People detection and re-identification for multi surveillance cameras

Etienne Corvee, Slawomir Bak and Francois Bremond
INRIA, Sophia Antipolis, Pulsar Team
{*etienne.corvee, slawomir.bak, francois.bremond*}*@inria.fr*

Keywords: people detection, people tracking, people re-identification, local binary pattern, mean Riemannian covariance

Abstract: Re-identifying people in a network of non overlapping cameras requires people to be accurately detected and tracked in order to build a strong visual signature of people appearances. Traditional surveillance cameras do not provide high enough image resolution to iris recognition algorithms. State of the art face recognition can not be easily applied to surveillance videos as people need to be facing the camera at a close range. The different lighting environment contained in each camera scene and the strong illumination variability occurring as people walk throughout a scene induce great variability in their appearance. In addition, people images occlud each other onto the image plane making people detection difficult to achieve. We propose a novel simplified Local Binary Pattern features to detect people, head and faces. A Mean Riemannian Covariance Grid (MRCG) is used to model appearance of tracked people to obtain highly discriminative human signature. The methods are evaluated and compared with the state of the art algorithms. We have created a new dataset from a network of 2 cameras showing the usefulness of our system to detect, track and re-identify people using appearance and face features.

1 Introduction

Recently, the person re-identification problem became one of the most important tasks in video surveillance. Only knowledge about identities of tracked persons can allow a system to fully extract semantic information about scene activities. The human re-identification problem can be defined as a determination whether a given person of interest has already been observed over a network of cameras.

Haar features have been studied intensely for the detection of objects, in particular for face detection (Viola and Jones, 2004). One another major feature used for object detection is provided by HOG as evaluated in (Dollar et al., 2009). Pedestrians, faces and bicycles are successfully detected when represented by HOG (Dalal and Triggs, 2005a; Adam et al., 2006). A boosting technique is often used to model and rapidly detect objects (Laptev, 2006) such as humans (Zhu et al., 2006). SVM coupled with HOG is also often used (Dalal and Triggs, 2005a) for this task. Although Covariance features can be computationally expensive to estimate, they have strong discriminative powers. Tuzel and al. (Tuzel et al., 2008) use a Logiboost algorithm on Riemannian manifolds where Covariance features in a Riemannian geometry are trained allowing the classification of pedestrians.

Many recent papers use body parts to enhance people detection performance. There are many ways to combine body parts; for instance Mohan et al. (Mohan et al., 2001) studied different voting combination of body parts classifiers. In (Mikolajczyk et al., 2004), Mikolajczyk et al. use 7 body part detectors independently trained to better detect humans using SIFT-like descriptors. Hussein and Porikli (Hussein et al., 2009) introduce the notion of deformable features in a Logiboost algorithm to allow body parts to have non fixed locations in a people image template. More recently, high performances were obtained by (Huang and Nevatia, 2010) using a highly trained set of granules.

There is a natural consequence of an invention of robust human detection algorithms to extend approaches for re-identification purposes. The appearance-based re-identification techniques were focused on associating pairs of images, each containing one instance of individual. These methods are named *single-shot* approaches (Bak et al., 2010b; Park et al., 2006; Wang et al., 2007) and until now they were the most popular techniques. Currently researches try to improve identification accuracy by integrating information over many images. The group of methods which employs multiple images of the same person as training data is called *multiple-shot*

approaches.

As to *single-shot* approaches, in (Park et al., 2006) the clothing colour histograms taken over the head, shirt and pants regions together with the approximated height of the person were used as the discriminative feature. Similarly, clothing segmentation together with facial features (Gallagher and Chen, 2008) were employed to re-identify individuals. Shape and appearance context model is proposed in (Wang et al., 2007). A pedestrian image is segmented into regions and their colour spatial information is registered into a co-occurrence matrix. This method works well if the system considers only a frontal viewpoint. For more challenging cases, where viewpoint invariance is necessary, the ensemble of localized features (*ELF*) (Gray and Tao, 2008) has been proposed. Instead of designing a specific feature for characterizing people appearance, a machine learning algorithm constructs a model that provides maximum discriminability by filtering a set of simple features. Enhancement of discriminative power of each individual signature with respect to the others was also the main issue in (Lin and Davis, 2008). Pairwise dissimilarity profiles between individuals have been learned and adapted into nearest neighbour classification. Similarly, in (Schwartz and Davis, 2009), a rich set of feature descriptors based on colour, textures and edges has been used to reduce the amount of ambiguity among human class. The high-dimensional signature was transformed into a low-dimensional discriminant latent space using a statistical tool called Partial Least Squares (PLS) in one-against-all scheme. Nevertheless in both methods, an extensive learning phase based on the pedestrians to re-identify is necessary to extract discriminative profiles what makes the approaches non-scalable. The person re-identification problem has been reformulated as a ranking problem in (Prosser et al., 2010). The authors presented extensive evaluation of learning approaches and show that a ranking relevance based model can improve the reliability and accuracy.

Concerning *multiple-shot* approaches, in (Gheissari et al., 2006) the spatiotemporal graph was generated for ten consecutive frames for grouping spatiotemporally similar regions. Then, clustering method is applied to capture the local descriptions over time and improve matching accuracy. In (Bak et al., 2010a), the AdaBoost was applied to extract the most discriminative and invariant Haar-like features. Here, again one-against-all learning scheme was used to catch human dissimilarities. In (Farenzena et al., 2010), the authors proposed to combine three features: 1) chromatic content (HSV histogram); 2) maximally stable colour regions (MSCR) and 3) recurrent highly structured patches (RHSP). The extracted features were weighted by the distance with respect to the vertical axis to minimize effects of pose variations. Recurrent patches were also proposed in (Bazzani et al., 2010). Epitome analysis was used to extract highly informative patches form the set of images.

2 Overview

Based on our previously evaluated appearance-based people re-identification system in (Bak et al., 2011), we here test our system in a 2 camera network associating people, head and face. The first task is to detect people using simplified LBP features in section 3. Tracked objects are extracted from a video using a temporal trajectory analysis algorithm (A. Avanzi and Thonnat, 2001) which are fed to the appearance based people re-identification system in section 4. We propose a simple face recognition algorithm in section 5. Re-identification results are given in section 6. A new database was created to simulate a 2 cameras network with non overlapping views. This database is referred to as the TSP database and contains 23 persons walking back and forth in both cameras as illustrated in figure 1. Two tracked persons with heads and faces are shown in camera 1 and 2 of the TSP camera setup in figure 2.

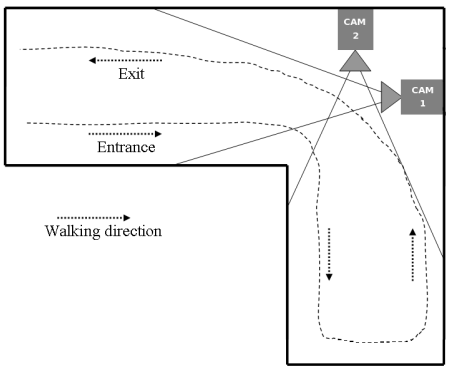

Figure 1: TSP camera setup: 2 non overlapping cameras in 2 joint corridors

3 People detection using simplified Local Binary Pattern

The standard LBP operator extract feature vector of size 256 (Trefny and Matas, 2010). We have implemented a new simplified version of the LBP operator

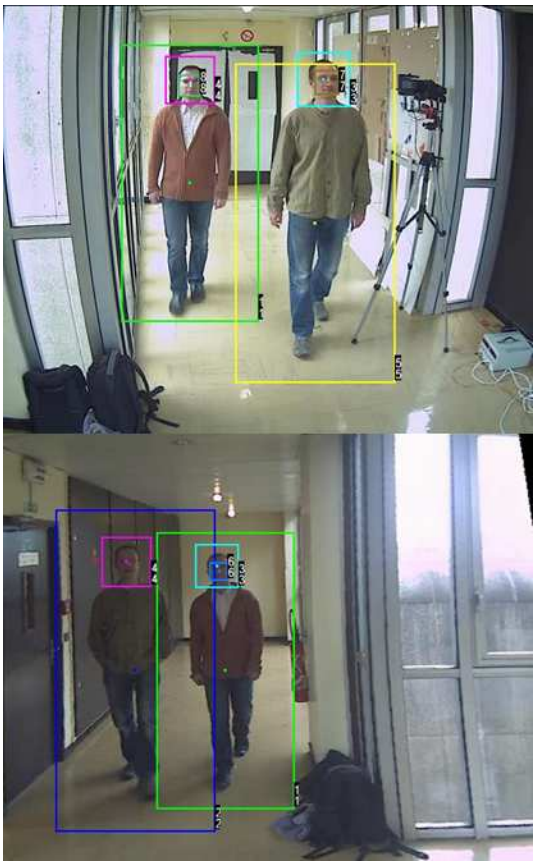

Figure 2: Tracked people, head and faces in up: TSP camera 1 and down: TSP camera 2

by reducing its dimensionality to 16. which we will refer to as the SLBP for the rest of the paper. This SLBP is calculated from 4 cells as illustrated in image 3. A pixel intensity average $v_{c,w,h}$ is calculated for each of the 4 cells c, $c = [1:4]$ as defined in equation 1 where E represents the expectation (average) function over pixel locations (x,y).

$$v_{c,w,h} = E_{x',y'}\left(I(x',y')\right) \quad (1)$$

$$c = \begin{cases} 1 : x' = [x : x+h],\ y' = [y : y+w] \\ 2 : x' = [x : x+h],\ y' = [y+w : y+2w] \\ 3 : x' = [x+h : x+2h],\ y' = [y : y+w] \\ 4 : x' = [x+h : x+2h],\ y' = [y+w : y+2w] \end{cases}$$

The SLBP feature is then calculated from 4 mean pixel intensity differences Δv_q as follows:

$$\begin{aligned} \text{SLBP}_{w,h}(\mathbf{x}) &= \sum_{q=0}^{3} s(\Delta v_q) 2^q \ ,\ s(.) = \begin{cases} 1 \text{ if } . > V \\ 0 \text{ else} \end{cases} \\ \Delta v_q &= v_i - v_j \\ (q,i,j) &= \{(0,1,0),(1,3,2),(2,2,0),(3,3,1)\} \end{aligned}$$

An Adaboost training scheme is adopted to train cell features across the a people image database. The training is performed on a multiple scale approach by varying the SLBP cell dimensions:

$$(w,h) = \{(1,1)(1,2)..(1,8),(2,1)..(8,8)\}$$

We apply the same training algorithm on head and face image datasets for head and face detection. The training for people detection is done using 7K positive images (5K from NICTA, 1K from MIT and 1K from INRIA people training dataset) and 50 background negative PAL images. Head were trained with 1K positive images (cropped head images from INRIA and TUD people training dataset) and 10 background negative images. Faces were trained with the standard CMU face training database.

The traditional Adaboost technique is used to train SLBP features. We reach a frame rate of 1fps for a minimum size of 160 pixels for people height in PAL images and when the scan is performed with an increase rate of 10% of its actual scanning window width and height. Simple rules are applied to the detected candidates in order to fuse overlapping candidates and eliminate stand alone noisy candidates: objects are merged if they overlap each other with a minimum union-over-intersection ratio of 50% and a final object requires a minimum number of 2 overlapping ones.

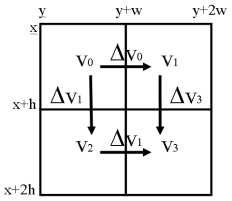

Figure 3: Simplified LBP operator.

4 Human re-identification

Recently, the person re-identification problem became one of the most important tasks in video surveillance. There is a natural consequence of an invention of robust human detection algorithms to extend approaches for recognition purposes. Person re-identification can be considered on different levels depending on information cues which are currently available in the system. Biometrics such as face, iris or gait can be used to recognize identities. Nevertheless, in most video surveillance scenarios such detailed information is not available due to video low-resolution or difficult segmentation (crowded environments, *e.g.* airports, metro stations). Therefore

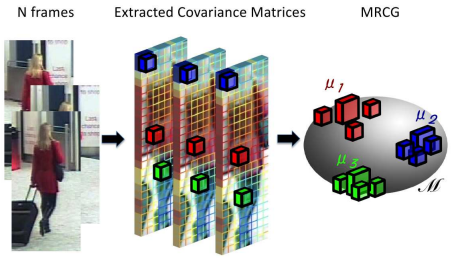

Figure 4: Covariances gathered from tracking results are used to compute the MRC using Riemannian manifold space (depicted with the surface of the sphere).

a robust modelling of a global appearance of an individual is necessary to re-identify a given person of interest. In these identification techniques (named *appearance-based approaches*) clothing is the most reliable information about an identity of an individual (there is an assumption that individuals wear the same clothes between different sightings). The model of an appearance has to handle differences in illumination, pose and camera parameters to allow matching appearances of the same individual observed in different cameras.

In (Bak et al., 2011), a highly discriminative human signature, called *Mean Riemannian Covariance Grid (MRCG)*, is proposed. This human signature has been designed to deal with low resolutions images and crowded environments where more specialized techniques (*e.g.* based on body parts detectors) might fail. Here, dense descriptors philosophy (Dalal and Triggs, 2005b) is combined with extremely effectiveness of the covariance descriptor. First, an image is divided into a dense grid structure with overlapping spatial square regions (*cells*). Such dense representation makes the signature robust to partial occlusions and it contains a relevant information about spatial correlations between *cells*. The authors (Bak et al., 2011) take advantage of the tracking and detection results combining information from multiple images.

Let $C_1^p,\ldots,C_N^p$ be a set of covariance matrices extracted during tracking of N frames corresponding to image square regions at position of the cell p. The MRC is defined as the mean covariance of these covariance matrices computed using Riemannian space (see Fig. 4). The mean covariance matrix as an intrinsic average blends appearance information from multiple images. This mean covariance matrix keeps not only information about features distribution but also carries out essential cues about temporal changes of the appearance related to the position of the cell p. All MRC *cells* compose a full grid, named as Mean Riemannian Covariance Grid (MRCG).

In our surveillance system, this appearance-based descriptor is used to match the same appearances between different camera views.

5 Face recognition

5.1 Face visual signature extraction

Similarly to (Huang and Nevatia, 2010) where people body parts are manually located before being independently trained, we manually choose 4 facial parts ($p=[1:4]$) to model a face i visual signature $S^{(i)}$ which are the left eye, right eye, nose and mouth:

$$S^{(i)} = \left\{ H_p^{(i)} \; ; \; p=[1:4] \right\} \tag{2}$$

Each signature $H_p^{(i)}$ is represented by a set of SLBP histograms $h_{p,w,h}^{(i)}(f)$:

$$H_p^{(i)} = \left\{ h_{p,w,h}^{(i)}(f) \; ; \; (w,h) = \{(1,1)..(1,8)..(8,8)\} \right\}$$

$$h_{p,w,h}^{(i)}(f) = \sum_{\mathbf{x}\in P_p} s'(\mathrm{SLBP}_{w,h}^{(i)}(\mathbf{x}), f) \tag{3}$$

$$s'(a,b) = \begin{cases} 1 \text{ if } a=b \\ 0 \text{ else} \end{cases} \tag{4}$$

where P_p represent the set of pixels within the predefined facial part area. Each histogram h is normalised over the SLBP feature value $f=[1:16]$.

5.2 Matching faces

Two faces i and j are matched if their signatures $S^{(i)}$ and $S^{(j)}$ are similar enough or in other terms if a distance measure between the two signatures is below a threshold. This similarity distance $D(i,j)$ is calculated by the mean similarity distance $\Delta H_p(i,j)$ of the 4 facial part signatures H_p defined in equation 2. Two facial parts similarity is calculated by a weighted mean of SLBP histogram differences as follows.

$$D(i,j) = E_p\left(\Delta H_p(i,j)\right) \tag{5}$$

$$\Delta H_p(i,j) = \frac{\sum_{(w,h)} \alpha_{w,h} E_f\left(\left(h_{p,w,h}^{(i)}(f) - h_{p,w,h}^{(j)}(f)\right)^2\right)}{\sum_{w,h}\alpha_{w,h}}$$

where cell weight α is inversely proportionally to the cell lateral distance.

6 Evaluation

6.1 People detection

We have evaluated our people detection algorithm on the test human dataset provided by INRIA against state of the art algorithms which we refer as HOG (Dalal and Triggs, 2005b) and LBP-HOG (Wang et al., 2009). The INRIA human dataset is composed of 1132 human images and 453 images of background scenes containing no humans. The results are displayed in figure 5 which shows that we obtain slightly better performances than the HOG-LBP technique in terms of missed detection rate vs. FPPI i.e. False Positive Per Image. In this figure, two extreme functioning modes could be chosen: approximately 2 noisy detections are obtained every 1000 background images for 50% true positive detections or 1 noisy detection every 2 frames for a detection rate of approximately 88%.

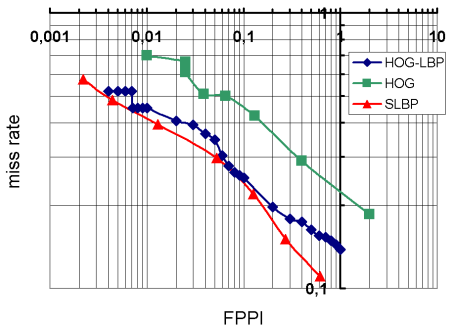

Figure 5: People detection evaluation: False Positive Per Image vs. miss detection rate for the INRIA test database

6.2 Appearance based people re-identification

Using the TSP dataset described in section 2, three kind of people are noted:

- type A - people wearing the same clothes in recording sessions.
- type B - people slightly change their appearances from one recording session to the other. For example, someone unzipping his/her jacket or someone taking his/her scarf off as illustrated by the top images of figure 6.
- type C - people with great change of appearance such as a person adding a coat with a hat as illustrated by the bottom images of figure 6.

We obtain 50 possible comparison of appearances that we evaluated in section 6.2.1 and 6.2.2. We can note that people detection and tracking results are noisy which makes the people re-identification task more challenging.

Figure 6: Top row: type B - people appearances are successfully re-identified despite their weak change of clothing. Bottom row: type C - people appearances are too distant to be re-identified

6.2.1 Same camera re-identification scenario

In this scenario, we aim to re-identify people in a camera who were previously seen in this same camera. The tracked people appearance signatures are recorded in a database during the first session. During the second session, the appearance signatures of all the tracked people are compared with the database. We have plotted in figure 7 the appearance matching distance between the same person appearances for type A, B and C people (defined above). The results show that people type A and B are successfully retrieve with rank 1, except for 1 person with rank 2, when matching distances are below 2.1. Type C people (i.e. strong change of appearance) are not re-identify. However, a uncertainty zone exists for a matching distance between 1.75 and 2.1.

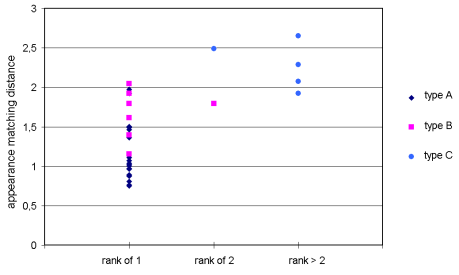

Figure 7: Evaluation of appearance matching distance in the same camera scenario

6.2.2 Different camera re-identification scenario

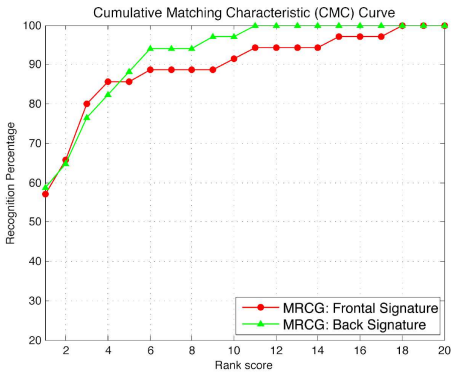

Figure 8: CMC curves obtained on TSP datasets.

In this section the evaluation of re-identification approach is presented in a 2 camera network. The performance is shown using the Cumulative Matching Characteristic (CMC) curve suggested in (Gray et al., 2007) as the validation method for the re-identification problem. The CMC curve represents the expectation of finding the correct match in the top matches. We evaluate the re-identification in the following way:

- Mono appearance mode - Frontal and back view of people tracked in camera 1 have their appearance signatures extracted separately.
- Multi appearance mode - Both frontal and back view of people tracked in camera 2 have their appearance signatures extracted.

Every signature from camera 1 is used as a query to the database (*multiple-appearance* signatures extracted from camera). The CMC curves for back and frontal view signatures are presented in Fig. 8. The results show that despite automatic inaccurate people detection and the different lighting conditions of the 2 cameras, the system shows promising people appearance based re-identification results.

6.3 Face recognition

The Olivetti Research Lab i.e. ORL database is used to evaluate our simple face recognition algorithm using SLBP features. 40 persons constitute this database with 10 pictures for each subject. All face pictures are taken by the same camera. Most state of the art algorithm compare face recognition rate using one picture at a time. In our scenario, we obtain tracks of faces for each person which we aim to compare. To simulate our 2 cameras system, we split the database in two: 5 images are used as our face database collected from one camera and the remaining 5 images are used as a request from camera 2. The images are challenging as they contain faces of different kind of poses and facial expressions.

technique	recognition rate %		
	ORL	TSP1	TSP2
DAISY	98.2		
PDSIFT	95.5		
SIFT Grid	95.2		
SLBP	85	91	100
SLBP(rank 1)	98	100	100
SLBP(rank 3)	100	100	100

Table 1: Face recognition rates

Table 1 shows the recognition rates comparison with 3 different techniques. Using an empirically set threshold (of 0.001), our system recognises only 85% of faces correctly whereas the DAISY technique (Velardo and Dugelay, 2010) recognises 98% of the faces correctly. Although our system is far from being the most performing system, it recognises 98% of the people when using a rank score of 1 and 100% for a rank of 3. In other terms, there is 98% chance that the face giving the closest distance D (in equation 5) is the correct face.

Using the TSP dataset, we obtained similar performances. Among the 23 persons present in the TSP database, only 11 people faced long enough camera 1 for a valid trajectory to be obtained: face images of less than 50x50 pixels were not taken into account (due to motion noise and image quality artefacts) and face tracks containing less than 5 face images were discarded.

In table 1, TSP1 and TSP2 refers to the same camera (camera 1 and camera 2 respectively) recognition scenario where half of a person face images is saved in the database and the other half as the request information. The results show that this simple SLPB face recognition technique is useful enough for such a database.

7 Conclusion

We have presented a multi-cue people re-identification framework. A novel simplified LBP feature is proposed to detect people, head and faces using the Adaboost training scheme. We obtain state of the art performance for people detection. We have also extended our approach on appearance based matching method to multi appearance based people re-identification. The proposed system tracks

people and their faces allowing the user to associate an appearance with a face. In most networks, cameras cannot provide the full people appearance view (e.g. strong occlusion) and faces are often not visible or only partly visible. Our proposed approach would allow a user to scan throughout a network of surveillance cameras the best matching candidates and to be able to track people of interest throughout this network.

Acknowledgements

We would like to thank the ANR project 'VideoId' who partially founded this work and the following partners of the project: Biometrics Groups at TELECOM SudParis, Multimedia Image processing Group of Eurecom and T3S (Thales Security Systems and Solutions S.A.S.).

REFERENCES

A. Avanzi, F. B. and Thonnat, M. (2001). Tracking multiple individuals for video communication. In *In IEEE Proc. of International Conference on Image Processing, Thessaloniki (Greece)*.

Adam, A., Rivlin, E., and Shimshoni, I. (2006). Robust fragment-based tracking using integral histogram. In *Computer Vision and Pattern Recognition - CVPR*.

Bak, S., Corvee, E., Bremond, F., and Thonnat, M. (2010a). Person re-identification using haar-based and dcd-based signature. In *2nd Workshop on AMMCSS*.

Bak, S., Corvee, E., Bremond, F., and Thonnat, M. (2010b). Person re-identification using spatial covariance regions of human body parts. In *AVSS*.

Bak, S., Corvee, E., Bremond, F., and Thonnat, M. (2011). Multiple-shot human re-identification by mean riemannian covariance grid. In *AVSS*.

Bazzani, L., Cristani, M., Perina, A., Farenzena, M., and Murino, V. (2010). Multiple-shot person re-identification by hpe signature. In *ICPR*, pages 1413–1416.

Dalal, N. and Triggs, B. (2005a). Histograms of Oriented Gradients for Human Detection. In *Computer Vision and Pattern Recognition - CVPR*.

Dalal, N. and Triggs, B. (2005b). Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893.

Dollar, P., Wojek, C., Schiele, B., and Perona, P. (2009). Pedestrian detection: a benchmark. In *CVPR*.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367.

Gallagher, A. C. and Chen, T. (2008). Clothing cosegmentation for recognizing people. In *CVPR*, pages 1–8.

Gheissari, N., Sebastian, T. B., and Hartley, R. (2006). Person reidentification using spatiotemporal appearance. In *CVPR*, pages 1528–1535.

Gray, D., Brennan, S., and Tao, H. (2007). Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. *PETS*.

Gray, D. and Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275.

Huang, C. and Nevatia, R. (2010). High performance object detection by collaborative learning of joint ranking of granules features. *IEEE - CVPR 2010*.

Hussein, M., Porikli, F., and Davis, L. (2009). Object detection via boosted deformable features. *IEEE International Conference on Image Processing (ICIP)*.

Laptev, I. (2006). Improvements of object detection using boosted histograms. In *Proceedings of the British Machine Vision Conference*.

Lin, Z. and Davis, L. S. (2008). Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In *ISVC*, pages 23–34.

Mikolajczyk, K., Schmid, C., and Zisserman, A. (2004). Human detection based on a probabilistic assembly of robust part detectors. In *ECCV*.

Mohan, A., Papageorgiou, C., and Poggio, T. (2001). Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:349–361.

Park, U., Jain, A., Kitahara, I., Kogure, K., and Hagita, N. (2006). Vise: Visual search engine using multiple networked cameras. In *ICPR*, pages 1204–1207.

Prosser, B., Zheng, W.-S., Gong, S., and Xiang, T. (2010). Person re-identification by support vector ranking. In *BMVC*, pages 21.1–21.11.

Schwartz, W. R. and Davis, L. S. (2009). Learning discriminative appearance-based models using partial least squares. In *SIBGRAPI*, pages 322–329.

Trefny, J. and Matas, J. (2010). Extended Set of Local Binary Patterns for Rapid Object Detection. In *Computer Vision Winter Workshop 2010 - CVWW10*.

Tuzel, O., Porikli, F., and Meer, P. (2008). Pedestrian detection via classification on riemannian manifolds. *PAMI*, 30(10).

Velardo, C. and Dugelay, J. (2010). Face recognition with daisy descriptors. In *MM'10 and Sec'10, ACM SIGMM Multimedia and Security Workshop, September 9-10, Rome, Italy*.

Viola, P. and Jones, M. (2004). Robust real-time face detection. In *International Journal of Computer Vision*.

Wang, X., Doretto, G., Sebastian, T., Rittscher, J., and Tu, P. (2007). Shape and appearance context modeling. In *ICCV*, pages 1–8.

Wang, X., Han, T., and Yan, S. (2009). An hog-lbp human detector with partial occlusion handling. *ICCV09 - International Conference on Computer Vision*.

Zhu, Q., Avidan, S., Yeh, M., and Cheng, K. (2006). Fast human detection using a cascade of histograms of oriented gradients. In *Computer Vision and Pattern Recognition - CVPR*.