A$  and  $S$ 
Result: Occurrence matrix  $\Theta$ 
1 begin
2   Use (25) to compute  $G$  from a single pass inspection of  $A$ 
   // Compute  $|D_s|$ ,  $\alpha_{D_s}$  and  $\nabla \cdot D_s$ 
3   foreach  $\mathbf{x} \in \Lambda$  do
4      $|D_{S(\mathbf{x})}| \leftarrow |D_{S(\mathbf{x})}| + 1$ 
5     if IsCorner( $\mathbf{x}$ ) then
6       Set  $\alpha_{D_{S(\mathbf{x})}}(\mathbf{x})$ 
7        $\nabla \cdot D_{S(\mathbf{x})} \leftarrow \nabla \cdot D_{S(\mathbf{x})} \cup \{\mathbf{x}\}$ 
   // Use (24) to compute  $\Theta$ 
   //  $|p|$ ,  $\alpha_p$  and  $\nabla \cdot p$  known a priori
8   foreach  $s \in S$  do
9     foreach  $p \in \mathcal{P}$  do
10      foreach  $\mathbf{x} \in \nabla \cdot D_s$  do
11        foreach  $\mathbf{y} \in \nabla \cdot p$  do
12           $\Theta(\cdot, s, p) \leftarrow \Theta(\cdot, s, p) +$ 
13             $\alpha_{D_s}(\mathbf{x})\alpha_p(\mathbf{y})G(\cdot, \mathbf{x} + \mathbf{y})$ 
14       $\Theta(\cdot, s, p) \leftarrow |D_s|^{-1}|p|^{-1}\Theta(\cdot, s, p)$ 
15   return  $\Theta$ 

```

8 Parts-based modeling by interest-point matching

This section introduces a parts-based appearance model where an interest operator is used to identify parts and establish correspondences between individuals. Given an image of a person, the Hessian affine invariant interest operator (Mikolajczyk et al. 2005) is used to nominate points of interest. The operator is limited to foreground patches extracted by a foreground-background separation algorithm (Gheissari et al. 2006). The Hessian operator is not stable over time. However, when compared to other methods (Mikolajczyk et al. 2005), it provides a large number of interest points and it is more informative with respect to color variation. This increases the probability of generating true correspondences between images of the same individual. For each interest point \mathbf{i} , a feature vector $h_{\mathbf{i}}$ is generated based on a circular support region $D(\mathbf{i}, \rho)$ of fixed radius ρ centered at pixel position \mathbf{i} . The feature is of the type described in Sect. 5 used for building the bounding box model, which is the histogram of the spatiotemporal HSV-edgel local descriptor. In order to limit the influence of foreground segmentation errors, interest points that contain large amounts of background are not considered.

When two images I and J are compared, an initial set of correspondences are nominated. The merit of a potential match ($\mathbf{i} \rightarrow \mathbf{j}$) is evaluated using Eq. 13. Inverse matching is used to ensure consistency of the correspondences. For each interest point \mathbf{i} in image I , the most likely correspondence \mathbf{i}' in image J is determined. If the distance between the signatures

of \mathbf{i} and \mathbf{i}' is below a threshold, then the most likely interest point \mathbf{i}'' in image I corresponding to point \mathbf{i}' is determined. If the Euclidean distance between \mathbf{i} and \mathbf{i}'' is smaller than a threshold, then the correspondence ($\mathbf{i} \rightarrow \mathbf{i}'$) is accepted.

A final validation stage is used to prune the initial correspondences. The support regions for corresponding interest points are expanded iteratively in the vertical direction. Again a feature vector for each expanded region is computed and compared with the distance (Eq. 13). The process continues for a fixed number of iterations or until there is too much overlap with the background. If the distance between the two signatures remains below a fixed threshold, the correspondence is accepted into the final set of correspondences.

The score given to the match between images I and J is based on the cardinality of the final set of correspondences. Two examples of this matching process are shown in Fig. 6. The efficacy of this matching algorithm is evaluated in Sect. 11.

9 Parts-based modeling via model fitting

In contrast to the interest operator algorithm, here it is considered a model-based approach that generates a correspondence between different body parts such as the head, arms, legs and torso. In other words, this allows to match the torso of an individual in one scene with the torso in the

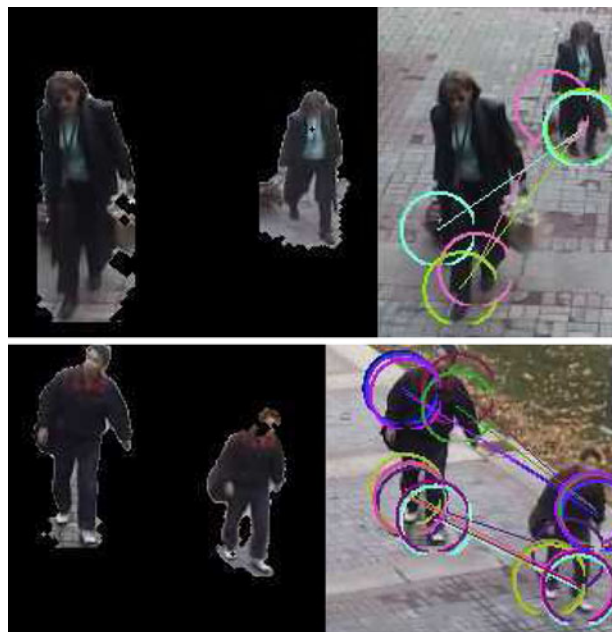


Fig. 6 Interest operator matching. Two examples of the interest operator matching algorithm. On the left are two images that are to be compared. On the right are the identified correspondences and associated circular support regions

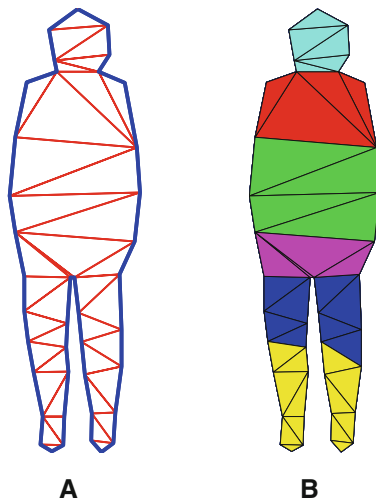


Fig. 7 Decomposable triangulated person model. **a** An example of a decomposable *triangulated* graph used as a person model. The *solid* (blue) edges correspond to the boundary of the person while the *light* (red) edges are interior edges. Note that only models without arms are considered, mostly because the individuals in the dataset have their arms next to their torso. **b** Partitioning of the person into body parts used for generating signatures for comparison (color figure online)

second scene and so on. This presents a significant challenge as the relative location of arms, legs and torso of an individual varies from one scene to another. This is addressed by using a model-based top-down segmentation of an individual in a scene where the different parts are accurately localized. This segmentation is used to establish the correspondence between the parts in two scenes, which facilitates a comparison of their appearances.

A decomposable triangulated graph is used here as the method for model fitting to people. See Fig. 7 for an example. Several researchers have used decomposable triangulated graphs to represent deformable shapes (Amit and Kong 1996; Song et al. 2003; Felzenszwalb 2005; Zhang et al. 2003; Doretto and Soatto 2006; Wu et al. 2008). These graphs are a collection of cliques of size three and have a perfect elimination order for their vertices, i.e., there exists an elimination order for all vertices such that (1) each eliminated vertex belongs only to one triangle, and (2) a decomposable triangulated graph results from eliminating the vertex. As these graphs support a perfect elimination order, the model optimization can be efficiently done by using a dynamic programming algorithm (Amit and Kong 1996).

The decomposable triangulated graph is used for modeling and segmenting people in a scene using an energy minimization approach that is now described. The starting point for the model-fitting algorithm is the bounding box of the person of interest. Let the model be a decomposable triangulated graph T made of a finite number of triangles $\{T_i\}$. The problem is to estimate a function g that maps the model to the image domain such that the consistency of the

model with salient image features is maximized, and deformations of the underlying model are minimized. This function g is restricted to being a piecewise affine map (Felzenszwalb 2005), where the deformation of each triangle $g_i(T_i)$ in the model is an affine transformation. The energy functional to be minimized $E(g, I)$ can then be written as a sum of costs, with one set of costs for each triangle in the model. Specifically,

$$E(g, I) = \sum_i E_i(g_i, I) = \sum_i E_i^{data}(g_i, I) + E_i^{shape}(g_i) \quad (26)$$

where I denotes the underlying image features. The data terms pull the model towards the salient image features, whereas the shape terms penalize large deformations of the model.

The shape and data costs for each triangle in the model are now formulated. The shape cost for each triangle is defined in terms of the decomposition of its affine transformation into the product of a rotation R and a scale-shear matrix S , defined as

$$A \doteq \begin{bmatrix} \cos \psi & -\sin \psi \\ \sin \psi & \cos \psi \end{bmatrix} \begin{bmatrix} s_x & s_h \\ s_h & s_y \end{bmatrix} \doteq R(\psi)S. \quad (27)$$

The rotation matrix R is the projection of A onto the special orthogonal group $\mathbb{SO}(2)$ (Ma et al. 2004). S and R can be computed in closed form via singular value decomposition. The shape term is then defined as

$$E^{shape} = \log\left(\frac{\lambda_1}{\lambda_2}\right)^2 + \log(1 + s_h)^2, \quad (28)$$

where λ_1 and λ_2 are the eigenvalues of the scale-shear matrix. The first term is the log-anisotropy term and penalizes changes in height-width ratio (Bookstein 1986; Felzenszwalb 2005), while the second term penalizes shear.

The data cost in the energy functional attracts the model to salient image features. Note that the decomposable triangulated graph has both boundary edges and interior edges, and that the data costs are defined only for boundary edges. The data cost for all interior edges is zero. Two complementary sets of image features are used to define the data cost: salient edges in the image which are detected using Canny (1986), and the foreground mask obtained from the foreground-background separation (Gheissari et al. 2006). Note that the model fitting approach is less sensitive to the presence of spurious edges than to missing ones, hence a combination of Canny edges and spatio-temporal foreground mask to compute the data cost is used. After the extraction of edges a Euclidean distance transform is applied to obtain the edge feature image L . The edge cost measures how far a boundary triangle edge is from Canny edges. The average edge feature value along the sampled triangle edge ε is used as the edge cost, i.e.,

$$E^{edge} = \frac{1}{|\varepsilon|} \int_{\mathbf{x} \in \varepsilon} L(\mathbf{x}) dx. \tag{29}$$

The foreground cost measures consistency of the model with the foreground mask and is defined by the relative number of foreground pixels in a window on either side of the boundary triangle edge

$$E^{fg} = 1 - \left| \frac{N_1^{fg}}{N_1} - \frac{N_2^{fg}}{N_2} \right|, \tag{30}$$

where N_1^{fg} and N_1 are the number of foreground pixels and total number of pixels on one side of the window. N_2^{fg} and N_2 are similarly defined for the other side. Note that this term is small when the boundary edge is along the foreground mask.

The dynamic-programming algorithm for computing the optimal deformation of the model is now described. The problem is to find g that maps the vertices of the model to image locations, and that minimizes the energy functional in Eq. 26. The dynamic programming approach does an exhaustive search of the candidate locations to find the global optimum. The search space for the candidate locations of the vertices of the model is restricted to the boundary of the foreground mask and Canny edges. Since the triangulated model used here has a perfect elimination order and the cost defined in Eq. 26 is extensible, a serial dynamic-programming algorithm (Amit and Kong 1996; Felzenszwalb 2005) can be used for optimization. At each iteration of the algorithm, the perfect elimination order is used to eliminate one vertex from the model, and its optimal location is encoded in terms of its two adjacent vertices. This process is repeated until all vertices are eliminated. The final location of all vertices in the model is computed by standard backtracking. Figure 8 shows the results for two representative cases.

Once the model fitting is done, the appearance model for an individual is computed as follows. The individual is partitioned into salient body parts using the fitted model as illustrated in Fig. 7. As an example, consider how the signature is generated for the upper torso. Using all the triangles that correspond to the upper torso (colored red in Fig. 7) the appearance and structure is described by a histogram of the same type that is used in the bounding box model of Sect. 5, which is based on the spatiotemporal HSV-edgel local descriptor. Similarly, the signatures for all the other body parts are computed and compared using Eq. 13.

10 Shape and appearance context modeling

This section introduces an extension of the appearance context model. Like in Sect. 6, the model aims at describing



Fig. 8 Model fitting. This figure illustrates two examples of fitting the decomposable triangulated model to individuals. The cropped image, edge feature image, foreground mask, and fitting results are shown from left to right. The (green) dots show the candidate locations for the model points. Observe that the model fits well to the individuals despite the presence of bags, shadows, additional interior edges due to different clothing etc (color figure online)

the spatial occurrence of appearance labels. However, this is done with respect to specific parts of the body of an individual. Doing so should make the appearance model more distinctive, as highlighted by the following example.

Let us examine the appearance labeled image of a person as it is depicted in Fig. 9(right), and denote with D_f and D_h the parts pertaining to the face and hand regions respectively. Notice that these are the only regions that have been assigned the label $a_1 \in \mathcal{A}$. By using the notation introduced in Sect. 6.1, D_{a_1} can be expressed as $D_{a_1} = D_f \cup D_h$. Notice also that the region of the torso, arm, and hair has been assigned the label $a_2 \in \mathcal{A}$. Given a_1 and a_2 the appearance context model can be divided in two components as follows

$$\Theta(a_2, a_1, p) = \frac{D_f}{D_{a_1}} E[h(a_2, p) | D_f] + \frac{D_h}{D_{a_1}} E[h(a_2, p) | D_h] \doteq h_f(p) + h_h(p). \tag{31}$$

The quantity $h_f(p)$ is an histogram that represents the occurrence of the label a_2 in regions of the plane that with respect to the face region are defined by p . Similarly $h_h(p)$ represents the occurrence of a_2 in regions of the plane that with respect to the hand region are defined by p . Fig. 9

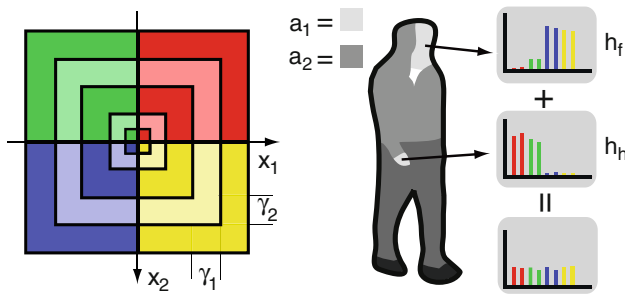


Fig. 9 L-shaped partition and appearance context averaging effect. From left to right. Sketch of the L-shaped plane partition used in Sect. 6 ($\gamma_1 = 4N/l$, $\gamma_2 = 4M/l$), and Sect. 10 ($\gamma_1 = 4Nd/\tau$, $\gamma_2 = 4Md/\tau$). Illustration of the averaging effect when appearance context descriptors are pooled from the entire body of an individual. As pointed out in Sect. 10, the sum $h_f + h_h$ is less informative than h_f and h_h alone. The notion of body parts maintains these contributions separate, which enhance the specificity of the descriptor

sketches a color coded version of $h_f(p)$, $h_h(p)$, and their sum, which matches the color coding of the plane partition \mathcal{P} with L-shapes depicted on the left. $h_f(p)$ highlights that with respect to the face, a_2 is mostly present in the blue and yellow quadrants of \mathcal{P} . $h_h(p)$ highlights that with respect to the hand, a_2 is mostly present in the red and green quadrants of \mathcal{P} . Finally, $h_f(p) + h_h(p)$ shows that with respect to where a_1 is located, a_2 is roughly uniformly distributed over all the quadrants.

The toy example points out that averaging h_f and h_h has caused a loss of information, and therefore descriptive power. From an information theoretic point of view, two unimodal distributions have merged to obtain an almost uniform distribution, with a significant increase of entropy. Thus, this observation is suggesting that *if it was possible to identify the parts of a given person, it would be more descriptive to capture the spatial occurrence of the appearance labels with respect to each of the parts rather than to each of the appearance labels.*

Following the idea just expressed, it arises the need to be able to identify body parts. This is done by introducing an appearance based approach that given the bounding box image I containing an individual, it computes a shape labeled image, where each label always attempts to identify the same part of the body. The procedure to compute it is inspired by the idea of shape context (Belongie et al. 2002; Mori and Malik 2006). Given the image I , this is processed according to a local descriptor operator that at every pixel \mathbf{x} computes a d -dimensional descriptor $\omega(\mathbf{x})$. The experimental Sect. 11 reports results with different choices of operators for computing ω . It was found that a fast and reliable option is such that $\omega(\mathbf{x}) = \text{HOG}(\nabla I_L, \mathbf{x})$, where I_L is the L channel of the Lab color space of I . From ω , at every pixel a form of *shape context descriptor* $\psi \in \mathbb{R}^\tau$ is computed according to the following expression

$$\psi(\mathbf{x}) \doteq (E[\omega|p_1(\mathbf{x})]; \dots; E[\omega|p_{\tau/d}(\mathbf{x})]). \quad (32)$$

Here $\{p_1, \dots, p_{\tau/d}\}$ indicates a plane partition of the same kind used for the appearance context descriptor, but with τ/d L-shaped regions rather than l (see Fig. 9). The shape context descriptor ψ is then vector quantized according to a quantization (labeling) function $q: \mathbb{R}^\tau \rightarrow \mathcal{S}$, with quantization levels defined by a *shape dictionary* $\mathcal{S} = \{s_1, \dots, s_n\}$, made of *shape labels* learnt off-line with unsupervised clustering. This produces the *shape labeled image* $S(\mathbf{x}) \doteq q \circ \psi(\mathbf{x})$. Neighboring pixels have similar shape context descriptors. Therefore, the quantization process produces a “piecewise-like” segmentation of the image into regions $D_{s_i} = \{\mathbf{x} | S(\mathbf{x}) = s_i\}$, which are meant to always identify the same region/part of the object of interest (see Fig. 3).

Given the image I inside the bounding box containing an individual, let A be its appearance labeled image, and let S (defined over Λ) be its shape labeled image, then the *shape and appearance context* model of I is the occurrence Θ computed over S and A , which is an $m \times n \times l$ matrix. Similarly to the appearance context model, the information carried by the descriptor in Eq. 14 is included in this one as well. Note that, as a welcome side effect concerning the computational complexity, because of the piecewise type of segmentation of the image, C_S ends up being very small, and the occurrence computation becomes much faster. Finally, as for the appearance context model, matching two shape and appearance context models is done via L_1 norm.

The shape and appearance context enjoys the same invariance properties of the appearance context. However, it should be noted that translation invariance is lost if the same parts of an object are labeled differently in different images, e.g. an arm labeled as a leg. This means that trivially using a fixed mask to roughly identify object parts under pose and viewpoint changes as well as object bounding box imprecisions, can significantly decrease the performance of the model. The decomposition into parts with the shape labeled image tries to compensate exactly those variations (see Fig. 3).

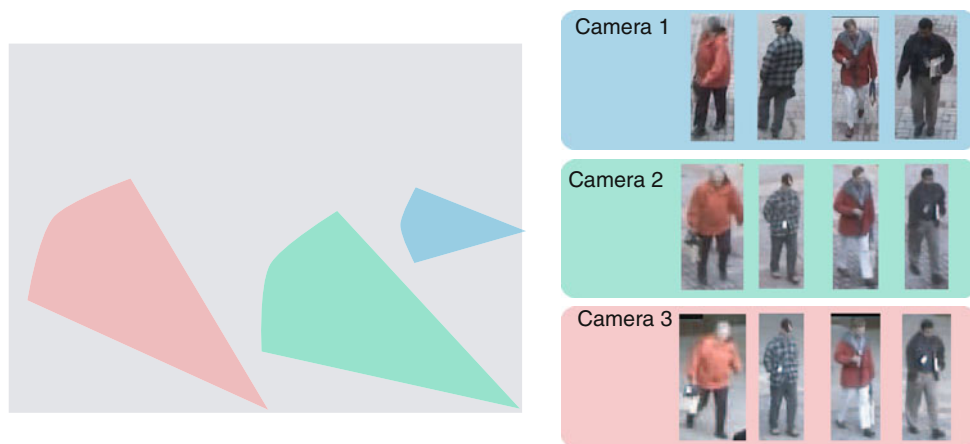
11 Experiments

This section describes a series of experiments that have been carried out by testing the reidentification, or matching capabilities, of the models introduced, on a dataset that was acquired by a camera network deployed in typical outdoor surveillance settings.

11.1 Camera network setup and dataset

In order to test the proposed models it has been used a video dataset acquired by a network consisting of three cameras.

Fig. 10 Camera network setup. *Left* the layout of cameras used for collecting the data for the experiments. Camera 2 is placed roughly at twice the height of Cameras 1 and 3. *Right* the representative samples of key frames for four individuals from all three camera views



The approximate layout is shown in Fig. 10, which shows that the cameras have non-overlapping field of views. The subjects were recorded while entering a corporate campus and were in no way coached or rehearsed.

For every person in each view, two to four key frames were selected to form the measurements of the identity of that individual. For every keyframe, a tight bounding box of the person is also recorded, the typical size of which is about 80×170 pixels. Such information is provided automatically by a multi-camera multi-target tracker (Tu et al. 2007). Some of the algorithms have been tested on a baseline dataset comprising 44 different identities, to which 99 more individuals were later added. Some images of the dataset are shown in Fig. 3, Fig. 10, and Fig. 11 from which the different poses in different views can be noticed.

11.2 Reidentification evaluation

Each image in the dataset is indexed as I_{ic} where i encodes the person id, t encodes the time or key frame number, and c encodes the camera view. All the reidentification models are evaluated in the following manner:

- Each image I_{ic} is compared against the set of images $\{I_{ic'}\}$ such that $c' \neq c$.
- For each person/camera combination, the maximum ranking true match for all its key frames is determined.
- The number of times that a maximum ranking true match is higher than a given value is then tabulated. This produces the so called cumulative match characteristic (CMC) curves (Moon and Phillips 2001), which tell the rate at which the correct match is within the top k ranked identities. Here the tabulation is done for k that varies from 1 to 20.

This evaluation scheme is analogous to a standard surveillance scenario where an operator would query a person reidentification system with multiple images of the same individual captured over a short period of time from a

particular camera. Any hits from these queries would result in a success.

11.3 Models based on the spatiotemporal HSV-edgel descriptor

The bounding box model (Sect. 5), the interest operator model (Sect. 8) and the model fitting approach (Sect. 9) share a couple of important characteristics. The first one is that they are all based on the spatiotemporal HSV-edgel descriptor. The second one is that they do not require any form of training before using them. Here they are evaluated together with the dataset and procedure explained above.

Figure 11 reports the performance of the three algorithms. The model fitting approach has the highest matching rate with approximately 60% of the queries achieving a top ranking true match and over 90% of the queries generating a true match in the top ten. The interest operator method achieves a top ranking true match 25% of the time and a true match is found in the top ten 65% of the time. It should be noted that the performance of this approach may improve with higher image resolution. The performance of the baseline bounding box approach is comparable to that of the interest operator approach. Figure 11 shows the top ranking images for a number of queries using the model fitting algorithm.

11.4 Training and testing

The bag-of-features model (Sect. 5), the appearance context model (Sect. 6), and the shape and appearance context model (Sect. 10) require a training phase in order to learn the appearance dictionary as well as the shape dictionary. This is done by applying k -means clustering¹² to the two sets of local descriptors computed on the training set

¹² Note that other clustering options could be explored, like (Jurie and Triggs 2005; Philbin et al. 2008).

Fig. 11 Reidentification matches. *Top* ten matches using the model fitting parts-based appearance model. The query image is shown in the *left column*, and the remaining *columns* are the *top* matches ordered from *left to right*. A *box* is used to highlight when a match corresponds to the query. The *third row* shows an example where the correct match is not present in the *top ten* matches



according to (32), which is the HOG Log-RGB local descriptor, and according to (2), which is the shape context local descriptor. For training, 30% of the images of the individuals are randomly selected. The rest of the data set is used for testing.

An appearance label of the appearance dictionary is assigned to a local descriptor $\varphi(\mathbf{x})$ using the L_1 norm. On the other hand, a shape label of the shape dictionary is assigned to a shape context descriptor $\psi(\mathbf{x})$ using the χ^2 distance¹³, because it was found to provide a higher part recognition performance. For the bag-of-features model, matching is done by comparing descriptors such as (14) via the intersection distance (13). Matching using the appearance, or the shape and appearance context is done by comparing the models with the L_1 norm.

11.5 Bag-of-features model

The bag-of-features model was tested extensively in order to chose the local descriptor that maximizes the

¹³ Note that ψ is a concatenation of histograms.

performance. Figure 13(left) shows a comparison between models where the local descriptor Φ is a color transformation combined with different types of quantization. In particular the Lab, RGB, and HSV color spaces are tested. Quantization along the three axes is done according to the number of bins reported in the figure. The suffixes “pls” and “mul” indicate whether channel quantization is performed independently or jointly, respectively. The experiment indicates that the Lab and the RGB color spaces perform better.

Figure 13(right) shows a test where the local descriptor operator Φ is the linear filter (LF) bank used in (Winn et al. 2005), which was proved to be very effective for category recognition. The CMC curves show that the bank of filters improves versus simple color quantization. In particular, several dimensions of the appearance dictionary are tested, and for the dataset used here the matching rate is maximized when the cardinality of \mathcal{A} is around $m = 60$.

Figure 14(left) shows that the HOG operator defined in Sect. 3 was tested with different color spaces, and in particular, with the L channel of the Lab space, the Log-RGB space, the RGB space, the logarithm of the color-ratio

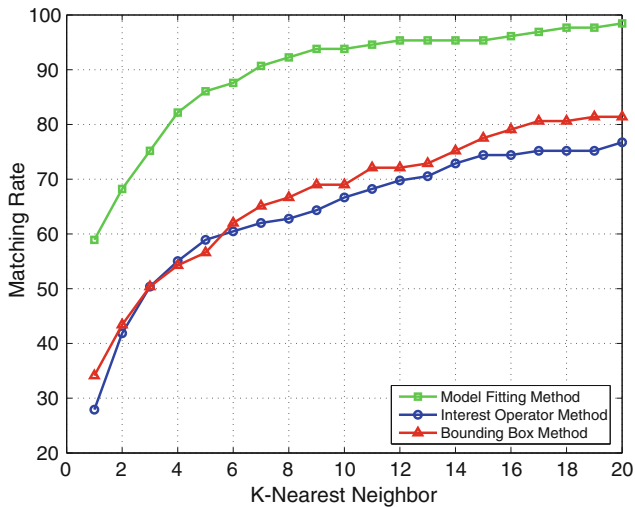


Fig. 12 Comparison between interest operator, model fitting, and bounding box models. The percent of correct detection rate is plotted (matching rate) versus the number of matches considered (k -nearest neighbors). The model fitting approach significantly outperforms the other two. The performance of the bounding box approach and the interest point approach is comparable

space, and the “C” invariance of the color invariant space (Geusebroek et al. 2001). The L channel and the Log-RGB color space are the best options, and the HOG with either of the choices almost double the performance with respect to the linear filter bank. From this experiment the HOG Log-RGB local descriptor was selected as the operator of choice for computing the appearance labeled image \mathcal{A} .

Figure 15 summarizes several experiments where the parameters defining the HOG Log-RGB operator are varied within a range. This information is useful for gaining insights about parameter sensitivity, and about how to set them. Four parameters are considered, one at a time, namely the number of quantizing orientations ℓ of the HOG, the amount of prior Gaussian smoothing to be applied to an image, identified by the standard deviation σ , the patch size w over which the HOG is computed, and the dimensionality of the appearance dictionary m . A good

operating point was found to be for $\ell = 16$, $\sigma = 0$, $w = 11$, and $m = 60$.

11.6 Shape and appearance context models

Since the bag-of-features model does not capture the spatial relationship between appearance labels, it is expected that doing so through either the appearance context, or the (parts-based) shape and appearance context model will increase distinctiveness and matching performance. Figure 16(left) shows a comparison between different shape and appearance context models, where different choices of the operator Φ are considered. The HOG is tested with three different color spaces (Log-RGB, RGB, and logarithm of the color-ratio space). The linear filter bank (LF) used in (Winn et al. 2005) is also tested alone, and in combination with the HOG of the L channel of the Lab color space. From the CMC curves it is noticeable how the configurations that include the HOG significantly outperform the ones with the LF bank only. In particular, the HOG in the Log-RGB color space still gives the best CMC curve.

One important component of the shape and appearance context model is the estimation of the body parts, which is done through the shape labeled image. This entails the computation and quantization of the shape context descriptors (Eq. 32). Figure 14(right) shows the CMC curves for two descriptors ω that were tested. One is the HOG computed on the L channel of the Lab color space of the image, obtained with $d = 8$ quantizing orientations, a patch size $w = 11$, and a shape dictionary of dimension $n = 18$. The second (Edge), is such that ω is a Canny edge detector, and ψ does an edge pixels count on the given partition element region. Both of the approaches work well, with a slight advantage to the first one, which was picked to be the default choice.

The implementation of the shape and appearance context descriptor in C++ can reach a speed of about 10fps on an image region of 250×100 pixels on a high-end PC. The implementation in the discrete domain of lines 5–7 of

Fig. 13 Bag-of-features color spaces and linear filters. Matching comparison using different color spaces and quantization schemes (left). Matching comparison of a linear filter bank against appearance dictionary size (right) (color figure online)

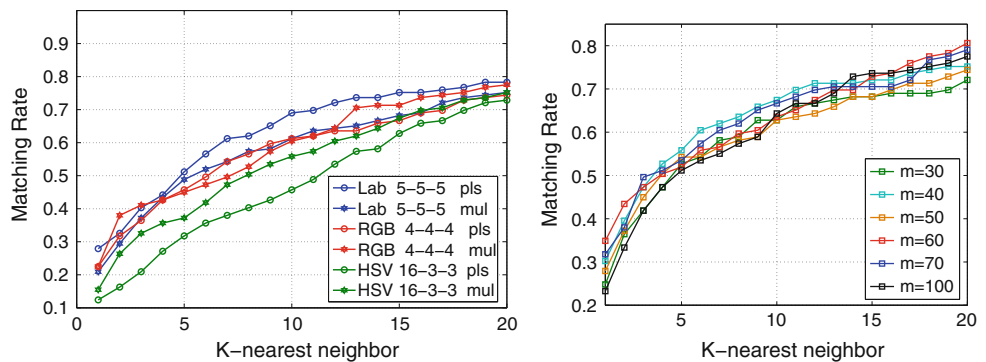


Fig. 14 HOG with different color spaces and shape descriptors. Application of the HOG to different color spaces (left). Comparison of different shape descriptors (right) (color figure online)

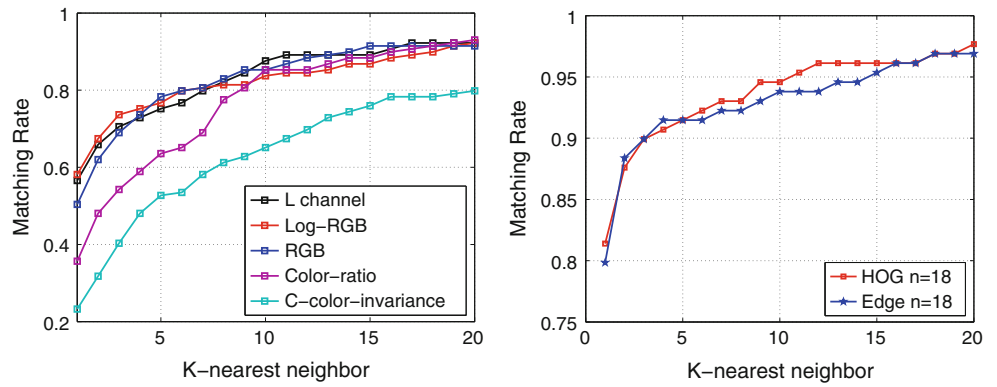
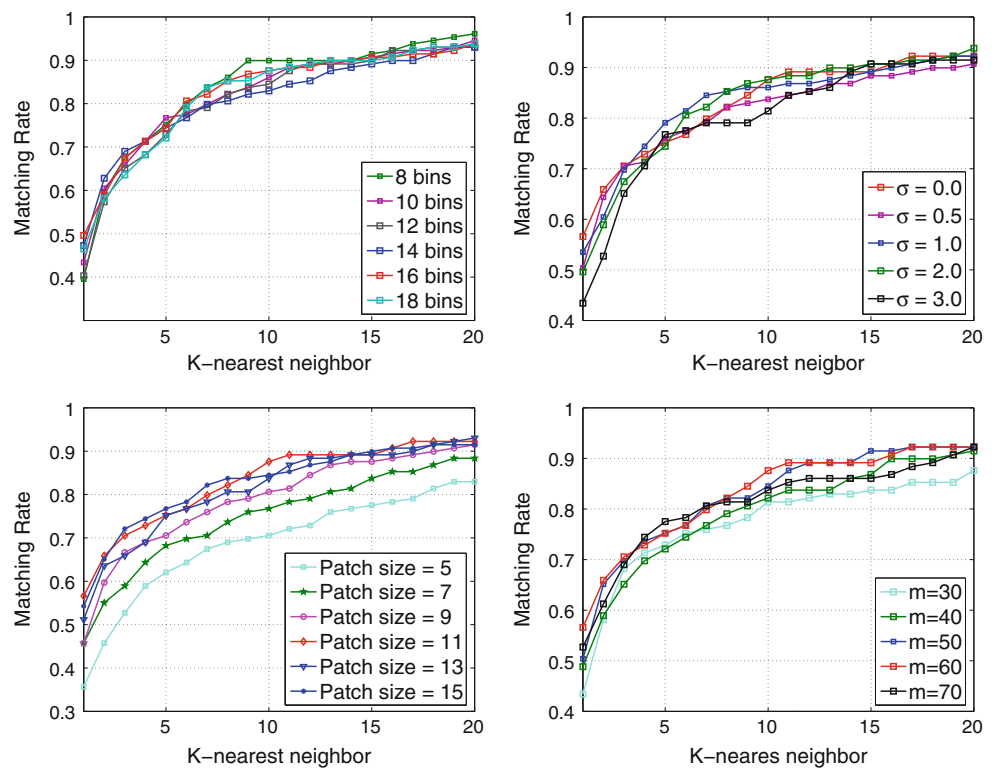


Fig. 15 Bag-of-features. HOG Log-RGB. Matching comparison by varying one of the operator parameters at a time, such as the quantizing orientations ℓ (top-left), the image smoothing σ (top-right), the patch size w (bottom-left), and the appearance dictionary size m (bottom-right)



Algorithm 1 requires particular care, and the interested reader can consult (Doretto and Wang 2007) for details.

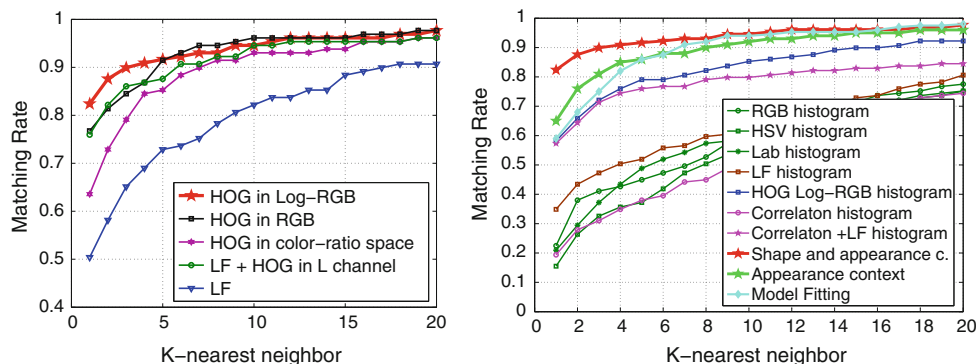
11.7 Comparison summary

Figure 16 (right) compares the matching performance of several approaches. It includes the CMC curves of many bag-of-features models corresponding to different choices of local descriptor operator Φ . In particular, they include the best results obtained where Φ is either a simple color transformation and quantization (to spaces such as RGB, HSV, Lab), or the linear filter (LF) bank used in (Winn et al. 2005), or the HOG Log-RGB operator. Included in the figure are also the CMC curves of an appearance model

that computes the histogram of the so called correlatons (Savarese et al. 2006). This can be seen as an approach related to the appearance context model in that it attempts to capture the co-occurrence of appearance labels¹⁴. The same histogram of correlatons is also used jointly with the histogram of linear filters, resulting in a significant performance increase. Moreover, Figure 16(right) includes the CMC plots of the appearance context model, the shape and appearance context model, and of the model fitting

¹⁴ Note that the approach in (Savarese et al. 2006) was originally designed for doing inter-category object recognition. Here it has been tested outside of its natural domain.

Fig. 16 Shape and appearance context and comparison summary. *Left* matching comparison between several approaches for computing appearance labels. Approaches using HOG outperform the ones that do not. *Right* matching comparison summary of several appearance models. Approaches that capture appearance spatial relationships outperform those that do not



approach, which was the best among the frameworks based on the spatiotemporal HSV-edgel descriptor.

The overall picture clearly indicates that capturing a description of the spatial distribution of the appearance results in more distinctive models for reidentification. This emerges from the difference between models that are based on the bag-of-features paradigm and the others. In particular, the model fitting and appearance context models show comparable results, with a slight advantage of the latter. However, it should be noted that one of the advantages of the former is that it does not require any form of off-line training. Finally, the model with the highest performance has resulted to be the shape and appearance context model with an 82% of rank-1 matching rate.

12 Conclusions

This work shows how the person reidentification problem is framed within the context of camera networks, typically deployed to address visual surveillance needs. In particular, it describes how solving this problem enables tracking people beyond blind gaps. This is a much needed feature in order to perform the automatic analysis of video archives acquired by surveillance apparatus. Rather than considering passive biometric cues, or camera calibration and kinematic behavior assumptions of the people imaged by the camera network, this work focusses on reidentifying people based on their whole body appearance. The representation of such cue can be built from one image (single-shot approach) of a person (like for the bag-of-features model, the appearance context, and the shape and appearance context model), or from many images (multiple-shot approach) (like the bounding box method, the interest operator method, or the model fitting approach). The latter methods allow to better “average out” undesired appearance variations due to the unpredictable appearance changes of clothing. The single-shot models introduced can be extended to multiple-shots by simply accumulating the visual information of many images.

The appearance of an individual may inherently be represented as a whole, leading to holistic models (like the bounding box, the bag-of-features, and the appearance context), or can be divided into regions at the outset, leading to parts-based models (like the interest operator, the model fitting, and the shape and appearance context). By nature, parts-based models describe a form of spatial configuration of the appearance. Conversely, that may not be the case for holistic models, like the bounding box, and the bag-of-features. However, when they can be made to capture the appearance spatial distribution, like in the appearance context, their matching rate significantly increases.

The highest performance boosts have been observed by switching from a holistic to a parts-based approach. This consideration is not in accordance with the general trend in appearance based person detection. There, the most robust approaches have proven to be the ones that use a holistic representation. This is indicative of how different these problems are, even though in person detection as well as in person reidentification people are subject to the same imaging conditions.

From an algorithmic standpoint, when training is not feasible, the model fitting approach represents a viable option for reidentification. On the other hand, the appearance context and the shape and appearance context models constitute an extension when off-line learning can be afforded. Also, the generalization of the integral representation to operate on generalized rectangular domains has allowed the derivation of very fast procedures to compute these models in realtime.

Finally, future directions might include the following aspects. The first one is concerned with the organization of multiple models into a hierarchy addressing the issue of reliable matching by narrowing down the search space at every level of the hierarchy. The second is the organization of such hierarchy to achieve complete invariance with respect to pose and viewpoint variations. The third is the use of such hierarchy for enabling efficient indexing schemes for database queries. This would be valuable for

forensic applications involving large numbers of cameras capturing imagery over extended periods of time.

Acknowledgments The authors are grateful to Xiaogang Wang, Niloofar Gheissari, and Richard Hartley for their valuable contributions to the development of the approaches outlined in this manuscript. This report was prepared by GE GRC as an account of work sponsored by Lockheed Martin Corporation. Information contained in this report constitutes technical information which is the property of Lockheed Martin Corporation. Neither GE nor Lockheed Martin Corporation, nor any person acting on behalf of either; a. Makes any warranty or representation, expressed or implied, with respect to the use of any information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or b. Assume any liabilities with respect to the use of, or for damages resulting from the use of, any information, apparatus, method or process disclosed in this report.

Appendix

This appendix first introduces some notation and then gives the proofs of Theorem 1, and Theorem 2. The variable \mathbf{v} can be interpreted as partitioning \mathbb{R}^k into its 2^k open orthants $O_{\mathbf{v}} \doteq \{\mathbf{x} \in \mathbb{R}^k | x_i > 0 \text{ if } v_i = 1, \text{ or } x_i < 0 \text{ if } v_i = 0, i = 1, \dots, k\}$ (see Fig. 5 for an example where $k = 2$). Let us introduce the notation $O_{\mathbf{v}}(\mathbf{x}) \doteq \{\mathbf{x} + \mathbf{y} | \mathbf{y} \in O_{\mathbf{v}}\}$, and also a function $\beta_{\mathbf{x}}(\mathbf{v}) : \mathbb{B}^k \rightarrow \mathbb{B}$, such that: (1) $\beta_{\mathbf{x}}(\mathbf{v}) = 1$ if \mathbf{x} is an adherent point¹⁵ for the open set $D \cap O_{\mathbf{v}}(\mathbf{x})$; (2) $\beta_{\mathbf{x}}(\mathbf{v}) = 0$ otherwise (see Fig. 5 for an example where $k = 2$). It is trivial to prove that: I) $\mathbf{x} \in D \setminus \partial D \iff \beta_{\mathbf{x}}(\mathbf{v}) = 1 \forall \mathbf{v}$; II) $\mathbf{x} \in \partial D \iff \beta_{\mathbf{x}}(\mathbf{v}) = 0 \forall \mathbf{v}$. Finally, if \mathbf{v}_j represents j out of the k components of \mathbf{v} , and if \mathbf{v}_{k-j} represents the remaining $k - j$ components, then *edges* and *corners* of the boundary ∂D are defined as follows: A point $\mathbf{x} \in \partial D$ lays on an *edge* if there exist j components of \mathbf{v} , with $1 \leq j \leq k - 1$, such that $\beta_{\mathbf{x}}(\mathbf{v})$ does not depend on \mathbf{v}_{k-j} , i.e. $\beta_{\mathbf{x}}(\mathbf{v}) = \beta_{\mathbf{x}}(\mathbf{v}_j), \forall \mathbf{v}$. If \mathbf{x} does not lay on an edge, it is a *corner*. The set of corners of D is indicated with $\nabla \cdot D$.

Proof of Theorem 5 Let $\{u_{i,1}, u_{i,2}, \dots | u_{i,j} \in \mathbb{R}, u_{i,j} < u_{i,j+1}, i = 1, \dots, k\}$, be the set of points along $\{x_i\}$, such that D is made of portions of hyperplanes passing through these points. Fig. 5 illustrates an example for $k = 2$. The intersection of the hyperplanes with D defines a partition $D \doteq \bigcup_i R_i$ into rectangular regions $\{R_i\}$, which allows to write $\int_D f(\mathbf{x}) d\mathbf{x} = \sum_i \int_{R_i} f(\mathbf{x}) d\mathbf{x}$, and apply Eq. 17 to each term of the summation. By rearranging the terms, and using the function $\beta_{\mathbf{x}}(\mathbf{v})$, the integral can be rewritten as $\sum_{\mathbf{x} \in \mathcal{D}} \alpha_D(\mathbf{x}) F(\mathbf{x})$, where \mathcal{D} is the set of all the corner points of the regions $\{R_i\}$ (note that $\nabla \cdot D \subseteq \mathcal{D}$), and $\alpha_D(\mathbf{x}) \doteq \sum_{\mathbf{v}} (-1)^{v^T \mathbf{1}} \beta_{\mathbf{x}}(\mathbf{v})$. Now recall that if $\mathbf{x} \in \mathcal{D} \setminus \partial D$,

¹⁵ A point \mathbf{x} is an adherent point for an open set B , if every open set containing \mathbf{x} contains at least one point of B . A point \mathbf{x} is an adherent point for B if and only if \mathbf{x} is in the closure of B .

then $\beta_{\mathbf{x}}(\mathbf{v}) = 1$, which implies $\alpha_D(\mathbf{x}) = 0$. On the other hand, if \mathbf{x} is on an edge, then one can write $\alpha_D(\mathbf{x}) = \sum_{\mathbf{v}_j} (-1)^{v_j^T \mathbf{1}} \beta_{\mathbf{x}}(\mathbf{v}_j) \sum_{\mathbf{v}_{k-j}} (-1)^{v_{k-j}^T \mathbf{1}} = 0$, and Eq. 23 is valid. When \mathbf{x} is a corner described by $\beta_{\mathbf{x}}$, one should proceed with a direct computation of $\alpha_D(\mathbf{x})$. For $k = 2$, $\alpha_D(\mathbf{x})$ is different then zero only for the 10 cases depicted in Fig. 5, in which it assumes the values indicated. \square

Proof of Theorem 2 According to Eq. 15, Θ can be computed by using Eq. 21, and subsequently applying Theorem 1, giving

$$\Theta(a, s, p) = |D_s|^{-1} \sum_{\mathbf{x} \in \nabla \cdot D_s} \alpha_{D_s}(\mathbf{x}) H(a, p(\mathbf{x})), \tag{33}$$

where

$$H(a, p(\mathbf{x})) = \int_{-\infty}^{\mathbf{x}} h(a, p(\mathbf{u})) d\mathbf{u}. \tag{34}$$

Now note that $h(a, p(\mathbf{u}))$ can be computed through the integral histogram of A

$$F(a, \mathbf{z}) = \int_{-\infty}^{\mathbf{z}} e \circ A(\mathbf{v}) d\mathbf{v}, \tag{35}$$

and by applying Theorem 1, resulting in

$$h(a, p(\mathbf{u})) = |p(\mathbf{u})|^{-1} \sum_{\mathbf{z} \in \nabla \cdot p(\mathbf{u})} \alpha_{p(\mathbf{u})}(\mathbf{z}) F(a, \mathbf{z}). \tag{36}$$

By combining Eq. 36 with Eq. 34 follows that

$$H(a, p(\mathbf{x})) = \int_{-\infty}^{\mathbf{x}} |p(\mathbf{u})|^{-1} \sum_{\mathbf{z} \in \nabla \cdot p(\mathbf{u})} \alpha_{p(\mathbf{u})}(\mathbf{z}) F(a, \mathbf{z}) d\mathbf{u}. \tag{37}$$

From the definition of $p(\mathbf{u})$, it follows that $|p(\mathbf{u})| = |p|$, and $\nabla \cdot p(\mathbf{u}) = \{\mathbf{u} + \mathbf{y} | \mathbf{y} \in \nabla \cdot p\}$, and also that $\alpha_{p(\mathbf{u})}(\mathbf{u} + \mathbf{y}) = \alpha_p(\mathbf{y})$. Therefore, after the change of variable $\mathbf{z} = \mathbf{u} + \mathbf{y}$, in Eq. 37 it is possible to switch the order between the integral and the summation, yielding

$$H(a, p(\mathbf{x})) = |p|^{-1} \sum_{\mathbf{y} \in \nabla \cdot p} \alpha_p(\mathbf{y}) \int_{-\infty}^{\mathbf{x}} F(a, \mathbf{u} + \mathbf{y}) d\mathbf{u}. \tag{38}$$

By substituting Eq. 38 into Eq. 33, and by taking into account Eq. 35 follows that Eqs. 24 and 25 are proved. \square

References

Amit Y, Kong A (1996) Graphical templates for model registration. IEEE Trans Pattern Anal Mach Intell 18(3):225–236
 Bak S, Corvee E, BrTmond F, Thonnat M (2010a) Person re-identification using spatial covariance regions of human body

- parts. In: Proceedings of IEEE international conference on video and signal based surveillance
- Bak S, Corvee E, BrTmond F, Thonnat T (2010b) Person re-identification using haar-based and dcd-based signature. In: Proceedings of the workshop on activity monitoring by multi-camera surveillance systems
- Bäumel M, Bernardin K, Fischer M, Ekenel HK (2010) Multi-pose face recognition for person retrieval in camera networks. In: Proceedings of IEEE international conference on video and signal based surveillance
- Bay H, Ess A, Tuytelaars T, Van Goo L (2008) Surf: Speeded up robust features. *Comput Vis Image Underst* 110(3):346–359
- Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. *IEEE Trans Pattern Anal Mach Intell* 24:509–522
- Bird N, Masoud O, Papanikolopoulos N, Isaacs A (2005) Detection of loitering individuals in public transportation areas. *IEEE Trans Intell Transport Syst* 6(2):167–177
- Bissacco A, Soatto S (2009) Hybrid dynamical models of human motion for the recognition of human gaits. *Int J Comput Vis* 85(1):101–114
- Blackman S, Popoli R (1999) Design and analysis of modern tracking systems. Artech House Publishers, Norwood
- Bookstein FL (1986) Size and shape spaces for landmark data in two dimensions. *Stat Sci* 1(2):181–242
- Cai Y, Huang K, Tan T (2008) Human appearance matching across multiple non-overlapping cameras. In: Proceedings of the international conference on pattern recognition
- Canny J (1986) A computational approach to edge detection. *IEEE Trans Pattern Anal Mach Intell* 8(6):679–698
- Cox JJ, Hingorani SL (1994) An efficient implementation and evaluation of reid's multiple hypothesis tracking algorithm for visual tracking. In: Proceedings of the international conference on pattern recognition
- Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 1. pp 886–893
- Damen D, Hogg D (2007) Associating people dropping off and picking up objects. In: Proceedings of the British machine vision conference
- Doretto G, Soatto S (2006) Dynamic shape and appearance models. *IEEE Trans Pattern Anal Mach Intell* 28(12):2006–2019
- Doretto G, Wang X (2007) Integral computations: a framework to compute fast region based features. Tech. Rep. 2007GRC593, GE Global Research. Visualization and Computer Vision Laboratory, Niskayuna
- Doretto G, Yao Y (2010) Region moments: fast invariant descriptors for detecting small image structures. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition
- Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2010) Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition
- Fei-Fei L, Perona P (2005) A Bayesian hierarchical model for learning natural scene categories. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2, pp 524–531
- Felzenszwalb PF (2005) Representation and detection of deformable shapes. *IEEE Trans Pattern Anal Mach Intell* 27(2):208–220
- Felzenszwalb PF, Huttenlocher D (2004) Efficient graph-based image segmentation. *Int J Comput Vis* 59(2):167–181
- Forssen PE (2007) Maximally stable colour regions for recognition and matching. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition
- Funt BV, Finlayson GD (1995) Color constant color indexing. *IEEE Trans Pattern Anal Mach Intell* 17:522–529
- Gandhi T, Trivedi MM (2007) Person tracking and reidentification: introducing panoramic appearance map (PAM) for feature representation. *Mach Vis Appl* 18(3–4):207–220
- Geusebroek J, Boomgaard R, Smeulders AWM, Geerts H (2001) Color invariance. *IEEE Trans Pattern Anal Mach Intell* 23:1338–1350
- Gheissari N, Sebastian TB, Tu PH, Rittscher J, Hartley R (2006) Person reidentification using spatiotemporal appearance. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2, pp 1528–1535
- Gray D, Tao H (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of the European conference on computer vision, pp 262–275
- Guo Y, Hsu S, Shan Y, Sawhney H, Kumar R (2005) Vehicle fingerprinting for reacquisition & tracking in videos. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2, pp 761–768
- Hamdoun O, Moutarde F, Stanculescu B, Steux B (2008) Person reidentification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In: Proceedings of the ACM/IEEE international conference distributed smart cameras
- Hu L, Wang Y, Jiang S, Huang Q, Gao W (2008) Human reappearance detection based on on-line learning. In: Proceedings of the international conference on pattern recognition
- Huang J, Kumar SR, Mitra M, Zhu WJ, Zabih R (1997) Image indexing using color correlograms. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, San Juan, pp 762–768
- Isard M, MacCormick J (2001) BraMBLe: aBayesian multiple-blob tracker. In: Proceedings of IEEE international conference on computer vision, pp 34–41
- Jaffré G, Joly P (2004) Costume: a new feature for automatic video content indexing. In: Proceedings of RIAO, pp 314–325
- Javed O, Rasheed Z, Shafique K, Shah M (2003) Tracking across multiple cameras with disjoint views. In: Proceedings of IEEE international conference on computer vision, pp 952–957
- Javed O, Shafique K, Shah M (2005) Appearance modeling for tracking in multiple non-overlapping cameras. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2, pp 26–33
- Javed O, Shafique K, Rasheed Z, Shah M (2007) Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Comput Vis Image Underst* 109:146–162
- Jurie F, Triggs B (2005) Creating efficient codebooks for visual recognition. In: Proceedings of IEEE international conference on computer vision
- Ke Y, Sukthankar R, Hebert M (2005) Efficient visual event detection using volumetric features. In: Proceedings of IEEE international conference on computer vision, vol 1, pp 166–173
- Khan SM, Shah M (2006) A multiview approach to tracking people in crowded scenes using a planar homography constraint. In: Proceedings of the European conference on computer vision, pp 133–146
- Krahnstoeber N, Tu P, Sebastian T, Perera A, Collins R (2006) Multi-view detection and tracking of travelers and luggage in mass transit environments. In: Proceeding of IEEE international workshop on performance evaluation of tracking and surveillance
- Kumar S, Hebert M (2006) Discriminative random fields. *Int J Comput Vis* 68:179–201
- Lazebnik S, Schmid C, Ponce J (2003) Affine-invariant local descriptors and neighborhood statistics for texture recognition.

- In: Proceedings of IEEE international conference on computer vision, pp 649–655
- Lin Z, Davis LS (2008) Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In: International symposium on visual computing, pp 23–34
- Lo Presti L, Sclaroff S, La Cascia M (2009) Object matching in distributed video surveillance systems by l₁-based appearance descriptors. In: Proceedings of the international conference on image analysis and processing
- Lowe D (2004) Distinctive image features from scale-invariant key points. *Int J Comput Vis* 60:91–110
- Ma X, Grimson WEL (2005) Edge-based rich representation for vehicle classification. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2, pp 1185–1192
- Ma Y, Soatto S, Kosecká J, Sastry SS (2004) An invitation to 3D vision: from images to geometric models. Springer, New York, Inc.
- Madden C, Cheng E, Piccardi M (2007) Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Mach Vis Appl* 18(3):233–247
- Makris D, Ellis TJ, Black JK (2004) Bridging the gaps between cameras. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 205–210
- Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. *IEEE Trans Pattern Anal Mach Intell* 27:1615–1630
- Mikolajczyk K, Schmid C, Zisserman A (2004) Human detection based on a probabilistic assembly of robust part detectors. In: Proceedings of the European conference on computer vision, pp 69–82
- Mikolajczyk K, Tuytelaars T, Schmid C, Zisserman A, Matas J, Schaffalitzky F, Kadir T, Van Gool L (2005) A comparison of affine region detectors. *Int J Comput Vis* 65(1–2):43–72
- Moon H, Phillips PJ (2001) Computational and performance aspects of PCA-based face-recognition algorithms. *Perception* 30(3):3003–3321
- Mori G, Malik J (2006) Recovering 3d human body configurations using shape contexts. *IEEE Trans Pattern Anal Mach Intell* 28(7):1052–1062
- Moscheni F, Bhattacharjee S, Kunt M (1998) Spatiotemporal segmentation based on region merging. *IEEE Trans Pattern Anal Mach Intell* 20(9):897–915
- Nakajima C, Pontil M, Heisele B, Poggio T (2003) Full-body person recognition system. *Pattern Recognit* 36(9):1997–2006
- Oliveira de Oliveira I, de Souza Pio JL (2009) People reidentification in a camera network. In: Proceeding of the IEEE international conference on dependable, autonomic and secure computing
- Ozcanli OC, Tamrakar A, Kimia BB, Mundy JL (2006) Augmenting shape with appearance in vehicle category recognition. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, New York, NY, USA, vol 1, pp 935–942
- Park A U Jain, Kitahara I, Kogure K, Hagita N (2006) ViSE: visual search engine using multiple networked cameras. In: Proceedings of the international conference on pattern recognition, pp 1204–1207
- Patras L, Hendriks EA, Lagendijk RL (2001) Video segmentation by MAP labeling of watershed segments. *IEEE Trans Pattern Anal Mach Intell* 23(3):326–332
- Pham TV, Worring M, Smeulders AWM (2007) A multi-camera visual surveillance system for tracking of recurrences of people. In: Proceedings of the ACM/IEEE international conference distributed smart cameras
- Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2008) Lost in quantization: improving particular object retrieval in large scale image databases. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition
- Phillips P, Flynn P, Scruggs T, Bowyer K, Chang J, Hoffman K, Marques J, Min J, Worek W (2005) Overview of the face recognition grand challenge. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 947–954
- Porikli F (2003) Inter-camera color calibration by correlation model function. In: Proceedings of IEEE international conference on image processing, vol 2, pp 133–136
- Porikli F (2005) Integral histogram: a fast way to extract histograms in Cartesian spaces. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 1, pp 829–836
- Prosser B, Gong S, Xiang T (2008) Multi-camera matching using bi-directional cumulative brightness transfer functions. In: Proceedings of the British machine vision conference
- Rahimi A, Dunagan B, Darrel T (2004) Simultaneous calibration and tracking with a network of non-overlapping sensors. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition
- Rasmussen C, Hager G (1998) Joint probabilistic techniques for tracking multi-part objects. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 16–21
- Savarese S, Winn J, Criminisi A (2006) Discriminative object class models of appearance and shape by correlators. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2, pp 2033–2040
- Schiele B, Crowley JL (2000) Recognition without correspondence using multidimensional receptive field histograms. *Int J Comput Vis* 36(1):31–50
- Schwartz WR, Davis LS (2009) Learning discriminative appearance-based models using partial least squares. In: Brazilian symposium on computer graphics and image processing
- Seigneur JM, Solis D, Shevlin F (2004) Ambient intelligence through image retrieval. In: International conference on image and video retrieval. Springer, Berlin, pp 526–534
- Senior A, Hsu MA R, Land Mottaleb, Jain AK (2002) Face detection in color images. *IEEE transactions on pattern analysis and machine intelligence* 24(5):696–706
- Shotton J, Winn J, Rother C, Criminisi A (2006) TextonBoost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: Proceedings of the European conference on computer vision, pp 1–15
- Song Y, Goncalves L, Perona P (2003) Unsupervised learning of human motion. *IEEE Trans Pattern Anal Mach Intell* 25(7):814–827
- Swain MJ, Ballard DH (1991) Color indexing. *Int J Comput Vis* 7(1):11–32
- Teixeira LF, Corte-Real L (2009) Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recognit Lett* 30(2):157–167
- Truong Cong DN, Achard C, Khoudour L, Douadi L (2009) Video sequences association for people re-identification across multiple non-overlapping cameras. In: Proceedings of the international conference on image analysis and processing
- Truong Cong DN, Khoudour L, Achard C, Meurie C, Lezoray O (2010) People re-identification by spectral classification of silhouettes. *Signal Process* 90(8):2362–2374
- Tu PH, Doretto G, Krahnstoeber NO, Perera AAG, Wheeler FW, Liu X, Rittscher J, Sebastian TB, Yu T, Harding KG (2007) An intelligent video framework for homeland protection. In: Carapazza EM (ed) Proceedings of SPIE defence and security symposium—unattended ground, sea, and air sensor technologies and applications IX, Orlando, vol 6562
- Tuzel O, Porikli F, Meer P (2006) Region covariance: a fast descriptor for detection and classification. In: Proceedings of the European conference on computer vision, pp 589–600

- Varma M, Zisserman A (2005) A statistical approach to texture classification from single images. *Int J Comput Vis* 62:61–81
- Vedaldi A, Soatto S (2006) Local features, all grown up. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2, pp 1753–1760
- Vincent L, Soille P (1991) Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans Pattern Anal Mach Intell* 13(6):583–598
- Viola P, Jones MJ (2004) Robust real-time face detection. *Int J Comput Vis* 57:137–154
- Wang L, Tan T, Ning H, Hu W (2003) Silhouette analysis-based gait recognition for human identification. *IEEE Trans Pattern Anal Mach Intell* 25(12):1505–1518
- Wang X, Doretto G, Sebastian TB, Rittscher J, Tu PH (2007) Shape and appearance context modeling. In: Proceedings of IEEE international conference on computer vision, pp 1–8
- Winn J, Criminisi A, Minka T (2005) Object categorization by learned universal visual dictionary. In: Proceedings of IEEE international conference on computer vision, vol 2, pp 1800–1807
- Wolf L, Bileschi S (2006) A critical view of context. *Int J Comput Vis* 69(2):251–261
- Wu H, Liu X, Doretto G (2008) Face alignment using boosted ranking models. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 1–8
- Zhang J, Collins R, Liu Y (2003) Representation and matching of articulated shapes. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp II:342–349
- Zhao Q, Tao H (2005) Object tracking using color correlogram. In: IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance, pp 263–270
- Zitnick CL, Jojic N, Kang SB (2005) Consistent segmentation for optical flow estimation. In: Proceedings of IEEE international conference on computer vision, pp 1308–1315