Color Based Tracing in Real-life Surveillance Data

Michael J. Metternich, Marcel Worring, and Arnold W.M. Smeulders

ISLA-University of Amsterdam,
Science Park 107, 1098 XG Amsterdam, The Netherlands
{M.J.Metternich,M.Worring,A.Smeulders}@uva.nl
http://www.science.uva.nl/research/isla/

Abstract. For post incident investigation a complete reconstruction of an event is needed based on surveillance footage of the crime scene and surrounding areas. Reconstruction of the whereabouts of the people in the incident requires the ability to follow persons within a camera's field-of-view (tracking) and between different cameras (tracing). In constrained situations a combination of shape and color information is shown to be best at discriminating between persons. In this paper we focus on person tracing between uncalibrated cameras with non-overlapping field-of-view. In these situations standard image matching techniques perform badly due to large, uncontrolled variations in viewpoint, light source, background and shading. We show that in these unconstrained real-life situations, tracing results are very dependent on the appearance of the subject.

Key words: Real-life Surveillance and Tracing

1 Introduction

The two major applications of camera surveillance are real-time crime prevention and crime investigation after an incident has occurred. For the first type of application event detection [23, 22] or aggression detection [6, 30] are used to understand people's actions. For post incident investigation a complete reconstruction of the event is needed which additionally requires to follow persons within a camera's field-of-view (tracking) and between different cameras (tracing). A system aiding a human user in this process should therefore be able to perform both tracking and tracing.

Person tracking is a very lively research area, with various workshops and challenges organized each year such as Performance Evaluation of Tracking and Surveillance (PETS) and People Detection and Tracking workshop of the 2009 IEEE International Conference on Robotics and Automation. Though this subject is far from solved, impressive results have been obtained in constrained as well as real-life situations. Note, however, that surveillance data is often time-lapsed and in these conditions tracking algorithms cannot be used as is, but need

reconsideration or adaptation[15]. For an excellent overview of available tracking methods and their advantages and disadvantages, see [29].

Up to now, studies regarding tracing have only been tested in controlled conditions. The major issues of real-life surveillance situations however, are the large, uncontrolled variation in viewpoint, light source and shading. These variations have great impact on the appearance of a person. Viewpoint changes affect the shape of the person as well as the colors of clothing as the observed color changes with the angle between light source and line of view. Changes in light source have direct impact on the color of any object and while shading does not change the color itself, it does change other characteristics like intensity or saturation. In constrained situations these challenges could be reduced by using pre-calibrated cameras [31, 2] or by calibrating the cameras afterward using information about the overlapping field-of-view [4]. If the cameras are not calibrated and either the cameras' field-of-view do not overlap or it is unknown what the overlap is, any description usable for tracing should be able to deal with the lack of calibration. Color changes between cameras can be addressed by using color constancy methods [27] and various color spaces can be used to be invariant to shadows. These methods have been designed based on solid theoretical foundations and proven to be optimal in lab conditions. A major question is whether these methods generalize well enough to deal with the challenges of surveillance data, or that other criteria play a role there.

The paper is organized as follows. First, a state-of-the-art person detection system is described, which results are used throughout the paper. Section 3 introduces the tracing methods and invariance properties to answer the questions: How are image regions best described to be able to distinguish between persons? And (how) can unwanted variation in image regions be suppressed? The results of applying these methods to a benchmark and a real-life dataset are discussed in Section 4.

2 System Description

Any post incident investigation system is composed of at least three major steps. The first step is detection of persons in each of the cameras. The second step is to track these persons within a single camera. Afterward, these tracks are used to find instances of these persons in other cameras.

The standard approach for person detection is to match certain candidate detections to a model which is previously trained on sample images. One method to select these candidate detections is by simply selecting all regions over all frames of all possible sizes and locations. This is known as the sliding window technique. This method has two serious problems though: object classification is a time-consuming method, so classifying every possible sub-region of all frames in a dataset is infeasible. Another issue is that relying solely on classification scores will result in a great dependency on the chosen threshold. Both issues can be addressed by applying object classification only to certain regions of interest. The standard approach to obtain these regions of interest is background

extraction; we use [32] for its ability to automatically optimize the number of components in a Gaussian Mixture Model. All regions of interest are described using Histograms of Oriented Gradients [5] and classified using a Support Vector Machine (SVM) classifier [28], previously trained on the INRIA dataset[1]. Regions of which the score exceeds some predefined threshold are then classified as persons. A known issue of only applying classification on non-static regions is that persons standing still cannot be detected. However, we focus on tracing persons; the only requirement on detections is therefore that each person was detected at least once.

To be able to combine the resulting detections into tracks either hysteresis thresholding can be used [15] or the detections can be used as initializations for a filtering method such as Kalman filtering [14]. While these methods will help in reducing the number of false positives and might lead to better representations, we do not use either method but focus on matching of singular image regions instead. The reason is that any method to combine the image regions will make errors, leading to paths containing two different persons or false positives as well as true positives. This might distract us from the goal of this paper, understanding how image regions should be matched.

If the camera's field of view is actively changed, results of various methods assuming static cameras are poor, e.g. background extraction. Before doing background extraction, we therefore automatically detect camera movement. The simple method we applied to detect these moments of camera movement is to measure the movement of salient points using sift features [19]. If the average movement is greater than some predefined threshold the camera is assumed to be moving. For the current paper these regions are excluded from further analysis. In practice, however, these parts could contain relevant information like zooming in on specific persons. Figure 1 shows an overview of this person detection system. A preliminary version of the system was published in [21].

3 Tracing

We now use the image regions obtained from the person detection system described earlier to find image regions depicting the same person in other cameras. This can be studied using three techniques: image matching, image classification and image pair analysis. For image matching the image regions are described using predefined features, after which a comparison metric is used to order all image regions. Image classification uses similar features to represent an image region, but a classifier is specifically trained to distinguish between objects. The last method, image pair analysis, learns the most descriptive parts of an image regions autonomously, using manually defined image-region pairs [20]. These image regions are then used to identify a new object. The last two methods need

[1] This publicly available dataset can be downloaded from *http://pascal.inrialpes.fr/data/human/*

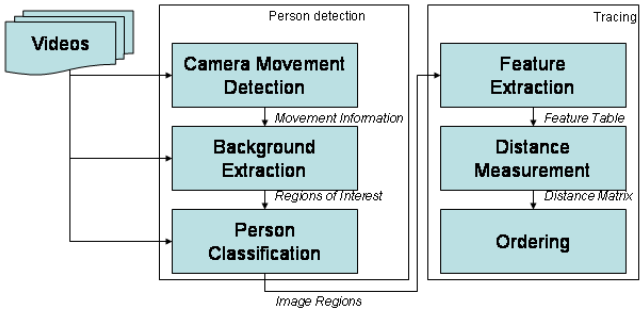

Fig. 1. Operational scheme of our person detection system.

specific training data of a person of interest. Instead of identifying persons, we aim to provide an overview of a dataset by helping the user to find multiple occurrences of a selected person. In these cases a classifier cannot be learned beforehand, so we focus on image matching techniques.

3.1 Feature Description

The simplest description of an image region is a global color histogram. This method ignores all location information and while this is very useful in many situations, it cannot make the vital distinction between a person wearing a red jacket and blue jeans and a person wearing a blue jacket and red jeans.

To obtain some location information Gray et al [12] introduced a 1D color histogram detected on predefined partitions of the image region. This method first divides the image in three parts: the top one fifth and middle and bottom two-fifths. For each sub-region three 1D histograms are calculated; the concatenated nine histograms are used as a descriptor for the complete image region.

In [1] an improvement of this descriptor is proposed which uses a collection of grids of region descriptors. Each grid segments the object into a different number of equally sized sub-rectangles. By combining fine regions with coarse regions, both local and global information is preserved.

Both methods are able to combine color and location information, but are unable to describe structure information such as patterns on a shirt. One method to capture shape, location and color information is the covariance matrix [25]. This method extracts a feature vector f_n for each pixel which are combined into the covariance matrix C of a region by:

$$C = \frac{1}{N-1}\sum_{n=1}^{N}(f_n - m)(f_n - m)^T \tag{1}$$

where N is the number of points in the region, m the mean vector of all the feature vectors and f_n the feature vector used to describe a position in the

region. To measure the influence of different descriptor types, we use three types of feature vectors:

$$f_{shape} = C(x, y, I, I_X, I_Y, I_{xx}, I_{yy}, mag, o) \tag{2}$$

$$f_{color} = C(x, y, Ch_1, Ch_2, Ch_3) \tag{3}$$

$$f_{combination} = C(x, y, Ch_1, Ch_2, Ch_3, I_x, I_y, mag, o) \tag{4}$$

where Ch_x indicates the x color channel which is dependent on the used color space. mag and o are based on the first order derivatives with respect to x and y:

$$mag(x, y) = \sqrt{I_x^2(x, y) + I_y^2(x, y)} \tag{5}$$

$$o(x, y) = \arctan\left(\frac{I_y(x, y)}{I_x(x, y)}\right) \tag{6}$$

f_{shape} uses shape information in the form of I_x, I_y, I_{xx}, I_{yy}, mag and $order$ and no color information. f_{color} considers only color information while $f_{combination}$ uses a combination of both shape and color: All feature vectors are used as a collection of grids of region descriptors.

A second method to capture both shape and color information in a single descriptor is the use of color SIFT features. To obtain fixed-length feature vectors per image, the bag-of-words model is used [24], which is also known as 'textons' [18], 'object parts' [8] and 'codebooks' [13, 17]. When using the bag-of-words model a large number of randomly sampled descriptors is clustered to obtain a visual codebook. In an image region all descriptors are then assigned to the codebook element which is closest in Euclidean space. To be independent of the total number of descriptors in an image, the feature vector is normalized to sum to 1. In this paper a visual codebook of 128 elements is constructed by applying k-means clustering to 20,000 randomly sampled descriptors from the set of images available for training. As descriptors we use both the traditional SIFT implementation without color information [19] and the concatenation of these descriptors calculated in each color channel separately.

3.2 Distance Metrics

To compare the resulting histogram features, several distance measures can be used, such as the Euclidean distance, intersection distance, quadratic cross distance and Bhattacharyya distance. Similar to [1] and [12] we use the Bhattacharyya distance:

$$D_{hist}(h_1, h_2) = -\ln\left(\sum_{x \in X} \sqrt{h_1(x)h_2(x)}\right) \tag{7}$$

This distance metric is unsuitable for measuring the distance between covariance matrices. So for comparing elements described using equation 1 we use

the metric proposed by Forstner and Moonen [10] which sums the generalized eigenvalues of the covariances:

$$D_{covar}(C_1, C_2) = \sqrt{\sum_i \ln^2 \lambda_i(C_1, C_2)} \tag{8}$$

3.3 Invariant descriptors

Changes in illumination and color of the light source can greatly affect matching results if the descriptors used are not robust to these changes. To make tracing robust to changes in shadows and shading, intensity invariance is needed. Various methods have been proposed to obtain invariance to these unwanted variations, where color models and color constancy focus on illumination and color respectively.

Color model To measure colors of objects independent of shadings van de Sande et al. [26] studied two aspects of intensity invariance in a non-surveillance setting: light intensity change and light intensity shift. Light intensity change stands for the constant factor in all channels by which the image values change while light intensity shift stands for an equal shift in image intensity values in all channels. Similar to their overview we compare the following models:

RGB histogram The RGB histogram is a 3-D histogram based on the R, G and B channels of the RGB color space. This histogram possesses no invariance properties and is the most basic represenation.
Opponent histogram The opponent histogram is a 3-D histogram based on the channels of the opponent color space YCbCr. This color space was designed to The color models of the first two channels are shift-invariant with respect to light intensity. The third channel posseses intensity information and has no invariance properties.
Hue histogram The HSV histogram is a 3-D histogram based on the Hue, Saturation and Value channels of the HSV color space. The H and the S color models are scale-invariant and shift-invariant with respect to light intensity.
XYZ, Lab and Luv histogram The XYZ, Lab and Luv histograms are 3-D histograms based on the XYZ, Lab and Luv colorspaces respectively. These colorspaces were designed to mimic the response of the human visual system.

Hue histogram and Opponent histogram can be used without the intensity channel. The Opponent histogram then becomes invariant to light intensity shift while the Hue histogram becomes invariant to intensity scale and shift. We aim for some level of intensity invariance, so normalized rgb, Hue histogram without intensity and Opponent histogram without intensity are expected to perform best for tracing.

Color constancy Color constancy is the ability to measure colors of objects independent of the color of the light source [11]. For each video, frame or de-

tection a correction is computed which virtually changes the color of the light source to white. For the RGB color space this leads to the following corrections:

$$R_{\text{output}} = \frac{R}{\sqrt{3} * R_{lightsource}} \tag{9}$$

$$G_{\text{output}} = \frac{G}{\sqrt{3} * G_{lightsource}} \tag{10}$$

$$B_{\text{output}} = \frac{B}{\sqrt{3} * B_{lightsource}} \tag{11}$$

where R, G and B are the input channels and $R_{lightsource}$, $G_{lightsource}$ and $B_{lightsource}$ are estimates of the light source. To estimate the light source the color constancy methods proposed by Forsyth [11] and Finlayson [9] are not applicable, due to their complex nature and dependency on calibration datasets. We thereore compare three methods to estimate the light source: Grey world [3], max-RGB [16] and Grey Edge [27].

Of the three models Grey edge is expected to perform best as Weijer et al. [27] showed that for real-world images color constancy based on the Grey-Edge hypothesis obtains better results than those obtained with the Grey-World method.

4 Experimental Results

In this section we assess the performance of the presented methods, color models and color constancy methods.

4.1 Evaluation Criteria

To assess the performance of the tracing methods, we consider two scenarios. The first is an investigator who wants to find clues to explore further. In this case it is important that some evidence of a person's presence in any camera is found. The second is the full reconstruction of the event where all instances of the person have to be found. In both secnarios there is an asymmetry between the camera used as starting point and other cameras. In the start camera the investigator can easily select the most appropriate detection to be used as query, and correct detections if needed. The detections in the other cameras cannot be controlled.

We resort to a method used for biometric identification systems that return ranked lists of candidates to express the performance of the proposed tracing methods: the Cumulative Matching Curve (CMC)[7]. Assuming we have a set of samples B_i with associated ground truth ID(B_i), two subsets of samples are created:

1. A *gallery* set G consisting of m samples of *different* subjects.

2. A *probe* set Q with n samples associated with the n subjects. The probe set Q can be from any set of individuals, but usually probe identities are presumed to be in the gallery G. The probe set may contain more than one sample of a given person and need not contain a sample of each subject in G.

In order to estimate the CMC, each probe sample is matched to every gallery sample resulting in a $n \times m$ similarity matrix S. The scores for each probe sample are ordered to assign a rank k to every gallery sample, where k_l is the rank of a gallery sample obtained from the same person as the probe sample. CMC(k) is then the fraction of probe samples that have rank $k_l \leq k$:

$$CMC(k) = \frac{1}{n}(\#k_l \leq k) \tag{12}$$

If an investigator is searching through a video archive it is unlikely that he or she is focused on finding all instances of that subject. It is more important that any instance of the subject in another camera is found as soon as possible as this might lead to more information about the subject. We therefore redefine the CMC curve. Again the scores for all probe samples are ordered to assign a rank k to every gallery sample, but for each query only the first match k_{first} is of importance. This metric is called the First Matching Curve or FMC. Formally, $FMC(k)$ is then the fraction of probe samples that have rank $k_{first} \leq k$:

$$FMC(k) = \frac{1}{n}(\#k_{first} \leq k) \tag{13}$$

4.2 Datasets

As benchmark we use the VIPeR dataset [12] to show the performance of all image matching techniques and invariance properties described in section 3. This dataset consists of 632 persons, where each person was photographed twice under various viewpoints, poses and lighting conditions. Sample pictures of this dataset are shown in figure 2. While the images in the VIPeR dataset are taken from surveillance cameras and a lot of variance is present, persons are always positioned in the center of the image and in an upward position.

In real-life, image matching techniques should not only be able to match these perfectly located persons, but also retrieve images of the same person with less than optimal detections. For that reason we recorded a dataset with the assistance of the Dutch police; twenty cameras are used without any overlap in field-of-view. These recordings were made as part of the regular surveillance process for that area. A ground-truth is obtained by manually labeling the positions of four persons who were asked to walk around in the area under surveillance.

The dataset contains several sources of variation. Most notably are the presence of a large number of pedestrians participating in traffic, changes in weather, camera angle, colors and texture of clothing and reflections in windows. In addition, the area where we collected data was under active surveillance leading in some cases to movement of the cameras. Sample frames from this dataset are shown in figure 3.

Fig. 2. Some examples of the VIPeR dataset.

Fig. 3. Sample frames of the used surveillance data.

4.3 Results

The results are divided in two sections, we first present the performance on the VIPeR dataset and then show how the best method performs in an unconstrained situation.

Results on the VIPeR dataset The average CMC curves on the VIPeR dataset can be found in figures 4, 5 and 6. Figure 4 shows that if a user is willing to observe the first 20 matches, in 70% of the cases the person he was looking for could be found.

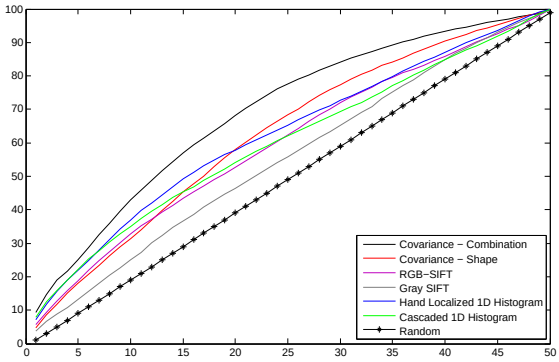

Fig. 4. CMC curves of the described matching techniques on the VIPeR dataset.

To be able to compare different color constancy methods we apply the methods implemented by [27]. Similar to [27] we vary the order of the method, the Minkowski norm and the local smoothing. Specific parameter settings can be found in table 1.

Method	Parameters		
	Order	Minkowski	Smoothing
Grey-World	0	1	0
Max-RGB	0	∞	0
Shades of Grey	0	7	0
general Grey-World	0	11	1
Grey-Edge	1	7	4
2nd order Grey-Edge	2	7	5

Table 1. Parameter settings for different color constancy methods.

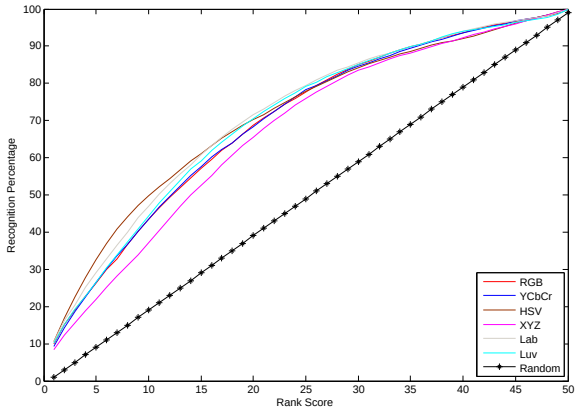

Fig. 5. CMC curves of color covariance matrix using different color spaces on the VIPeR dataset.

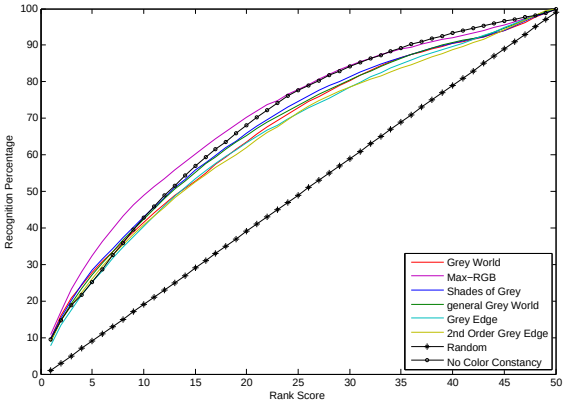

Fig. 6. CMC curves of color covariance matrix using different color constancy methods on the VIPeR dataset.

In conclusion, the covariance matrix with both color and shape information performs best on this benchmark. Adding color constancy slightly improves results, with Max-RGB performing best. Different color models did not influence results greatly, but HSV slightly outperforms all other methods.

Results on the real-life dataset Sample results of applying our person detection system to the real-life surveillance dataset are shown in figure 7. The detection results are thresholded in such a way that for each camera all persons in that camera's field-of-view were detected at least once. As a result, the number of false positives is large. Please note the large variation in amount of background within each bounding box.

Fig. 7. Sample frames of the used surveillance data with their detection results.

The automatic detections generated by the person detection system described in section 2 were matched to the ground truth. All detections with an overlap larger than 75% with a ground truth region are labeled as that specific person. Elements in the ground-truth larger than 6000 pixels, showing a complete, unobstructed body and not over or under saturated are then used as queries. We order the automatic detections in other cameras based on similarity to the query. A selection of these queries is given in figure 8. Automatic detections with the same label as the query are considered a match. Sample orderings using Cascaded 1d Histograms and covariance matrices with a combination of color and shape features are given in figure 9.

For each person separately the average CMC curve is measured to show the influence of different clothing. Initially we use covariance matrices with the combination of color and shape information, Max-RGB color constancy and HSV color space since this method performed best on the VIPeR dataset. We use a binary, person shaped mask to reduce the influence of backgrounds. Results are shown in figure 10.

A clear difference in performance can be observed between the four subjects; when person one, two and three are used as queries approximately random performance is obtained. This means that if a person is to be found, it is in general

Fig. 8. The four persons used to query the automatic detections.

Fig. 9. Sample orderings using Cascaded 1D Histograms (top row) and covariance matrices (bottom row). The query image is shown left with the first 7 results after the line.

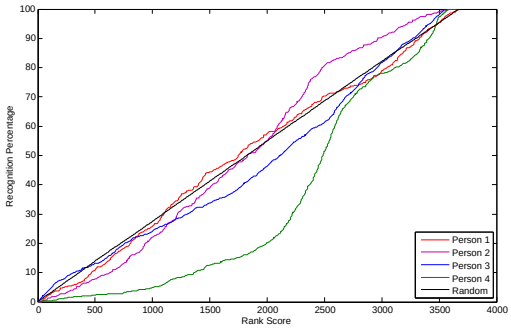

Fig. 10. CMC curves for different persons in real-life surveillance videos using Covariance matrices with a combination of shape and color information.

better to search the dataset chronologically. For person four however, performance is much worse than random. Since this person is visually very distinctive by the red jacket, a direct conclusion is that a representation more focused on color is more suitable for this type of data.

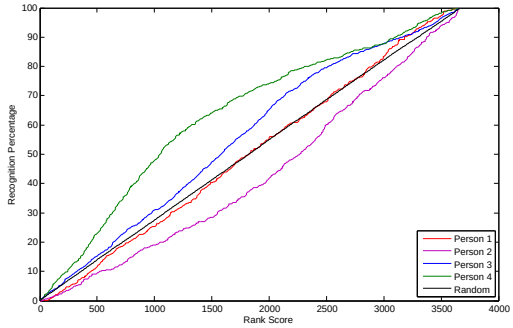

Fig. 11. CMC curves for different persons in real-life surveillance videos using Cascaded 1D Histograms.

Results of applying the Cascaded 1D Histogram are given in figure 11. Again Max-RGB color constancy is used, with the YCbCr colorspace and a binary, person shaped mask. As expected, person four is retrieved more easily than the three other subjects. For person one the performance is similar to random which is mostly due to the large difference in back and frontal appearance. Person two appears to be more easily traced using covariance matrices than color histograms. This is due to the fact that this person wore mostly gray colored clothes while the pattern on the jacket is more distinctive. Lastly, while person three wore mostly black clothes the combination of the black jacket with the white shirt underneath proves to be a strong enough visual cue to be able to trace ths person. In situation where the shirt was not visible it is unfeasible to find this person.

As mentioned earlier a person using the described person tracing system might not be interested in finding all instances of the person he or she is looking for but is more interested in finding a single instance of that person in another camera as fast as possible. We therefore apply the FMC metric to the same data and method. We show the results in figure 12.

The same observations we made after figure 11 apply here, but now the best results are obtained for person three instead of person four. The reason for this is that the image regions where for person three both his jacket and shirt were visible had a very high ranking. A user of a tracing system as described in this paper would then be able to obtain extra information leading to easier searches for other instances.

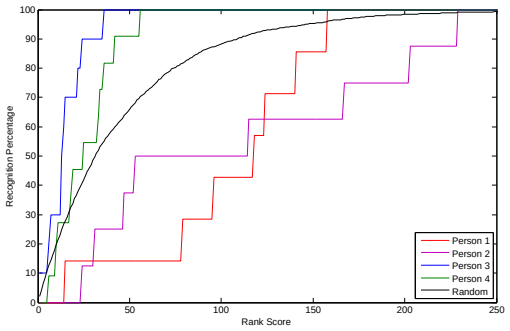

Fig. 12. FMC curves based on the first match for different person using Cascaded 1D Histograms. Only the first 250 ranks are shown out of 4000 image regions for readability.

5 Conclusion

In this paper we showed various methods to describe an image region which were used to trace a person over multiple cameras. We showed that a combination of color and shape information is needed for effective tracing on a dataset simulating perfect detection results. In situations where the detections are not so good however, this combination proved unsuccessful. A method focussing solely on color information was effective for two of the four subjects. This leads to the conclusion that before searching for a particular person the defining characteristics should be determined. If that person is wearing black clothes, any color based feature representation will fail for its inability to represent the color distribution properly. In these cases either more information should be used or a simple chronological search is recommended. If, however, the subject wears clothes with one or more distinctive colors, enough visual cues are present to be able to search through all videos based on a color-based feature representation.

To deal with changes in light and shadings, various color models and color constancy methods were applied. We showed that color constancy can slightly improve results by reducing the influence of color changes between cameras. Secondly, the use of a color space invariant to intensity is shown to improve tracing results by reducing the influence of shades.

We would like to point out that instead of using a single image of a person as query, multiple detections of the same person can be combined to obtain a track representation. Since such a representation can combine spatial and temporal information with the appearance information used in this paper, tracing performance could be greatly improved. Another point of interest is the large number of background regions falsely classified as persons. This issue can be dealt with in two ways. First of all the background extraction method can be improved. Background extraction is a very challenging task, but great results have been

achieved. However, there is still enough room for improvement. Secondly, the model classifying regions of interest as pedestrians can be improved. We described a method which is generally considered the state-of-the-art in person detection, but the SVM classifier was trained on a general pedestrian dataset. We expect a big improvement by training the model on a dataset more focused on real-life surveillance data.

In conclusion, methods for automating the process of incident reconstruction are promising, but it is a major step from the lab to real-life surveillance data. Our results give some guidelines for optimizing the process and for seeing under which conditions automatic analysis should be pursued.

References

1. A. Alahi, P. Vandergheynst, M.Bierlaire, and M.Kunt. Cascade of descriptors to detect and track objects across any network of cameras. Preprint submitted to Computer Vision and Image Understanding., 2009.
2. J. Black and T. Ellis. Multi camera image tracking. *Image Vision Comput.*, 24(11):1256–1267, 2006.
3. G. Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin Institute*, 310(1):337–350, 1980.
4. S. Calderara, A. Prati, and R. Cucchiara. Hecol: Homography and epipolar-based consistent labeling for outdoor park surveillance. *Computer Vision and Image Understanding*, 111(1):21–42, 2008.
5. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR '05: Preoceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:886–893, 2005.
6. D. Datcu, Z. Yang, and L. Rothkrantz. Multimodal workbench for automatic surveillance applications. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–2, 2007.
7. T. Dunstone and N. Yager. *Biometric System and Data Analysis: Design, Evaluation, and Data Mining.* Springer Publishing Company, Incorporated, 2008.
8. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–271, 2003.
9. G. Finlayson and S. Hordley. Gamut constrained illumination estimation. *International Journal of Computer Vision*, 67(1):93109, 2006.
10. W. Forstner and B. Moonen. A metric for covariance matrices. *Qua vadis geodesia*, 1:113128, 1999.
11. D. Forsyth. A novel algorithm for color constancy. *International Journal of Computer Vision*, 5(1):5–36, 1990.
12. D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Performance Evaluation of Tracking and Surveillance (PETS).*, 2007.
13. F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 604–610, Washington, DC, USA, 2005. IEEE Computer Society.
14. R. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering 82*, 1:3545, 1960.

15. P. Koppen and M. Worring. Multi-target tracking in time-lapse video forensics. In *MiFor '09: Proceedings of the First ACM workshop on Multimedia in forensics*, pages 61–66, New York, NY, USA, 2009. ACM.
16. E. Land and J. McCann. Lightness and retinex theory. *The Journal of the Optical Society of America A.*, 61(1):111, 1971.
17. B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *British Machine Vision Conference*, pages 759–768, 2003.
18. T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int. J. Comput. Vision*, 43(1):29–44, June 2001.
19. D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 20:91–110, 2003.
20. E. Nowak and F. Jurie. Learning visual similarity measures for comparing never seen objects. *Computer Vision and Pattern Recognition*, 1:1–8, 2007.
21. T. Pham, M. Worring, and A. Smeulders. A multi-camera visual surveillance system for tracking of reoccurences of people. *International Conference on Distributed Smart Cameras*, 1:164–169, 2007.
22. C. Piciarelli, C. Micheloni, and G. Foresti. Trajectory-based anomalous event detection. 18(11):1544–1554, 2008.
23. C. Rao, A. Ray, S. Sarkar, and M. Yasar. Review and comparative evaluation of symbolic dynamic filtering for detection of anomaly patterns. 3(2):xx–yy, 2009.
24. J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. *Computer Vision, IEEE International Conference on*, 2:1470, 2003.
25. O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. 9th European Conf. on Computer Vision*, 2006.
26. K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluation of color descriptors for object and scene recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, USA, 2008.
27. J. van de Weijer, T. Gevers, and A. Gijsenij. Edge-based color constancy. *IEEE Transactions on Image Processing*, 16(9):2207–2214, 2007.
28. V. N. Vapnik. *The nature of statistical learning theory.* Springer-Verlag New York, Inc., 1995.
29. A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13, 2006.
30. W. Zajdel, J. D. Krijnders, T. Andringa, and D. M. Gavrila. Cassandra: Audio-video sensor fusion for aggression detection. In *IEEE Int. Conf. on Advanced Video and Signal based Surveillance (AVSS)*, 2007.
31. Q. Zhou and J. K. Aggarwal. Object tracking in an outdoor environment using fusion of features and cameras. *Image Vision Comput.*, 24(11):1244–1255, 2006.
32. Z.Zivkovic and F. der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27:773–780, 2006.