Color Invariants for Person Re-Identification

Igor Kviatkovsky

Color Invariants for Person Re-Identification

Research Thesis

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science

Igor Kviatkovsky

Submitted to the Senate of
the Technion — Israel Institute of Technology
Shebat 5772 Haifa February 2012

The research thesis was done under the supervision of Dr. Amit Adam, Prof. Ehud Rivlin and Prof. Michael Heymann in the Computer Science Department.

I would like to thank Dr. Amit Adam, Prof. Ehud Rivlin and Prof. Michael Heymann for their devoted guidance and help. I would also like to thank my family and friends for their support.

The generous financial support of the Technion is gratefully acknowledged.

Contents

Abstract **1**

Abbreviations and Notations **2**

1 Introduction **3**

1.1 Background . . . 3

1.2 Our Approach . . . 5

1.2.1 Contributions . . . 7

1.3 Thesis Outline . . . 7

2 Related Work **9**

2.1 Object Recognition . . . 9

2.1.1 Invariant Properties and Feature Spaces . . . 9

2.1.2 Parts Decomposition . . . 11

2.1.3 Alignment . . . 11

2.2 Mainstream Approaches in Object Identification . . . 12

2.3 Person Re-Identification . . . 12

2.4 Color Invariance . . . 15

3 Color Invariant Signature **18**

3.1 Motivation . . . 18

3.2 Illumination Invariance . . . 19

3.2.1 Diagonal Model . . . 19

3.2.2 Normalized Color Space (rgb) . . . 21

3.2.3 Log-Chromaticity Color Space (log) . . . 22

3.2.4 Signature Invariance . . . 23

3.3 Using Shape Context . . . 27

3.4 Comparing Signatures . . . 30

3.5 Discussion . . . 31

4 Using Standard Invariants 33

4.1 Absolute Color Signatures 33

4.1.1 Histograms 33

4.1.2 Sampling and EMD 34

4.1.3 Gaussian Mixture Model 34

4.2 Color Invariants 35

4.3 Covariance Descriptor 35

5 Signature Extraction 37

5.1 Patches Extraction and Sampling 37

5.2 Re-Identification from Multiple Frames 38

6 Results 39

6.1 Experiments 39

6.1.1 Experiment 1 40

6.1.2 Experiment 2 42

6.2 Comparison With The State Of The Art 43

6.2.1 VIPeR Dataset 43

6.2.2 BOSS Dataset 45

7 Discussion 53

7.1 Major Contributions 53

7.2 Future Work 53

8 Appendix 55

8.1 Diagonal Model Validation 55

Abstract in Hebrew א

List of Figures

1.1 Some examples from our database. Notice the large variations in each person's appearance, due to uncontrolled changes in illumination, viewing direction, camera and pose. 4

1.2 Images of persons captured from four different surveillance cameras. Below each image the distributions of pixel values in *rg* and *log* chromaticity color spaces are shown. Pixels coded in red originate from the upper part (usually shirt) and those in blue from the lower part. 6

2.1 Images of the same person captured by two different cameras under various illumination conditions. Notice the change in colors. 15

2.2 Example of a dataset used for color invariants evaluation in [22]. Left: A set of reference images stored in the database. Right: Corresponding images from the query set. 17

3.1 'Color flows' in rg color space for four instances of α, β, γ parameters. Arrows describe changes in the coordinates due to illumination change. The shift of each pixel depends both on the pixel value and the parameters. 22

3.2 'Color flows' in log color space for four instances of α, β, γ parameters. Arrows describe changes in the coordinates due to illumination change. All arrows are parallel and have the same length meaning that the shift is dependent on the parameters and not on the pixel values. 23

3.3 Pixels pairs in rg color space. (a)- p_1, p_2 under illumination o. (b)-p_1, p_2 under illumination c, after applying the diagonal transformation D. 24

3.4 $Diff(L, L')$ distributions as a result of different factors. Red - varying the illumination according to the diagonal model while keeping the same surface reflectance ($L = L_1^o, L' = L_1^c$). Blue - varying the surface reflectance while keeping the same illumination ($L = L_1^o, L' = L_2^o$). Please note that the red distributions have sharp peaks near 0 while the blue distributions are much smoother. (a)- $log(\Delta l)$ distributions. (b)-$\Delta\theta$ distributions in degrees. 25

3.5 Colored objects and their chromaticity signatures. (a) - columns from left to right - reference image, rg signature, log signature. (b) - columns from left to right - query image, rg signature, log signature. The query image was taken under different illumination and from different angle. Notice the greater similarity between log rather that rg signatures. 27

3.6 Shape Context descriptor. (a) Log-polar histogram bins overlaid on top of the data points (5 bins for $log(r)$ and 12 bins for θ). (b) SC descriptor of the reference point (the center of the log-polar histogram). The darker the color the higher is the bin count. . . 28

3.7 Parts-based shape context examples. Row 1 - Pairs of images demonstrate three people imaged in two different cameras. Row 2 - observations x_i color-coded by spatial origin. We added the log-polar quantization of colorspace used in the shape context descriptor. Row 3 - shape context descriptor for a single specific point (the darker the color, the higher the value). 30

3.8 Parts-based shape context examples. Row 1 - Pairs of images demonstrate three people imaged in two different cameras. Row 2 - observations x_i color-coded by spatial origin. We added the log-polar quantization of colorspace used in the shape context descriptor. Row 3 - shape context descriptor for a single specific point (the darker the color, the higher the value). 31

3.9 Matching pairs of clusters. Notice that only part of the matches is shown for viewing convenience. Red/Blue '+' represent upper/lower parts of the first person and Red/Blue 'o' represent upper/lower parts of the second person. Matched person images are depicted on the left upper corner of the graph. (a) The results of matching images of the same person captured by different cameras. The cost of the matching is $d = 0.46$. (b) The results of matching images of the different persons. The cost of the matching is $d = 1$. 32

5.1 (a) - The bounding box of the detected person. (b) - Detected person silhouette. (c) - Two masks for differentiating upper part and lower part. (d) - The upper and lower patches from which red-coded and blue-coded observations were sampled. 38

6.1 Experiment 1. (a) - Barnard calibration boards under two different illuminations. (b) - Comparison of three approaches for re-identification with one/two parts variation. 41

6.2 Experiment 1.(a) - Comparison of PARTS-SC and SC re-identification performance for 5 different color spaces. (b) - Comparison of HI (with one/two parts variation) re-identification performance for 5 different color spaces. 42

6.3 Experiment 2. (a) - Comparison of the three approaches for re-identification with one/two parts variation. (b) - Comparison of two parts variation of *PARTS-SC* and *HI* methods with *Cov* method and their combination, the *Comb* method. 43

6.4 Experiment 2. Five top ranked candidate matches returned by *PARTS-SC* (fig. a) *HI* (fig. b) and *Cov* (fig. c) methods for twelve different queries. For each one of the figures (a)-(c), the leftmost column shows the query image and to the right of each query, five top ranked matches are displayed. Highlighted images represent a correct match. 46

6.5 Experiment 2. Five top ranked candidate matches returned by *Comb* method for twelve queries from each one of the four cameras (a)-(d) . For each one of the figures (a)-(d), the leftmost column shows the query image and to the right of each query, five top ranked matches are displayed. Highlighted images represent a correct match. In the bottom rows of figures (a)-(c) there are examples of a queries which do not have any correct match in the five top ranked candidates. 47

6.6 Examples of pedestrian image pairs from the VIPeR dataset. Note the strong variations in illumination and pose. 48

6.7 VIPeR dataset. (a) - Comparison of the three approaches for re-identification with one/two parts variation and the *GMM* approach. (b) - Comparison to the best baseline approach reported in [26]. 48

6.8 VIPeR dataset. (a) - Comparison of two parts variation of *PARTS-SC* and *HI* methods with *Cov* method and their combination, the *Comb* method. (b) - Comparison of the *Comb* method with the ELF and SDALF methods. 49

6.9 VIPeR dataset. Examples of matches using the *Comb* signature. For each one of the figures (a) and (b), the leftmost column shows the probe image and to the right of it the ten top ranked matches out of the 316 gallery images are displayed. Highlighted images represent a correct match. The last row in column (b) is an example of a query which does not have any correct match in the ten top ranked. 50

6.10 VIPeR dataset. Accuracy as a function of the sample size. Four different sample sizes were used - 50,70,85 and 100. The best results are obtained for sample size of 85. 51

6.11 Examples of passenger images in the BOSS dataset. Each row shows images viewed by different camera. Note the difference in the viewing directions of the cameras and the different illumination conditions. 51

6.12 BOSS dataset. (a) - Comparison of the three approaches for re-identification with one/two parts variation and the *GMM* approach. (b) - Comparison of two parts variation of *PARTS-SC* and *HI* methods with *Cov* method and their combination, the *Comb* method 52

6.13 BOSS dataset. Performance evaluation of spatiogram using several illumination normalization methods reported by [53] compared with Comb signature. 52

8.1 Error and distance measures of the diagonal and the linear mappings. We differentiate between three cases: 1. Both images are taken indoors. 2. Both images are taken outdoors. 3. No constraints on the scene type. 57

8.2 (a-d) Examples of images from database by Gehler *et al.* [20]. (e-f) Distribution of translation vectors due to illumination change. (a),(b) Images of indoors scenes. (c),(d) Images of outdoors scenes. (e) - Both scenes are imaged under either outdoor or indoor illumination. (f) - One scene is imaged under outdoor illumination and one is imaged under indoor illumination. Note the large differences in illumination conditions. 58

Abstract

We revisit the problem of specific object recognition using color distributions. In some applications - such as specific person identification - it is highly likely that the color distributions will be multimodal and hence contain a special structure. Although the color distribution changes under different lighting conditions, some aspects of its structure turn out to be invariants. We refer to this structure as an intra-distribution structure, and show that it is invariant under a wide range of imaging conditions while being discriminative enough to be practical. Our signature uses shape context descriptors to represent the intra-distribution structure. Assuming the widely used diagonal model, we validate that our signature is invariant under certain illumination changes. Experimentally, we use color information as the only cue to obtain good recognition performance on publicly available databases covering both indoors and outdoors conditions. Combining our approach with the complementary covariance descriptor, we demonstrate results exceeding the state of the art performance on the challenging VIPeR database.

Abbreviations and Notations

SC	—	Shape Context
PARTS-SC	—	Parts Based Shape Context
HI	—	Histogram Intersection
EMD	—	Earth Movers Distance

Chapter 1

Introduction

In this work we revisit object recognition using color. In contrast with most current research which is concerned with category recognition, here we focus on specific object recognition. Even more specifically, we are interested in video surveillance applications, where specific person recognition is an important application. This problem is also known as person re-identification and has received a lot of attention recently [31, 24, 58, 36, 28, 27, 53, 1, 11].

Person re-identification in general, and using color in particular, is very challenging. Figure 1.1 shows several examples of people imaged by different cameras in a surveillance context. In such an uncontrolled environment, appearances of the same person are highly variable due to changes in illuminations, cameras, geometry and pose. In addition, surveillance cameras are often of a low resolution.

1.1 Background

Color Invariants. In early work on color-based recognition by Swain and Ballard [52], color histograms were used as the discriminating feature between different objects. Colors were used ”as is” with no attempt to incorporate invariance to different illuminations, cameras and geometry which greatly affect the perceived colors. Funt and Finlayson [19] used indexing of color ratios computed from neighboring points instead of indexing the color values themselves, in order to achieve some invariance. Gevers and Smeulders [22] derived several color invariants, using physics-based modeling of the image acquisition process. Different invariants were designed to handle different types of variabilities in the imaging conditions. For example, the *rgb* colorspace[1] is invariant to illumination intensity and to changes in the illuminant-object-camera geometry,

[1] *rgb* color space is defined by $r = \frac{R}{R+G+B}$, $g = \frac{G}{R+G+B}$, $b = \frac{B}{R+G+B}$.

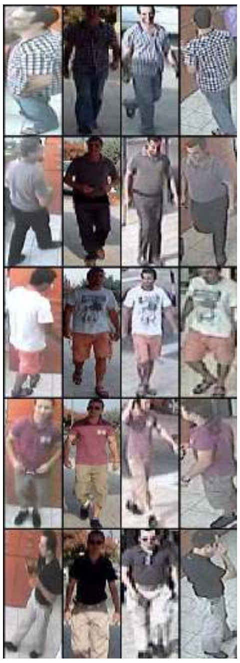


Figure 1.1: Some examples from our database. Notice the large variations in each person's appearance, due to uncontrolled changes in illumination, viewing direction, camera and pose.

but is not invariant to illumination color changes. It was recognized that no single colorspace was able to achieve invariance to all the encountered imaging conditions. These invariants were successfully used by the authors in [23, 21] for image retrieval, segmentation and tracking. In a more recent work, Gevers and Stokman [51] describe a method for selection and fusion of different color invariants, with an application to image feature detection.

Person reidentification. The challenge in person re-identification resulted in several approaches to the problem, relying on different cues. Usually color was not the only cue used. Gheissari, Sebastian and Hartley [24] combine several ideas to achieve impressive performance. They use color and structural information extracted locally around key-points to generate a discriminative and robust signature. Furthermore, they demonstrate that using spatio-temporal alignment of the object considered contributes significantly to performance. In a follow-up work by Wang *et al.* [58], improved performance is demonstrated using co-occurrence matrices of quantized appearance features and quantized shape features which correspond to object parts. In order to overcome illumination changes, Javed, Shafique and Shah [31] model explicitly the brightness transfer function between pairs of different cameras. One of the disadvantages using this approach is that a training phase is necessary to learn the brightness transfer function. Moreover, this training phase has to be repeated each time the illumination conditions change. The use of discriminative learning techniques for person re-identification is seen recently more and more often. Gray and Tao [27] used AdaBoost to select the most discriminative cues out of a large pool of color and texture features. Lin and Davis [36] learn pairwise appearance-based classifiers to separate pairs of people, using a joint color and height histogram as a feature vector. We refer in more detail and provide additional examples of the related work in Chapter 2.

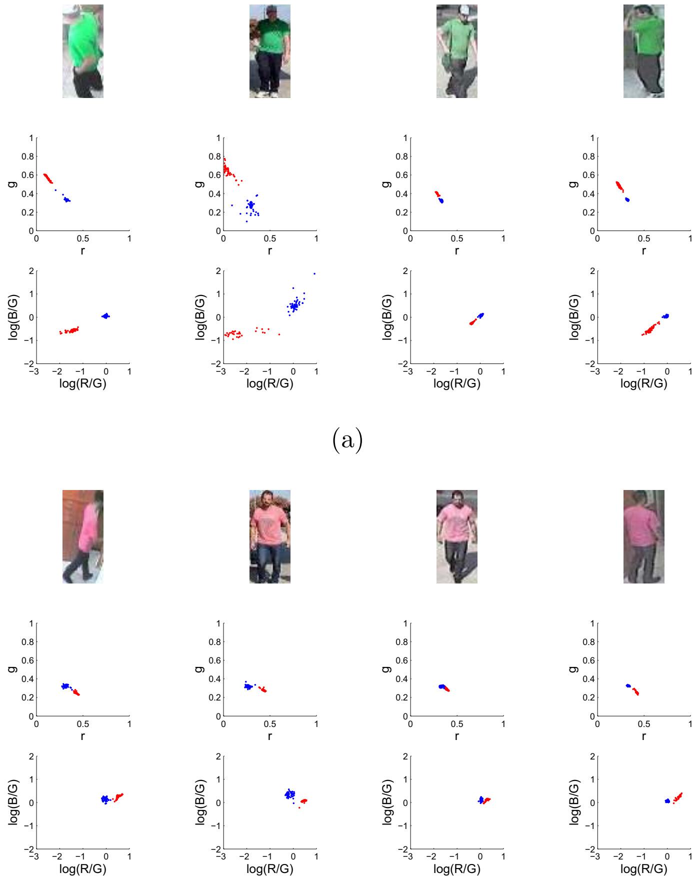
1.2 Our Approach

Returning to describing colors in an invariant way, our approach considers the distribution of observed colors in the object we try to describe. We fix the colorspace in which we work - for example the *rg* colorspace, and observe the resulting distribution of object pixel values in this colorspace. Because of the nature of our objects - people - the distributions we will observe will generally be multimodal with two significant modes or clusters. These modes/clusters correspond to different natural parts in the object - usually legs and torso. Figure 1.2 demonstrates this observation. In each of the eight examples in the figure, one may see two clear clusters of the color distribution in the *rg* and *log-chromaticity*[2] color spaces we used. We colored one cluster in red and the other in blue. The red modes arise from pixels associated with the torsos, and the blue modes arise from observations generated by the legs. We will refer to these modes/clusters as ”color clouds” in this work.

Our intuition, validated in this work, is that the shapes and relative configurations between the ”color clouds” are invariant under a wide range of imaging conditions. In addition to being invariant, we hope that these shapes and relative configurations of ”color clouds” will also be discriminative. Assuming this is true, we use shape context [5] as a non-parametric descriptor of this ”color clouds” based signature.

We remark that Matas [41, 42] argued that the relation between color patches in multicolored object is an important discriminative and invariant cue for recognition. Unlike Matas [41], who considered a problem of recognizing an arbitrary

[2] *rg* color space is defined by $r = \frac{R}{R+G+B}$, $g = \frac{G}{R+G+B}$, *log-chromaticity* color space is defined by $\xi_1 = \ln \frac{R}{G}$, $\xi_2 = \ln \frac{B}{G}$.



(a)

(b)

Figure 1.2: Images of persons captured from four different surveillance cameras. Below each image the distributions of pixel values in *rg* and *log* chromaticity color spaces are shown. Pixels coded in red originate from the upper part (usually shirt) and those in blue from the lower part.

multicolored object, we focus on a specific type of objects having a specific nature of color distribution in them. This allows us to effectively incorporate spatial information by giving a well defined meaning to clusters in color space (the upper/lower parts). Moreover, we encode the object's color content differently. Matas builds a graph based representation of object's chromatic content, the color adjacency graph (CAG), whose nodes correspond to clusters in color space and the edges represent colors spatial adjacencies. Hence, whereas Matas disregards the color distribution shape and relies on its mean value, we exploit all (or a sampled subset) of cluster points, extracting a somewhat richer description of color distributions. The choice of *rg* color space by Matas was motivated by invariance to illumination intensity and viewing geometry while changes in illumination color were not accounted for. Berwick and Lee [6] showed that *log-chromaticity* color space is more suitable than the *rg* for representing colors, assuming the diagonal model of illumination change. We adopt this finding in our work.

1.2.1 Contributions

Our approach differs from previous works in several important aspects:

1. The idea of using the intra-distribution structure as an invariant descriptor, is novel. The distributions encountered in our application usually do indeed contain discriminative structure, as a result of their multi-modal nature (due to different clothing for lower and upper parts). This is in contrast with works which considered the distribution as a whole [52, 22, 19, **?**].

2. We use non-parametric shape descriptors to describe the intra-distribution structure.

3. We apply the approach to the problem of person re-identification and demonstrate that color as a single cue does indeed have good discriminative properties, in spite of the required high invariance due to largely varying imaging conditions.

Additional contributions of this work are the experimental validation of our approach on publicly available datasets and our own privately collected dataset. Unlike most of the currently available evaluation datasets, which are designated with either indoors or outdoors person re-identification scenarios, our dataset contains images taken from two indoor and two outdoor surveillance cameras.

1.3 Thesis Outline

The outline of this work is as follows. In Chapter 2 we cover most of the relevant literature related to object recognition and identification. In Chapter 3 we explain the considerations behind the design of a color invariant signature for person re-identification. We refer to the diagonal model of illumination change and its impact on the selection of color space. In Chapter 4 we discuss the use of standard invariants for the task of person re-identification. Chapter 5 elaborates on the additional processing envelope around the signature extraction. Results and discussion follow in Chapters 6 and 7.

Chapter 2

Related Work

2.1 Object Recognition

Object recognition is one of the most challenging and broadly studied areas in the field of computer vision, referring to either *classification* or *identification*. Classification stands for assigning a certain class label, chosen from a pool of class labels, to a given object. The pool of class labels may comprise of two or more classes. In the former case the problem is called *binary classification problem* and in the later case it is called *multiclass classification problem*. Face detection is a well known binary classification problem when the first class refers to the human faces and the second class refers to anything besides the human faces. Handwritten digits recognition is an example of a multiclass classification problem when the classes are the handwritten digits 0 up to 9. Unlike classification, the identification task is about recognizing the identity of an object. For example - face recognition, as opposed to face detection, is recognizing the identity of the object in hand, given that it is a human face. Ullman claimed [55] that the task of identification is easier to perform by an artificial system than the classification task, while exactly the opposite holds for biological systems.

Three major principles are being widely used by most of the recognition systems, either on their own or in various combinations [55]. The first principle uses *invariant properties*, the second exploits *part decomposition* and the third principle is *alignment*. We will now review each one of them and present representative papers.

2.1.1 Invariant Properties and Feature Spaces

It is almost impossible to recognize an object under arbitrary viewing conditions and thereof some regularities in the object views are usually assumed. This approach assumes that several properties of the object views are preserved under the most commonly encountered transformations that the object may undergo. These properties are usually referred to as *invariant properties* or *invariant features*. Typically a feature is a real number computed using the information extracted from a single image[1]. In most cases a single feature is not enough for capturing the invariant properties of the object, thus a set of features is used. In case of n features an object's view is mapped to a single point in n-dimensional feature space R^n. It is important to ensure that this mapping can be easily computed using the information extracted from the object's view, otherwise it may turn to be as complicated as the recognition task itself. When choosing the feature space for object representation one has to bear in mind the tradeoff between its invariant and discriminative properties. For example, mapping all objects into a constant value will result in perfect invariance but no discriminative capabilities whatsoever. Rather than storing different views for each object in a database, their representations in the feature space are stored, forming subspaces in the R^n. Depending on the problem domain these subspaces may have similar or different structures. A query image is classified by assigning its representation to one of the subspaces in the features space.

Invariant features can be based on global appearance of the object, or on multiple local descriptors. The work by Lamdan and Wolfson [35] is an early example for global invariant feature. The authors use *geometric hashing* for representing object's structure in a viewpoint invariant manner and present an application of this method for efficient recognition of a 3D object from its 2D views. Apart from being invariant to a viewpoint, geometric hashing is also invariant to partial occlusions. Color histogram [52] is another example of global invariant which is invariant to scale and rotation but is not invariant to occlusions. Kliot and Rivlin [33] presented a method for efficient trademark logos retrieval based on matching shape contours. The method uses global geometric invariants in combination with local invariant signatures and is robust to various viewpoint transformations and missing shape parts.

Local descriptors have the advantage of being more robust to occlusions. In recent years many invariant local descriptors have been proposed and evaluated. A summary of these efforts has been presented in [44]. The SIFT descriptor [38] has emerged as one of the most reliable descriptors. Scale invariant locations (*keypoints*) are detected on the image by analyzing its scale space and gradient based descriptors are extracted around these keypoints. These descriptors are invariant to rotation, scale and illumination intensity changes. Recognition is done by matching SIFT descriptors extracted from a query image to descriptors stored in the database and because of the large amount of local descriptors involved, the matching process is robust to occlusions and background clutter.

[1] Notice that in general features are not restricted to be derived from a single image. Sometimes they are computed using several views or learned models.

More recently the SURF [4] descriptor has been proposed. SURF's advantage over SIFT is its computational efficiency in part due to the use of integral image [57].

Sometimes different types of invariant features are found to be useful under various conditions while a single invariant is not sufficient for good accuracy. Thus typically combinations of different descriptors are used.

2.1.2 Parts Decomposition

This approach assumes that the object can be decomposed into a set of parts or components. The recognition process begins at the part level, detecting various parts in the image. Then it proceeds with verifying the spatial relationships between the parts and recognizing or rejecting the object as a result. The parts/components can be generic ones like boxes and cylinders, or specific ones like eyes, nose and mouth in the case of face detection. Representative works employing this approach are [8, 12].

2.1.3 Alignment

Let $M = \{M_1, ..., M_n\}$ be a set of n object models stored in a database. Given an object image V, the goal is to correctly match it to one of the models. While the image is usually in 2D, the models may be either in 2D or 3D. A given image of an object may differ from all its previously seen images, thus a set of transformations T that the object model may undergo is defined to compensate for these differences. The set of allowable transformations depends on the dimension of the models and the images as well as on the problem domain. The alignment approach seeks to find the model $M_i \in M$ and the transformation $T_i \in T$ which compensate for the appearance differences between the transformed model $T_i(M_i)$ and the given object image V. The alignment process is divided into two stages. At the first stage, hypotheses are generated for possible alignment transformations between the model and the image and at the second stage these hypotheses are verified. Each hypothesis is generated using minimal amount of information, such as two pairs of corresponding points in the case of 2D planar models. Pairs of corresponding points are located using local features such as corners and holes, and only those correspondence pairs resulting in a legal alignment transformation are used to form a hypothesis. In the verification stage, the models' edges are transformed to the image plane and compared with the image edges. The best alignment is the one having the highest number of matching edges. Classic papers using this approach are [30, 37, 56].

2.2 Mainstream Approaches in Object Identification

Local descriptors, *e.g.* SIFT, are very common in identifying a specific object in still images or video. Sivic and Zisserman [50] use local descriptors for image retrieval and object recognition in video. The authors learn a visual vocabulary from a training video by extracting several types of local descriptors from its frames and clustering them. This vocabulary is later used for indexing the content of each frame in the test videos. Object recognition is done by computing the cross correlation between words frequency vectors of the query image and the indexed video frames.

2.3 Person Re-Identification

Works on person re-identification differ in approaches they exploit and usually use several approaches in combination. The invariant features approach is one of the most commonly used. Hamdoun *et al.* [28] present an approach which uses local descriptors. In this work the authors perform person re-identification by matching interest points (SURF) accumulated through short video sequences. In order to speed up the matching process the authors store the interest points in a KD-tree and match the query and model sequences by counting the number of points which fall close enough to each other. Unlike from [28] where the local invariant points are extracted using each frame independently, in [24] the authors extract key-points which are invariant in spatio-temporal domain. They use color and structural information around each key-point to generate a discriminative and robust signature. Furthermore, they demonstrate that using spatio-temporal alignment of the object considered contributes significantly to performance. In a follow-up work by Wang *et al.* [58], improved performance is demonstrated using co-occurrence matrices of quantized appearance features and quantized shape features which correspond to object parts. Further improvement is reported by Zheng *et al.* [60] where the authors demonstrate that utilizing group context information (information about the people around the individual) greatly improves the performance.

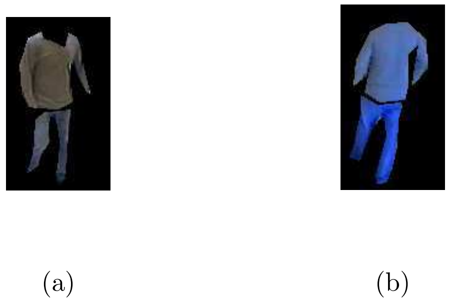
Bak *et al.* [1] propose using parts decomposition for person re-identification. In this works the authors divide the person's body into five parts (the top, the torso, legs, the left arm and the right arm) and train a detector for each part. After detection, each part is represented using the region covariance descriptor [54] while each pixel is represented by its coordinates, color and texture information. The pyramid matching [25] is used to match the parts in query and model images. Reported results are superior to those in [60].

In the recent work by Farenzena *et al.* [11] SDALF (Symmetry-Driven Accumulation of Local Features) approach was introduced. SDALF is a methodology for constructing an invariant and discriminative signature using symmetry-driven accumulation of local features. First, the person's body is divided into three parts (head, torso, legs) by computing its horizontal asymmetry axes. Then, torso's and legs' vertical symmetry axes are estimated and used for weighting the extracted features representing each part. The idea is to weight the extracted features according to their distance from the symmetry axis in order to minimize the effects of pose variation. Three types of visual cues are utilized - color histogram, MSCR (Maximally Stable Color Regions) and RHSP (Recurrent High-Structured Patches) - a novel texture descriptor. For each individual, three descriptors are extracted using these cues. The similarity between two individuals is computed using a weighted combination of distances between the descriptors. Reported results on three public datasets constitute the current state of the art.

All of the aforementioned works share the same general idea - feature set representing the person and distance measure for signatures comparison are chosen manually. When selecting the feature set one has to make sure that the representation is both discriminative and invariant. As opposed to using handcrafted features, Gray and Tao [27] proposed to use AdaBoost for selecting the most discriminating ones out of a large pool of color and texture features. An interesting insight emerged from this work - over 75 percent of the classifier weight were devoted to color based features, with the highest weight given to hue and saturation. This fact supports the assumption that color indeed is a most powerful cue for person re-identification.

Several attempts were made to tackle the person re-identification as a multi-class classification problem. In an early work by Nakajima *et al.* [46] several classification schemes for recognizing the person and his pose are presented. The authors train SVM classifiers using color-based and shape-based features and combine them using either one-vs-all or pairwise strategy to form a real time person recognition system. The features were extracted from video sequences of a number of individuals taken in a constrained environment indoors. Moreover, the evaluation dataset included only eight individuals which makes it difficult to estimate the accuracy and scalability of the approach. The color-based features used by the authors were color histograms extracted from the entire region of person's body detected using background subtraction. In order to gain invariance to illumination intensity the authors utilized the *rg* colorspace, which indeed resulted in better recognition rates than those while using the standard RGB. Lin and Davis [36] learn appearance-based classifiers to separate pairs of people. In order to enable scalability of the scheme to a large number of categories, pairwise dissimilarity profiles (function of spatial location) are learned and integrated into a nearest-neighbor classification. Experiment on a real surveillance data, including 61 individuals, showed promising results in recognition performance and scalability.

As mentioned already, perceived colors change significantly as a result of multiple cameras, illumination, geometry and pose, see Figure 2.1. In early works by Porikli [48] and Javed *et al.* [31] the authors explicitly model the brightness transfer function between different cameras to compensate for illumination variations. However, the main drawback of this approach is the assumption of being able to measure the cameras brightness response in advance. Most recent works on person re-identification do not rely on being able to perform these calibration steps, but rather design descriptors which are inherently invariant to photometric changes. Color histogram is the most commonly used descriptor for representing color distribution and hence many efforts were put to make it more robust to illumination variations. Madden *et al.* [39] proposed to use the histogram equalization technique [16], to reduce the effect of illumination variations on histograms representing person's appearance. After equalizing the histograms of each RGB channel independently, the authors suggest to represent the target by its major colors clusters' means, rather than by its joint RGB histogram. In a follow up work [40] Madden *et al.* compare the effectiveness of histogram equalization technique to several other methods *i.e.* histogram stretching and illumination filtration and conclude that histogram equalization has the best performance. The main drawback of histogram representation is that colors spatial origin is lost. This may result in an incorrect match of two persons, one wearing for example a red top and blue pants and the other one a blue top and red pants. Park *et al.* [47] proposed partitioning of the detected person's silhouette into three parts (the head, the torso and the legs) and representing the person's appearance based on two histograms summarizing the colors extracted separately from the torso and the legs. Yu *et al.* [59] proposed to use a joint 4D histogram of color and spatial features. The spatial feature they use is the novel *path-length* feature which is the length of the shortest path from a reference point, chosen to be the top of the head, to the pixel. The color features chosen by the authors were the *color rank features*. Given a set of sampled pixels, the color rank features encode the relative value of a pixel in each one of the R, G and B channels separately, ignoring their absolute values. In [16] it is shown that this relative ranking is preserved under a wide range of illumination changes. In a follow up work by Lin and Davis [36], the authors replace the path-length feature with a much simpler feature - the normalized height - which is a normalizing vertical coordinate of a pixel. Even though this feature is simpler and more computationally efficient than the path-length feature, the authors achieved better recognition performance. Truong Cong *et al.* [53] evaluate the performance of three signatures, varying in the way they exploit spatial information, for the task of re-identifying a person between a pair



(a) (b)

Figure 2.1: Images of the same person captured by two different cameras under various illumination conditions. Notice the change in colors.

of cameras installed on a train. The authors compare an RGB histogram with an RGB-path-length joint histogram and a spatiogram [7] combined with several illumination normalization techniques including the histogram equalization described earlier. Their conclusion is that spatial information and illumination normalization techniques, especially histogram equalization, indeed contribute to the re-identification performance.

2.4 Color Invariance

Although color is a powerful cue for recognizing the identity of the object, differences in illumination cause measurements of object colors to be biased towards the color of the light source. Fortunately, humans have the ability of color constancy: they perceive the same color of an object despite large differences in illumination. Much efforts have been invested in developing automatic color constancy algorithms, which use illuminant estimation procedures, see [29] for an overview. In these procedures, the illuminant is estimated given the image data and appropriate corrections are made to this data to make it illumination invariant. Contrary to the above are the color invariant approaches where color constancy is achieved by transforming the pixels data to explicitly derived color spaces. Assuming a certain model for image formation process, invariant properties are mathematically proved for these color spaces. In [13] by Finlayson *et al.* the image is summarized using three angles computed between three color channels of the image (stretched out as vectors). This description is illumination invariant but not discriminative enough to deal with large number of objects. In early work on color-based recognition by Swain and Ballard [52], color histograms were used as the discriminating feature between different objects. The approach worked well as long as illumination stayed fixed, but when illumination changed, its performance degraded drastically. It was suggested to preprocess the image using color constancy methods to gain some illumination invariance. But involving color constancy algorithms significantly reduces the simplicity and efficacy of the approach, and no single algorithm was found to perform well enough. Funt and Finlayson [19] used indexing of color ratios computed from adjacent pixels instead of indexing the color values themselves, and achieved much better illumination invariance.

Color invariants preserving the original image structure are sometimes called invariant color spaces since each pixel RGB value independently undergoes an algebraic transformation mapping it to different color space while preserving the original structure of the image. Gevers and Smeulders [22] derived several such color invariants, using physics-based modeling of the image acquisition process. Different invariants were designed to handle different types of variabilities in the imaging conditions. For example, the rgb color space is invariant to illumination intensity and to changes in the illuminant-object-camera geometry, but is not invariant to illumination color changes. It was recognized that no single color space was able to achieve invariance to all the encountered imaging conditions. Berwick and Lee [6] suggested using log-chromaticity color space to compute a signature which is invariant to illumination color, assuming a diagonal model of illumination change [14]. This signature was used for image retrieval and detection of specularites in objects' images. Finlayson and Hordley [17] used the log-chromaticity color space to derive a single color coordinate, a function of RGB values, depending on surface reflection only. Histograms based on this single invariant coordinate demonstrated superior results in color based object recognition comparing to chromaticity histograms.

Most of the above mentioned works applied color invariants to recognize objects such as branded products imaged under constrained conditions, see Figure 2.2. Several attempts were made to use them in less constrained environment [23, 21, 51]. Typically these works use features based on object's global appearance *e.g.* histograms of color invariants. Matas *et al.* [42], on the other hand, introduce a local color invariant feature for image retrieval and object recognition. First, image neighborhoods having a multimodal color distribution are detected using the mean shift algorithm [9]. From each color mode several color invariants are computed and are used jointly to describe the neighborhood. Concatenated vector of the invariants extracted from color distribution modes in the neighborhood is referred to as a Multimodal Neighborhood Signature (MNS) of this neighborhood. Thus, an image or an object in the image are represented as a set of signatures, and the recognition task is carried out by matching test image signatures to those of stored models. The authors decided on which color invariants to use assuming the diagonal model of illumination

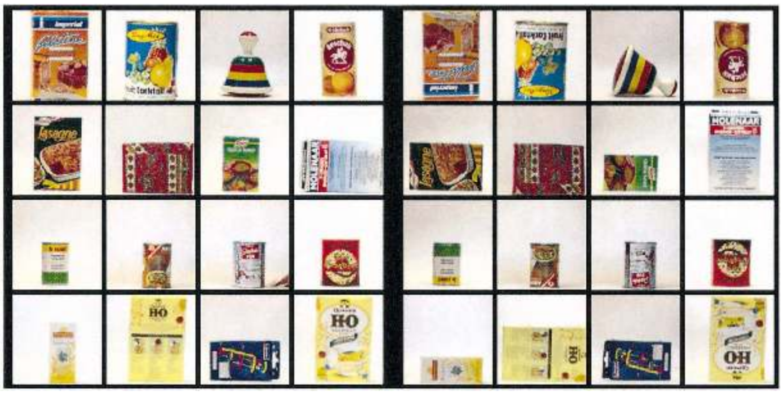


Figure 2.2: Example of a dataset used for color invariants evaluation in [22]. Left: A set of reference images stored in the database. Right: Corresponding images from the query set.

change. Impressive results in recognizing objects in videos shot both indoors and outdoors support the choice of the diagonal model for coping with such diverse illumination conditions.

In this work we evaluate the use of color invariants for person re-identification both indoors and outdoors, and introduce our own invariant color based signature for person re-identification.

Chapter 3

Color Invariant Signature

3.1 Motivation

In this chapter we describe the color based invariant signature for person re-identification. First we review the diagonal model for illumination change and discuss the invariant properties of several chromaticity colorspaces. Then we introduce the signature describing the distribution of colors in person's clothing and prove its invariance under illumination changes assuming the diagonal model.

Figure 1.2 shows different images of two people taken from different surveillance cameras. Below each image we show the corresponding distribution of object colors, in *rg* and *log chromaticity*[1] color spaces. We mark the observations generated by the upper part of the object in red, while the observations generated by the lower part of the object are marked in blue. One may note the following:

1. All the distributions have multimodal structure in them. Our coloring of observations emphasizes this structure. Generally we may see two modes in the distribution - we refer to these as "color clouds" or clusters.

2. The structure of the "color clouds" is sufficiently preserved for the same person.

3. For different people, the structures of the color clouds are different when we consider also the spatial origin of the clouds (notice that red and blue switched places).

These observations motivated us to consider a signature based on the shape of the "color clouds" which constitute the color distribution of an object. As

[1] See next section for *rg* and *log chromaticity* colors spaces definitions.

noted previously, in our surveillance application distributions will indeed contain multimodal shape.

Why do we expect the intra-distribution structure to be invariant under wide imaging conditions ? We list here a few intuitive arguments:

- The number of modes corresponds to the number of different colors observed in the object, and this is a strong invariant.
- Relative positioning of modes - if one part of clothing is more "red" than another, then under most encountered transformations - it will stay more "red" than the other part. This was validated in practice for a wide range of illuminants and imaging devices, assuming the diagonal model of illumination change [16].
- If one piece of clothing is much more uniform in color than another (leading to a more condensed mode or color cloud), then likewise under most encountered transformations, it will stay more condensed than the other mode.
- If we think of a "color cloud" as elliptic, then its orientation will not change significantly. This is due to the distribution of actual values in the diagonal model that are encountered in practice [2].

These arguments are very intuitive and may be considered more of a speculation. Thus, we will now present more rigorous considerations involved in designing the signature. In section 3.2 we elaborate on choosing the most suitable invariant color space under the assumption of the diagonal model of illumination change. In section 3.3 we introduce the signature which captures the shape of the "color clouds" for person re-identification.

3.2 Illumination Invariance

3.2.1 Diagonal Model

The majority of works on color invariants [22, 19, 13, 17, 6] assume that the image acquisition process can be described using the following model:

$$\rho_k = \int_\omega E(\lambda)S(\lambda)Q_k(\lambda)d\lambda \quad (k = 1, 2, 3) \tag{3.1}$$

where $Q_k(\lambda)$ is a function of wavelength λ, characterizing the proportion of color signal absorbed by the sensor k. The color signal is basically a product of the illuminant energy function $E(\lambda)$, measuring the amount of energy the illumination source emits at each wavelength λ, and surface reflection function

$S(\lambda)$, characterizing the surface reflecting properties. Integration is performed over the visible spectrum ω (400-700nm). ρ_k are the responses of three color sensors of the given imaging device. Most commonly these are the well known RGB color channels.

A common goal is to determine the reflectance $S(\lambda)$ of a point in the scene given its image, namely a pixel value comprising of three sensor responses ρ_k. Suppose that we know the sensor spectral response functions $Q_k(\lambda)$. But given a sensor response ρ_k, there are many combinations of possible functions $S(\lambda)$ and $E(\lambda)$ that could account for it, resulting in a severely underconstrained problem. Assuming, as usually done, that these functions are not continuous but are sampled at a discrete set of points, Eq. 3.1 becomes

$$\rho_k = \sum_{i=1}^{n} E(\lambda_i)S(\lambda_i)Q_k(\lambda_i)\Delta\lambda \quad (k = 1, 2, 3) \tag{3.2}$$

where n and $\Delta\lambda$ are the number of sample points and a sampling interval, respectively. Each surface is characterized by a discrete reflectance function $S(\lambda_i), i = 1, ..., n$. Assuming we have m distinctly colored surfaces, denote $S_j(\lambda_i)$ a reflectance at wavelength λ_i by a surface j when $j = 1, ..., m$. Hence, the number of reflectance parameters is nm. Assuming that the illumination, $E(\lambda_i)$, is constant through the imaged scene, we have n illumination parameters. Thus, the total number of unknowns *i.e.* illumination and reflectance parameters is $n(m+1)$ and the number of measurements *i.e.* sensor responses is $3m$. Typically n is much larger than 3, thus the problem is still underconstrained.

In many practical applications it is important to be able to compare images of points in the scene independently to the illumination, rather than determining the reflectance $S(\lambda)$ of these points explicitly. To solve this, one needs to determine a mapping that transforms RGB responses to an object imaged under some reference illumination c to corresponding responses under another illumination o. The most widely used mapping type is a linear transformation. Even more specifically, the majority of applications adopt the *diagonal model* of illumination change which has been proved by Finlayson *et al.* [14] to suffice for illumination variations encountered in practice. The diagonal model is a simple relation between pixel values of object imaged under some reference illumination c and another illumination o, expressed as a scaling of each color channel independently. The assumption under which the diagonal model holds is that the sensors of imaging device are sufficiently narrow-band, otherwise a special *sharpening transformation* [15] may be applied to make them more narrow-band. If camera's sensor response functions were completely narrow-band *i.e.* sensitive to a single wavelength, the diagonal model would model the illumination variation with perfect accuracy. But since in practice such sensors do not exist, the diagonal model provides an approximation which is precise to the extant that the sensors are narrow-band. A simple diagonal model may be written as a multiplication of RGB vector by a diagonal matrix D,

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \gamma \end{pmatrix} \begin{pmatrix} R^o \\ G^o \\ B^o \end{pmatrix}. \tag{3.3}$$

Despite its simplicity, good results were reported while using the diagonal model for illumination change indoors and outdoors [14, 22, 42]. Throughout this work we will adopt the diagonal model for illumination change. Thus from now on, unless stated otherwise, the term "illumination invariance" will refer to "illumination invariance under the assumption of diagonal model for illumination change". Section 8.1 presents experimental results validating the model under various changes in illumination conditions both indoors and outdoors.

We can easily notice that the original RGB color space is not invariant to applying the diagonal model, unless D is an identity. We will now review two chromaticity color spaces introduced in [22, 6], the normalized *rgb* and the *log*-chromaticity spaces, and describe their invariant properties.

3.2.2 Normalized Color Space (rgb)

The normalized color space is defined as:

$$r = \frac{R}{R+G+B},\ g = \frac{G}{R+G+B},\ b = \frac{B}{R+G+B} \tag{3.4}$$

Assuming that the diagonal matrix D in Eq. 3.3 is of the form $D = sI$, *i.e.* all three channels are scaled by the same factor s, the pixel rgb coordinates remain unchanged:

$$\begin{pmatrix} r^c \\ g^c \\ b^c \end{pmatrix} = \begin{pmatrix} \frac{sR^o}{sR^o+sG^o+sB^o} \\ \frac{sG^o}{sR^o+sG^o+sB^o} \\ \frac{sB^o}{sR^o+sG^o+sB^o} \end{pmatrix} = \begin{pmatrix} r^o \\ g^o \\ b^o \end{pmatrix} \tag{3.5}$$

Hence, the rgb color space is invariant to changes in the illuminant intensity. However, assuming D is of the form $D = diag(\alpha, \beta, \gamma)$, *i.e.* each channel is scaled differently, the transformed rgb coordinates are $r^c = \frac{\alpha R^o}{\alpha R^o+\beta G^o+\gamma B^o}$,$g^c = \frac{\beta G^o}{\alpha R^o+\beta G^o+\gamma B^o}$ and $b^c = \frac{\gamma B^o}{\alpha R^o+\beta G^o+\gamma B^o}$, meaning that the rgb color space is not invariant to changes in illuminant color. Due to the equality $b = 1 - r - g$, the b coordinate is redundant and therefore is typically disregarded. Figure 3.1 demonstrates the shifts in pixel rg coordinates as a result of four different changes in the illuminant color. We can see that the coordinates 'shift' depends on the illumination parameters and on the pixel value.

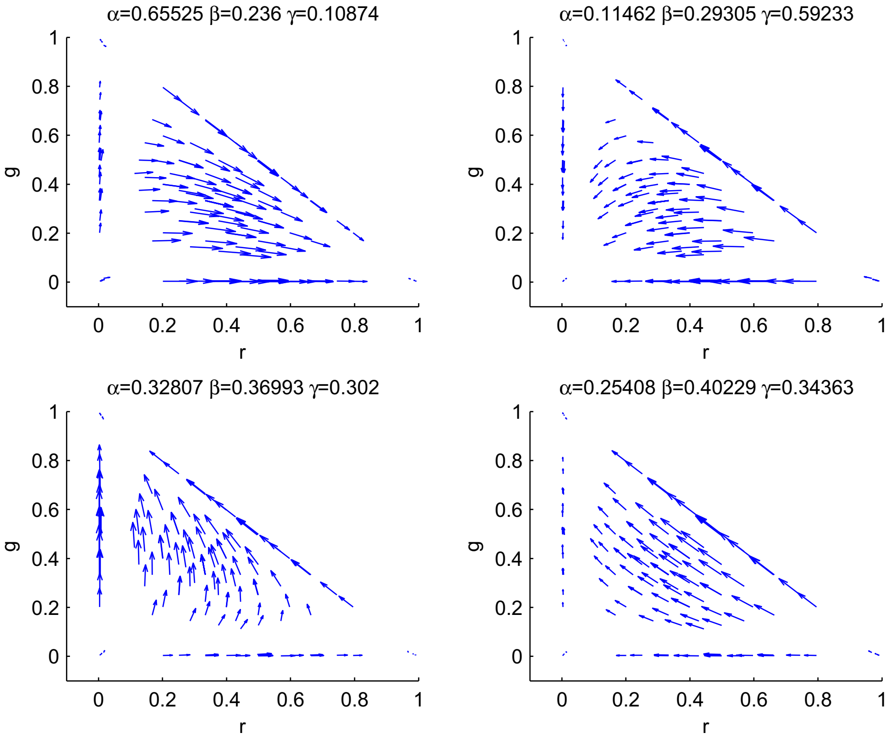


Figure 3.1: 'Color flows' in rg color space for four instances of α, β, γ parameters. Arrows describe changes in the coordinates due to illumination change. The shift of each pixel depends both on the pixel value and the parameters.

3.2.3 Log-Chromaticity Color Space (*log*)

The log-chromaticity color space is defined as:

$$\xi_1 = \ln \frac{R}{G}, \qquad \xi_2 = \ln \frac{B}{G} \tag{3.6}$$

Applying the diagonal transformation, Eq. 3.3, on a pixel $(R^o, G^o, B^o)^T$ yields a shift in the *log* coordinates:

$$(\xi_1^c, \xi_2^c) = (\ln \frac{\alpha R^o}{\beta G^o}, \ln \frac{\gamma B^o}{\beta G^o}) = (\xi_1^o, \xi_2^o) + (\ln \frac{\alpha}{\beta}, \ln \frac{\gamma}{\beta}) \tag{3.7}$$

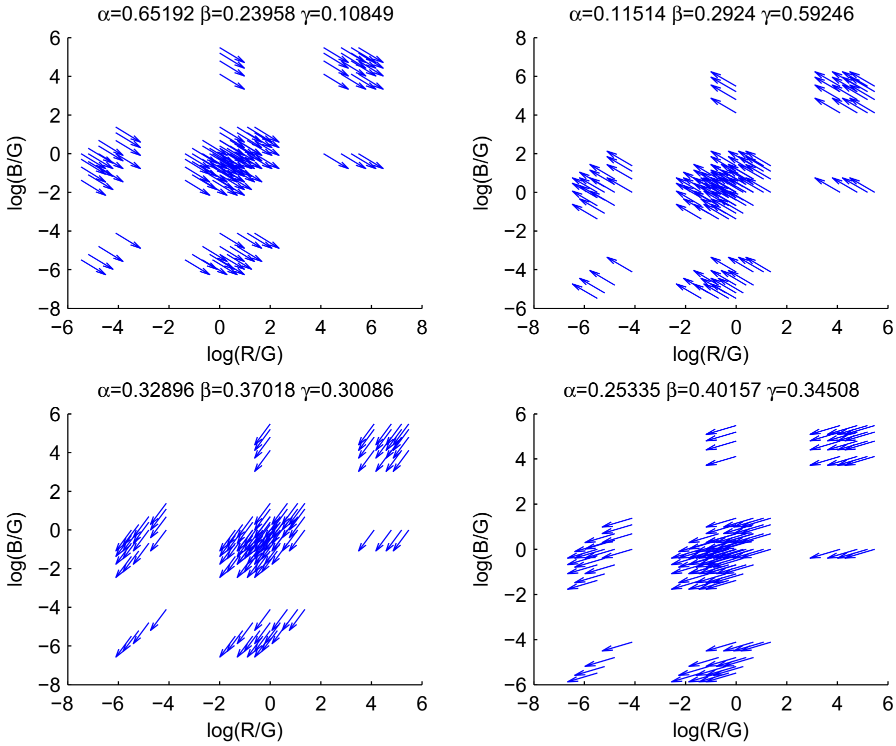
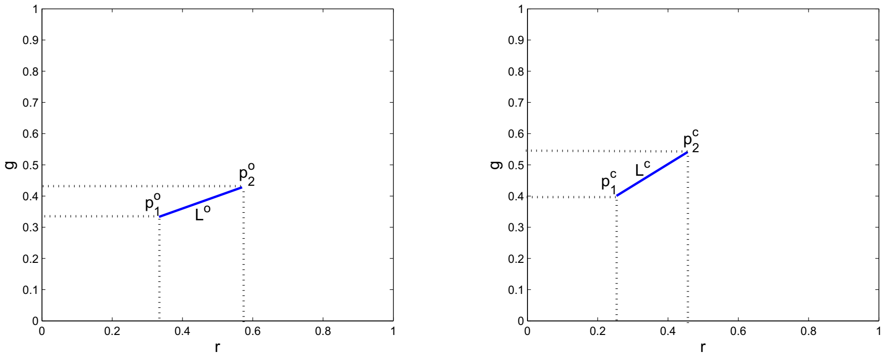


Figure 3.2: 'Color flows' in *log* color space for four instances of α, β, γ parameters. Arrows describe changes in the coordinates due to illumination change. All arrows are parallel and have the same length meaning that the shift is dependent on the parameters and not on the pixel values.

The shift in coordinate values is determined by the actual values in D and is independent of the pixel values, meaning that the *log* color space is invariant to illumination intensity and color up to translation. Figure 3.2 demonstrates the shifts in pixel *log* coordinates as a result of four different changes in the illuminant color.

3.2.4 Signature Invariance

As previously mentioned we are basing our signature on the shape of the multimodal distribution of color measurements taken from two parts of person's clothing - the legs and the torso. We suggest to represent this shape by a set of vectors in color space connecting the points originating in one part, *e.g.* legs,



(a) (b)

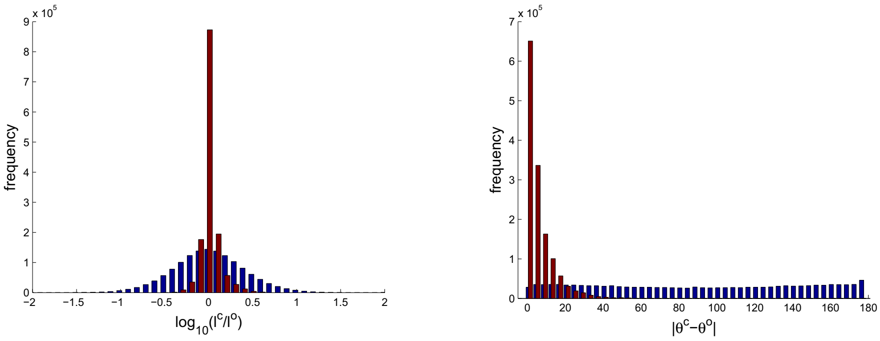
Figure 3.3: Pixels pairs in rg color space. (a)- p_1, p_2 under illumination o. (b)-p_1, p_2 under illumination c, after applying the diagonal transformation D.

to the points of second part, *e.g.* torso. We will now analyze the invariant properties of such vectors induced by each one of the color spaces (rg and log).

rg Color Space

We will start our examination with the rg color space which is not invariant to illumination color change, as shown in section 3.2.2. Let $p_1^o = (r_1^o, g_1^o)$, $p_2^o = (r_2^o, g_2^o)$ denote a pair of pixels sampled from two distinctly colored patches, which may represent the upper and lower parts of person's clothing imaged under some arbitrary illumination conditions o. Figure 3.3(a) depicts the pixels pair as two points in rg color space. Let L^o denote the vector connecting these two points, *i.e.* $L^o = p_2^o - p_1^o$. Similarly, let $p_1^c = (r_1^c, g_1^c)$, $p_2^c = (r_2^c, g_2^c)$ denote a pair of pixels imaging the same patches under different illumination conditions c. The vector L^c is connecting p_1^c and p_2^c, *i.e.* $L^c = p_2^c - p_1^c$, see Figure 3.3(b). We assume that Eq. 3.3 models the transformation between the illumination o and the illumination c.

We measure the variation in "color cloud" shape using the rotation and scaling of the vectors set. Thus, we represent these vectors in polar coordinates, $L = (l, \theta) = (\sqrt{(r_2 - r_1)^2 + (g_2 - g_1)^2}, \arctan(\frac{g_2 - g_1}{r_2 - r_1}))$. Let $(\Delta\theta, \Delta l)$ denote the



(a) (b)

Figure 3.4: $Diff(L, L')$ distributions as a result of different factors. Red - varying the illumination according to the diagonal model while keeping the same surface reflectance ($L = L_1^o, L' = L_1^c$). Blue - varying the surface reflectance while keeping the same illumination ($L = L_1^o, L' = L_2^o$). Please note that the red distributions have sharp peaks near 0 while the blue distributions are much smoother. (a)- $log(\Delta l)$ distributions. (b)-$\Delta\theta$ distributions in degrees.

difference between the vectors L^o and L^c:

$$\begin{aligned} Diff(L^o, L^c) &= (\Delta l, \Delta\theta) \\ &= (\frac{l^c}{l^o}, |\theta^c - \theta^o|) \\ &= (\frac{\sqrt{(r_2^c - r_1^c)^2 + (g_2^c - g_1^c)^2}}{\sqrt{(r_2^o - r_1^o)^2 + (g_2^o - g_1^o)^2}}, \left|\arctan(\frac{g_2^c - g_1^c}{r_2^c - r_1^c}) - \arctan(\frac{g_2^o - g_1^o}{r_2^o - r_1^o})\right|) \end{aligned}$$

Figure 3.4 depicts the empirical distribution of $\Delta\theta$ and Δl over some of the commonly encountered diagonal transformations for illumination published by Barnard [2] and by Gehler *et al.* [20]. $\Delta\theta$ measures the degree by which the vector L^o is rotated as a result of applying the diagonal transformation and Δl measures the scaling factor by which L's magnitude is multiplied.

Two distributions are depicted on each graph. The red one describes variations in vectors connecting pairs of pixels imaged under two different illuminations while the blue one describes the differences in vectors connecting two different pairs of pixels under the same illumination. Although L's orientation and scale are not completely preserved as a result of applying diagonal transformations, they are sufficiently preserved for basing the invariant properties upon them.

log Color Space

Vector orientation and scale invariance under a diagonal transformation can be rigorously proved in case of *log* color space. Similarly to *rg* color space we define two pairs of points $p_1^o = (\xi_1^{1^o}, \xi_2^{1^o}),\ p_2^o = (\xi_1^{2^o}, \xi_2^{2^o})$ and $p_1^c = (\xi_1^{1^c}, \xi_2^{1^c}),\ p_2^c = (\xi_1^{2^c}, \xi_2^{2^c})$ in *log* color space referring to pairs of pixels sampled from two colored patches under illumination o and c respectively. Likewise, we define the vectors L^o and L^c connecting these pairs of points. We will now see that these vectors are identical assuming the diagonal model Eq. 3.3:

$$\begin{aligned}
L^c &= p_2^c - p_1^c \\
&= (\xi_1^{2^c} - \xi_1^{1^c}, \xi_2^{2^c} - \xi_2^{1^c}) \\
&= (\ln \frac{R_2^c}{G_2^c} - \ln \frac{R_1^c}{G_1^c}, \ln \frac{B_2^c}{G_2^c} - \ln \frac{B_1^c}{G_1^c}) \\
&= (\ln \frac{\alpha R_2^o}{\beta G_2^o} - \ln \frac{\alpha R_1^o}{\beta G_1^o}, \ln \frac{\gamma B_2^o}{\beta G_2^o} - \ln \frac{\gamma B_1^o}{\beta G_1^o}) \\
&= (\ln \frac{R_2^o}{G_2^o} - \ln \frac{R_1^o}{G_1^o} + \ln \frac{\alpha}{\beta} - \ln \frac{\alpha}{\beta}, \ln \frac{B_2^o}{G_2^o} - \ln \frac{B_1^o}{G_1^o} + \ln \frac{\gamma}{\beta} - \ln \frac{\gamma}{\beta}) \\
&= (\ln \frac{R_2^o}{G_2^o} - \ln \frac{R_1^o}{G_1^o}, \ln \frac{B_2^o}{G_2^o} - \ln \frac{B_1^o}{G_1^o}) \\
&= (\xi_1^{2^o} - \xi_1^{1^o}, \xi_2^{2^o} - \xi_2^{1^o}) \\
&= p_2^o - p_1^o = L^o
\end{aligned}$$

This proves that the vectors L in *log* color space are invariant to anisotropic scaling caused by the illumination change.

Relying on the presented analysis we conclude that *log* color space is more suitable than *rg* for describing the shape of multimodal color distributions. Moreover, a work by Berwick and Lee [6] fully supports this choice of color space. The authors presented an object recognition approach based on describing the object in terms of its color distribution. The distributions are represented as binary masks and compared using cross correlation. Reported experiments confirmed that an object's signature in *log* color space is much more robust to variations in imaging conditions than its signature in *rg* color, see Figure 3.5. The main difference from our work is that Berwick and Lee encode the color distribution shape of the whole object while we encode the relation between distribution shapes corresponding to different object parts.

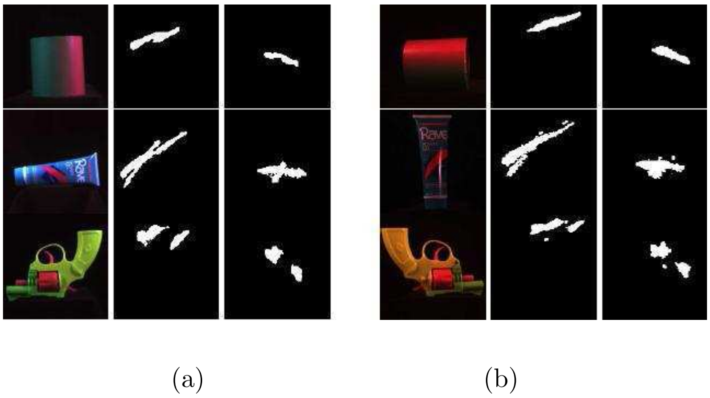


Figure 3.5: Colored objects and their chromaticity signatures. (a) - columns from left to right - reference image, *rg* signature, *log* signature. (b) - columns from left to right - query image, *rg* signature, *log* signature. The query image was taken under different illumination and from different angle. Notice the greater similarity between *log* rather that *rg* signatures.

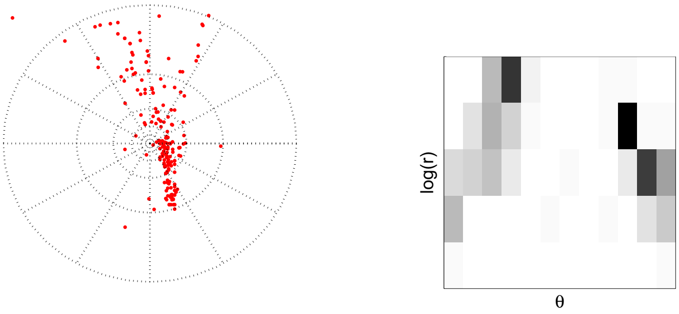
3.3 Using Shape Context

In [5] Belongie *et al.* introduced an alignment method for shape matching. They define a novel local descriptor which incorporates global shape information, called *Shape Context (SC)*, and find correspondences between points on two shapes using this descriptor. Given a set of correspondences a proper alignment transformation is obtained. This method was successfully applied for matching sampled contours of binarized letters and therefore proved itself useful for shape based object recognition.

For a set of points (without loss of generality think of them as points on a plane), the shape context descriptor at a given reference point is a log-polar histogram centered at this point, counting the number of the remaining points falling in each bin, see Figure 3.6. Such a histogram captures the spatial distribution of the remaining points with respect to the reference point. In this work we use the shape context descriptor to describe intra-distribution structures in color space.

Assume that we are given a set of color observations $O = \{x_1, \ldots, x_N\}$ (in some specific color space) that were extracted from an object. We will differentiate between two cases:

1. The observations are labeled with binary spatial information: each observation x_i is labeled with $l_i = 1$ if it came from the upper part of the object, and with $l_i = 0$ if it came from the lower part of the object. "Upper" and



(a)

(b)

Figure 3.6: Shape Context descriptor. (a) Log-polar histogram bins overlaid on top of the data points (5 bins for $log(r)$ and 12 bins for θ). (b) SC descriptor of the reference point (the center of the log-polar histogram). The darker the color the higher is the bin count.

”lower” have a well-define meaning we will describe later on.

2. The observations are given without any spatial information.

We will extract two different signatures corresponding to these two cases. In the first case, let $O_U = \{x_i | l_i = 1\}$ denote those observations generated from the upper part of the object. Let $O_L = \{x_i | l_i = 0\}$ be the actual observations generated by the lower part of the object. Denote by $sc(x, O)$ the shape context descriptor of the points in the set O with respect to the reference point x. The parts-based shape context signature is

$$\mathrm{PARTS - SC(O_L, O_U) = \{sc(x, O_U) | x \in O_L\}} \tag{3.8}$$

In other words, we encode the distribution of upper-part colors, with respect to colors appearing in the lower part of the object. This signature captures the shape of the upper-part color cloud, the shape of the lower-part color cloud and the relative positioning of the two color clouds.

In the second case in which we have no spatial information at all, we define

$$SC(O) = \{sc(x, O) | x \in O\} \tag{3.9}$$

where O is the set of observations we have. We note that this is the standard use of shape context for shape descriptions (e.g. [5]). However, in our work the set O is made of points in color-space, whereas usually points in O are spatial points (for example, contours of binarized letters).

In chapter 5 we will explain how we obtained the spatial information required for extracting the *PARTS-SC* signatures.

We note that the shape context descriptor includes a normalization of the radial distances between the reference point and the set of other points. To achieve scale invariance, all radial distances are divided by the mean distance between all point pairs in the shape. We included this normalization too and thus obtained invariance to contrast or dynamic range of observed colors.

Figure 3.7 shows several examples of parts-based shape context. The first row shows different views of several people from the database. Below each image we have the observations x_i color-coded according to their label l_i, with red coding observations from the upper part of the image ($l_i = 1$ or $x_i \in O_U$). We show the log-polar quantization of the color space, centered on one of the blue points (the closest to the center of mass of the blue cloud). The third row of the figure shows the shape context descriptor for this specific blue point. The actual *PARTS-SC* signature is the collection of all such descriptors for all the blue points of a given image. Looking at the second row, please note that the structure of color clouds corresponding to two images of the same person, are quite similar. In comparison with Figure 1.2 we may see here increased similarity due to the normalization described above.

In section 3.2.4 we have discussed the invariant properties of log and rg colorspaces and came to the conclusion that the log colorspace is more suitable for describing the shapes of color distributions assuming the diagonal model. We would like to remind that the log color space is not invariant to translation, which results under the diagonal model assumption, but due to the inherent translation invariant nature of the shape context descriptor, this disadvantage vanishes.

We would like to stress that the *PARTS-SC* signature encodes the relation between colors in the target rather than the absolute color values. Looking at figure 3.8, please note that the fourth and the fifth columns from the left, depict two persons dressed completely differently but having very similar *PARTS-SC* signatures. We will show that, as expected, combining *PARTS-SC* with signature encoding the absolute color values will result in an enhanced descriptor. Looking at the third and fourth columns of figure 3.8 reveals that the *PARTS-SC* signature is not necessarily invariant to all illumination changes. Notice the different angle between the red and the blue color clouds resulting in different signatures extracted for the same person imaged under two different illuminations. The reason for this difference is that the diagonal model is not always capable of precisely describing the illumination change.

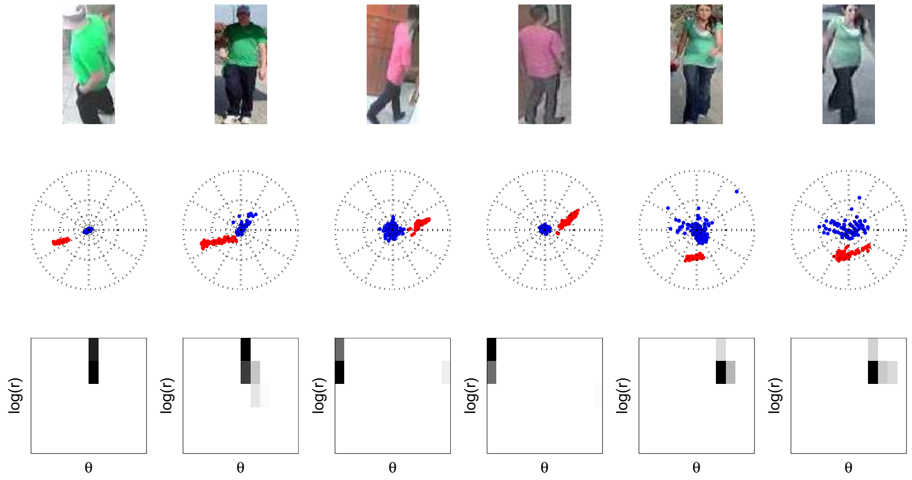


Figure 3.7: Parts-based shape context examples. Row 1 - Pairs of images demonstrate three people imaged in two different cameras. Row 2 - observations x_i color-coded by spatial origin. We added the log-polar quantization of colorspace used in the shape context descriptor. Row 3 - shape context descriptor for a single specific point (the darker the color, the higher the value).

3.4 Comparing Signatures

A signature S extracted from a given image is a set of shape context descriptors we will generically denote by $S = \{\mathrm{sc_i}\}$. If we are using *PARTS-SC* then i runs only on the indices of observations with lower spatial origin. If we are using the *SC* signature, then i runs over all observations.

We now describe how to compute the distance between two signatures $S = \{s_1, \ldots, s_N\}$ and $S' = \{s'_1, \ldots, s'_N\}$. (As explained in section 5.1, we always have the same number N of descriptors in each signature). Let C_{ij} be the Chi-squared distance between s_i and s'_j. Recall that each descriptor s_i, s'_j is a histogram over the log-polar bins we placed on colorspace. The matrix C with entries C_{ij} describes the cost matrix of matching shape context descriptors from the first signature S with those from the second signature S'. We define the distance between the two signatures as the minimal cost of matching their elements:

$$d(S, S') = \text{The minimal cost of matching all elements in}$$
$$S \text{ with all elements in } S\text{' using the cost matrix } C.$$

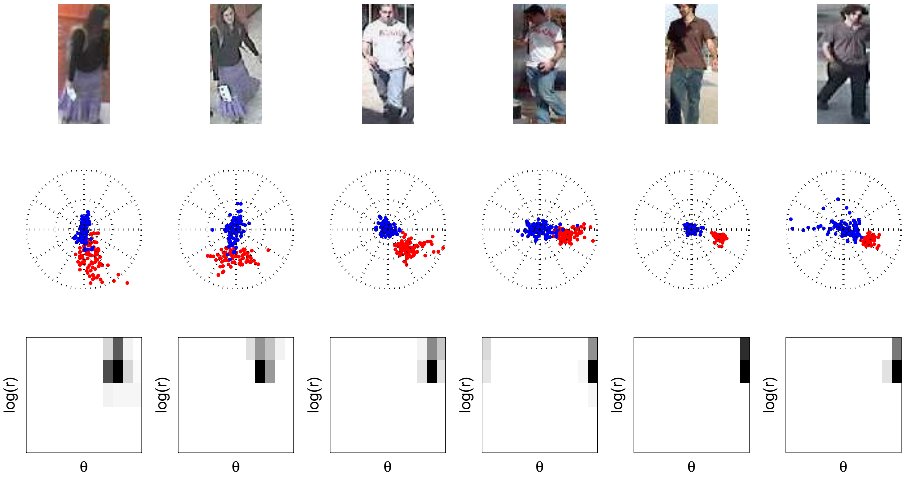


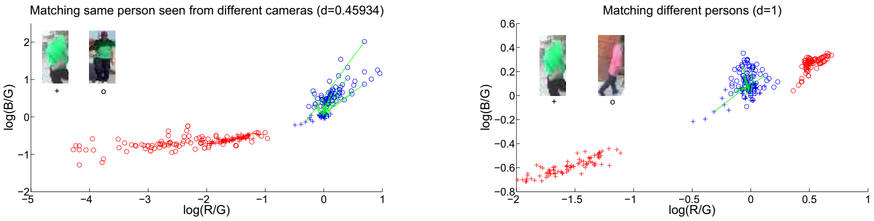
Figure 3.8: Parts-based shape context examples. Row 1 - Pairs of images demonstrate three people imaged in two different cameras. Row 2 - observations x_i color-coded by spatial origin. We added the log-polar quantization of colorspace used in the shape context descriptor. Row 3 - shape context descriptor for a single specific point (the darker the color, the higher the value).

We compute the minimal cost by using Rubner's EMD code[2] [49]. Figure 3.9 shows the actual optimal matching between two signatures. The first person's observations are marked with a + sign, and o marks the second person's observations.

3.5 Discussion

Encoding the red points distribution with respect to different blue points describes the red distribution shape seen from various locations of the blue distribution. Thus, the multiplicity of such reference points ensures that the shape of the blue distribution is encoded as well. Obviously, the more reference points are used, the more precise is the description of the blue distribution shape. Nevertheless, in case of a simple distribution shape, many reference points will result in redundant descriptors. Moreover, a large number of reference points, N, will negatively affect the running time of the procedure computing the distance between signatures. Recall that the computation of minimum cost perfect matching is of complexity $O(N^3)$. In chapter 6 we show the impact of N on the

[2] The EMD code is available at http://ai.stanford.edu/∼rubner/



(a) (b)

Figure 3.9: Matching pairs of clusters. Notice that only part of the matches is shown for viewing convenience. Red/Blue '+' represent upper/lower parts of the first person and Red/Blue 'o' represent upper/lower parts of the second person. Matched person images are depicted on the left upper corner of the graph. (a) The results of matching images of the same person captured by different cameras. The cost of the matching is $d = 0.46$. (b) The results of matching images of the different persons. The cost of the matching is $d = 1$.

re-identification accuracy evaluated on the VIPeR dataset.

Chapter 4

Using Standard Invariants

In this chapter we will summarize some of the most commonly used color based signatures for object identification. The majority of signatures encode the distribution of absolute color values either in the original color space (RGB) or in one of the invariant color spaces [22]. The distribution may be encoded using a histogram, a collection of sampled values or a parametrized model *e.g.* Gaussian Mixture Model. Some signatures encode ratios of color values, to describe the object color in illumination invariant way. Examples of such signatures may be found in [19] and the *m1m2m3* signature in [22]. While these signatures are extracted from pairs of neighboring pixels and are most typically used on regions borders, the *PARTS-SC* signature describes the relation between colors of two remote regions.

4.1 Absolute Color Signatures

4.1.1 Histograms

Color histogram is the most widely used technique for color distribution description, since first introduced by Swain and Ballard [52]. In their work the authors suggested to use *histogram intersection* as a similarity measure between the histograms. Given two normalized color histograms h_1 and h_2, each containing n bins, a histogram intersection similarity measure between them is defined as:

$$H(h_1, h_2) = \sum_{k=1}^{n} \min\left(h_1(k), h_2(k)\right) \tag{4.1}$$

Since the histograms are normalized *i.e.* sum of all their elements equals 1, the histogram intersection similarity measure may be transformed to a distance measure in the [0..1] range:

$$d_{HI}(h_1, h_2) = 1 - \sum_{k=1}^{n} \min\left(h_1(k), h_2(k)\right) \tag{4.2}$$

The main advantage of a histogram intersection is its high speed. The main disadvantage though is that it is a bin wise metric, and as such is sensitive to the fixed histogram quantization. Cross bin metrics such as the EMD (Earth Mover's Distance) [49] may be used to overcome this downside, but on the expense of a significant reduction in speed.

We have used a pair of color histograms in a *log* color space to describe each part of the person clothing - the upper and the lower parts. We used the histogram intersection method for measuring the distance between a pair of histograms corresponding to each part. The overall distance between representations is computed by summing the lower part distance and the upper part distance.

4.1.2 Sampling and EMD

Another way of describing the color distribution in a target is by uniformly sampling N pixels inside its silhouette boundaries and using the collection of N pixel values as a descriptor. We have evaluated this approach as well, representing the target's upper and lower parts as a collection of N points in *log* color space. We have used EMD to measure the distance between the collections of points, while the collection itself represents the signature and the cost matrix is calculated using the Euclidean distance in the *log* color space.

4.1.3 Gaussian Mixture Model

An alternative way for modeling the color distributions is a Gaussian Mixture Model (GMM), a widely used parametric technique for modeling data distributions. For example in [32] the authors used GMM for representing the color distribution in person's clothing for person re-identification. Since we assume that the color distribution has roughly two modes - the clothes upper and lower parts - we may use two unimodal Gaussians rather than the more general GMM. We fit one Gaussian to each one of the upper and the lower parts of the target. Each Gaussian is represented by the mean μ and the covariance matrix Σ. Given a pair of d-dimensional Gaussians $g_1 = (\mu_1, \Sigma_1)$ and $g_2 = (\mu_2, \Sigma_2)$, the distance between them is calculated using the Kullback-Leibler divergence [34]:

$$d_{KL}(g_1||g_2) = \frac{1}{2}[tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1}(\mu_2 - \mu_1) - log(\frac{|\Sigma_1|}{|\Sigma_2|}) - d] \tag{4.3}$$

For symmetrizing the distance measure we will use the following distance:

$$d(g_1, g_2) = d_{KL}(g_1||g_2) + d_{KL}(g_2||g_1) \tag{4.4}$$

Given a pair of Gaussians describing the person's upper/lower parts appearance, the distance between the overall appearance descriptions is simply the sum of differences of the parts. We have evaluated this approach in the 2D *log* color space.

4.2 Color Invariants

Besides the chromaticity color spaces discussed in the previous chapter (rg and log), several other color spaces were proposed by Gevers and Smeulders [22], for example:

- **c1c2c3**: $c_1 = \arctan \frac{R}{\max(G,B)}, c_2 = \arctan \frac{G}{\max(R,B)}, c_3 = \arctan \frac{B}{\max(R,G)}$
- **l1l2l3**: $l_1 = \frac{(R-G)^2}{(R-G)^2+(R-B)^2+(G-B)^2}, l_2 = \frac{(R-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2}, l_3 = \frac{(G-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2}$
- **o1o2**: $o_1 = \frac{R-G}{2}, o_2 = \frac{B}{2} - \frac{R+G}{4}$

We have evaluated the *PARTS-SC* performance in these color spaces, on the calibration boards of Barnard [2], and compared them to the performance obtained when using a regular histograms, as suggested in [22]. The results are described in Experiment 1 in chapter 6. The results show that the *PARTS-SC* signature performs better than the histogram based signature in all tested color spaces, suggesting that signature based on relations between colors are sometimes more discriminative than those based on the color absolute values.

4.3 Covariance Descriptor

The region covariance descriptor for object recognition was first introduced in [54]. Each pixel in the image is represented by a point in feature space. Possible features are the spatial coordinates of the pixel, its color and gradients, just to name a few. A region R in the image consisting of n pixels is described using the covariance matrix of the corresponding feature points $\{z_k\}_{k=0}^{n}$

$$C_R = \frac{1}{n-1} \sum_{k=1}^{n} (z_k - \mu)(z_k - \mu)^T \tag{4.5}$$

where μ is a mean value of z_k. An image is described by a set of covariance descriptors corresponding to image regions. Comparison between covariance descriptors is done using a metric for covariance matrices [18]

$$d_{Cov}(C_1, C_2) = \sqrt{\sum_{i=1}^{d} \ln^2 \lambda_i(C_1, C_2)} \tag{4.6}$$

where $\lambda_i(C_1, C2)$ are the generalized eigenvalues of C_1 and C_2 computed from $\lambda_i C_1 u_i = C_2 u_i$ where $u_i \neq 0$ are the generalized eigenvectors.

The region covariance descriptor has proved to be useful for texture classification and specific object recognition. It was also successfully applied for person re-identification in a parts based approach [1] and in combination with invariant color spaces [43].

We have evaluated a signature based on the region covariance descriptor. The features we use to describe each pixel are its color in the original RGB colorspace and its vertical spatial coordinate in respect to the bounding box:

$$z_i = [R(x_i, y_i), G(x_i, y_i), B(x_i, y_i), y_i] \tag{4.7}$$

where x_i, y_i are the pixel coordinates and $R(\cdot,\cdot)$, $G(\cdot,\cdot)$ and $B(\cdot,\cdot)$ are its RGB color values. We found that adding x_i coordinate to the feature vector reduces the accuracy. A probable explanation may be the symmetry of color distribution with respect to the vertical axis which is typically present in clothing we wear. Another conclusion was that using feature space based on color invariants rather than the original RGB values resulted in a reduced performance.

We have also found that using only those pixels belonging to the most dominant color of each part when computing the covariance descriptor improves the accuracy of the method. Therefore we find the largest segment of each part using Mean Shift clustering [9] and use only pixels belonging to this segment. The covariance descriptor signature provides us with an additional aspect of the appearance of person's clothing. It captures the texture missed by the signatures describing the absolute colors and the relation between colors. Therefore combining this signature with the other two is expected to enrich the description.

Chapter 5

Signature Extraction

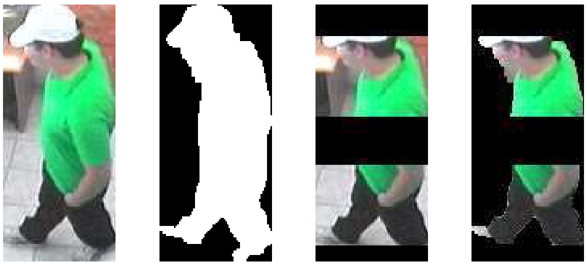
We now outline the additional processing envelope around the signature *i.e.* steps applied from the moment a surveillance video or image is obtained until the signature is extracted. First we describe how we obtain the spatial information required for the *PARTS-SC* signature, and then how we exploit multiple frames acquired for each person.

5.1 Patches Extraction and Sampling

Bounding boxes and silhouettes of the people in our experiments were extracted from the surveillance videos semi-automatically, and resized to a fixed template size. A fully automatic procedure for obtaining these bounding boxes and silhouettes may be based on a combination of pedestrian detectors [10, 61] and background subtraction. A robust solution for this part is not straightforward and is out of the scope of this work. We remark that our semi-automatic extraction of the silhouettes was not perfect and indeed does not have to be perfect.

Once the silhouette of a person is obtained, we extract two patches created by the silhouette intersection with each one of the fixed masks shown in Figure 5.1. These masks were defined in order to minimize effects of mixing colors from different clothing articles, or partial inclusion of head or feet into the patches. The patch intersecting with the upper mask generates the observations in O_U (those color-coded red) and the patch which intersects the lower mask generates the blue-colored observations.

For computational efficiency we do not use all the observations, but a random sample of $N = 85$ observations from each part. This sample size was empirically chosen to optimize the tradeoff between signature expressibility, robustness and computational efficiency.



(a) (b) (c) (d)

Figure 5.1: (a) - The bounding box of the detected person. (b) - Detected person silhouette. (c) - Two masks for differentiating upper part and lower part. (d) - The upper and lower patches from which red-coded and blue-coded observations were sampled.

5.2 Re-Identification from Multiple Frames

Every person viewed by each camera is represented by a number of images (frames). Let $I_{p,c} = \{I^i_{p,c}\}^K_{i=1}$ denote the collection of K images representing person number p captured by camera number c. We define the distance between two person/camera pairs, (p_1, c_1) and (p_2, c_2) as a *median* distance between images representing those pairs

$$D(I_{p_1,c_1}, I_{p_2,c_2}) = median\{\{d(I^i_{p_1,c_1}, I^j_{p_2,c_2})\}^{K_1}_{i=1}\}^{K_2}_{j=1} \tag{5.1}$$

Distance between images, $d(\cdot,\cdot)$, is the distance between signatures extracted from them. Using median distance provides robustness to descriptor failures.

Chapter 6

Results

6.1 Experiments

We present experimental validation of our approach on four databases. The first is a controlled database publicly provided by Barnard [2]. In that database colored patches were imaged under different illuminations in controlled experiments. We compiled the second database by extracting images from actual surveillance videos collected in an uncontrolled environment. It contains images of 31 people taken from 4 cameras under varying illumination conditions. Then we compare our approach to those published by others using two publicly available datasets - the VIPeR [26] and BOSS [53].

In our experiments, the criteria we use to measure recognition performance is similar to [24, 58, 53, 27, 11]. We report the results using cumulative match characteristic (CMC) [45]. The CMC is a curve plotting the probability of the correct match to be in the first n top ranked matches, for every n. This evaluation technique is motivated by the surveillance scenario where a query target, represented by sequence of images captured by a particular camera, is input to the system. The system then has to return matching targets (from all the videos of other people in all cameras) ranked in descending order of their similarity to the query.

All experiments are carried out using the log color space and a fixed set of parameters for all our signatures. We have used 120 bins for shape context descriptor (6 bins for $log(r)$ and 20 bins for θ). When building the 2D joint histogram in log color space we have used a uniform partitioning of the space into 10 bins. We have found that for person re-identification purposes, the most informative region in the log color space is the $[-1..1] \times [-1..1]$ square, thus we partition this region into 10 bins. Values outside this range are counted by the closest bordering bin of the histogram.

We compare the performance of several approaches:

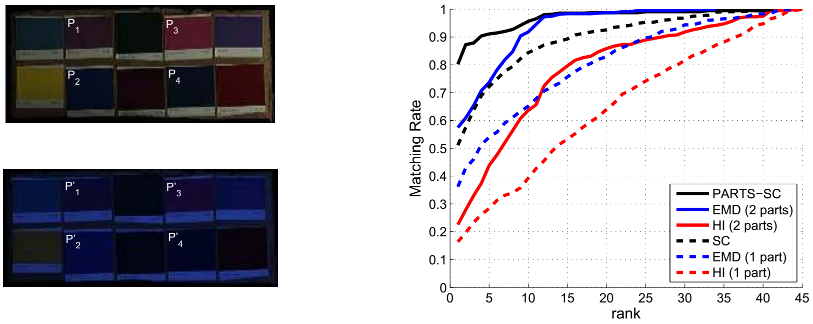
1. **HI** - the target signature is a histogram over *log* colorspace. Signatures are compared using histogram intersection [52] as a distance measure as described in section 4.1.1.
2. **EMD** - the target signature is the actual observations in *log* colorspace. Two signatures are compared using the Earth Mover's Distance (EMD) [49] between them. The cost matrix C_{ij} is computed as Euclidean distance between points p_i and p_j in *log* colorspace.
3. **SC** - the target signature is the *SC* signature without using spatial information regarding the origins of the observations.
4. **PARTS-SC** - the target signature is the *PARTS-SC* signature (using spatial information regarding the origins of the observations).
5. **Cov** - the target signature is the covariance descriptor. Signatures are compared using the metric for covariance matrices defined in section 4.3.
6. **GMM** - the target signature is a Gaussian describing the color distribution in *log* colorspace. Signatures are compared using the KL divergence as described in section 4.1.3.
7. **Comb** - the target signature is a combination of three signatures - the *HI*, *PARTS-SC* and *Cov*. Signatures are compared using the average distance between the component signatures:

$$D_{Comb}(S_1,S_2)=\frac{1}{3}(D_{HI}(h_1,h_2)+D_{PARTS-SC}(SC_1,SC_2)+D_{Cov}(C_1,C_2)) \tag{6.1}$$

Regarding the *HI* and *EMD* methods, we evaluate each of them in two variations. One is as described above, and the other uses the parts division of the target as follows: a separate signature is extracted for the lower part and for the upper part, and distances between corresponding signatures are added. For *Cov*, *GMM* and *Comb* signatures we evaluate only the parts based variation. Overall, we evaluate nine possible approaches - three approaches are evaluated in two variations and three approaches are evaluated using only one variation.

6.1.1 Experiment 1

In this experiment we evaluated the invariance of SC based signature to severe illumination changes. The database (provided by Barnard [2]) contains 11 images of a color calibration board. In all the images the viewing position is preserved. The illumination changes significantly (see Figure 6.1(a) for example).

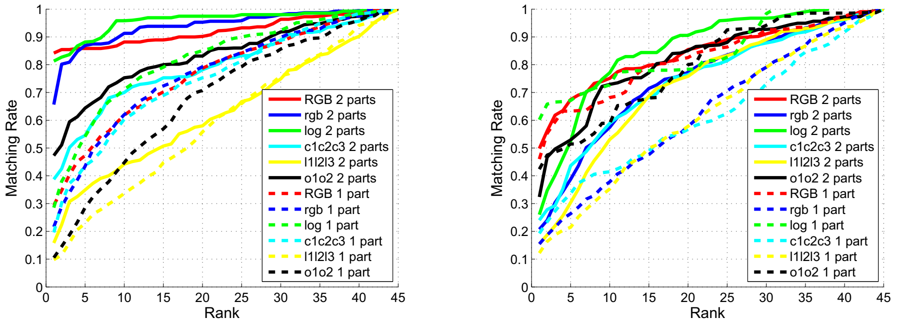


(a) (b)

Figure 6.1: Experiment 1. (a) - Barnard calibration boards under two different illuminations. (b) - Comparison of three approaches for re-identification with one/two parts variation.

We defined each column on the calibration board images as a target. For example, (P_1, P_2),(P_3, P_4),(P'_1, P'_2) and (P'_3, P'_4) on Figure 6.1(a) are different targets. Therefore we have $5 \times 11 = 55$ targets in the database. For a given query (specific column under a specific illumination) we removed from the database all its other images except one. The images of the different columns were all included. We then ranked all these 45 objects (44 incorrect matches and one correct) by their distance from the query object. Figure 6.1(b) shows the matching rate for our three methods and two variations (parts-based and single component) of each. As can be seen, *PARTS-SC* signature has the best performance with 80% of queries resulting in top ranked correct match. *EMD (2 parts)* is second with approximately 57% of the queries resulting in top ranked correct match and 92% in the ten top ranked. For each method the parts-based approach outperforms the single component approach. The *HI* method under-performs significantly with respect to the *SC* and *EMD* methods.

We have evaluated the *PARTS-SC* signature performance on several color spaces besides the *log* color space. Figure 6.2(a) shows the performance on the *RGB*, *rg*, *c*1*c*2*c*3, *l*1*l*2*l*3 and *o*1*o*2 color spaces. Figure 6.2(b) shows the performance for the same color spaces using the *HI* signature. We can see that the performance of both *PARTS-SC* and *HI* signatures is better when used with the *log* color space than with alternative color spaces.



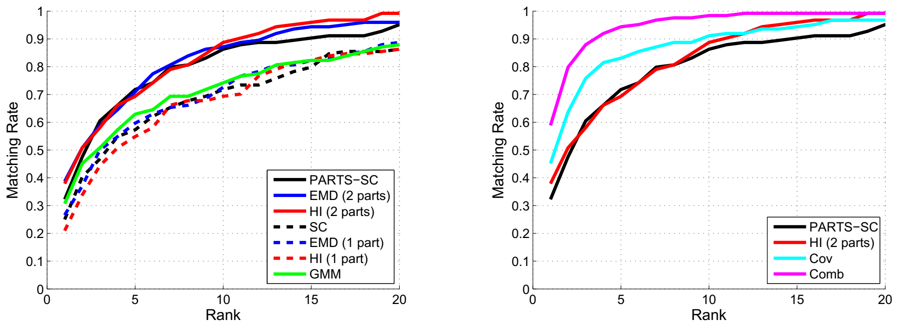
(a) (b)

Figure 6.2: Experiment 1.(a) - Comparison of PARTS-SC and SC re-identification performance for 5 different color spaces. (b) - Comparison of HI (with one/two parts variation) re-identification performance for 5 different color spaces.

6.1.2 Experiment 2

The database in this experiment contains images of 31 different individuals captured across four different cameras - a low resolution outdoor surveillance camera, two indoor surveillance cameras and a personal high quality camera. Please see Figure 1.1 for an impression of the variations in appearance. Every person in each view is represented by four to ten images extracted from the videos. The distance between two pairs of person/camera is defined in section 5.2.

Figure 6.3(a) shows the matching rate for the methods used in Experiment 1 and the *GMM* method. Again parts-based approaches perform better than the corresponding single component approaches. The parts based *EMD*, *HI* and *PARTS-SC* approaches perform relatively similarly. For each one of the three parts based approaches about 87% of queries result in a correct match within the first ten targets. We can see that the *GMM* signature performance is lower than the first three parts based signatures. Figure 6.3(b) shows the matching rate for the *PARTS-SC*, *HI* and *Cov* methods and for their combination, the *Comb* method. The *Cov* method has the best performance out of the first three methods, and the combined method significantly improves the overall performance. Figure 6.4 shows the five top ranked candidates for a number of queries, using the *PARTS-SC HI* and *Cov* methods. Different methods give different ranks to the candidates emphasizing the difference in aspects captured by each one of them. Each horizontal section in figure 6.4 features those queries resulting in a top ranked correct match using one of the methods while other



(a) (b)

Figure 6.3: Experiment 2. (a) - Comparison of the three approaches for re-identification with one/two parts variation. (b) - Comparison of two parts variation of *PARTS-SC* and *HI* methods with *Cov* method and their combination, the *Comb* method.

methods fail. The first, second and the third horizontal section show those queries resulting in a top ranked correct match using the *PARTS-SC HI* and *Cov* methods respectively. As expected, the combined method *Comb* is significantly better than any of its components, since it benefits from aspects captured by all of the three. Figure 6.5 shows five top ranked candidates returned by the *Comb* method for twelve exemplar queries generated by each one of the four cameras.

6.2 Comparison With The State Of The Art

6.2.1 VIPeR Dataset

VIPeR[1] is an evaluation dataset for viewpoint invariant person recognition introduced by Gray *et al.* [26]. The dataset contains 632 image pairs of pedestrians taken from arbitrary viewpoints under varying illumination conditions. All the images are normalized to a fixed size of 128×48 pixels, see examples in figure 6.6.

VIPeR is considered the most challenging dataset for person re-identification due to significant changes in illumination and pose, relatively low resolution and very limited information for modeling the pedestrian appearance. The authors have also published performance evaluation of several baseline approaches including joint histograms, concatenated histograms and hand localized histograms introduced in [47], just to name a few. For evaluation, pedestrians in

[1] VIPeR dataset is available at http://vision.soe.ucsc.edu/?q=node/178

the dataset were split into training and testing sets using random partitions. Each set had 316 pedestrian image pairs, while each pair order (which image in the pair is the gallery and which one is the probe image) was randomly chosen.

In a follow up work by Gray and Tao [27] the ELF (Ensemble of Localized Features) approach was proposed, suggesting to use feature selection to automatically choose the most discriminating features out of a large pool of color and texture features. ELF by far outperformed the baseline methods reported in [26]. The recently published SDALF (Symmetry-Driven Accumulation of Local Features) approach [11] made a step further, improving the ELF performance. We follow the methodology described in [26, 11] while evaluating the performance of our approach on VIPeR. In order to make the comparison to SDALF as close as possible we have averaged the results over the exact same 10 partition sets of pedestrians used for evaluation of SDALF[2].

Figure 6.7 shows the performance of our color based signatures and the baseline approaches reported in [26]. Our *PARTS-SC* and *HI* approaches by far outperform the best baseline approach, hand-localized histogram, even though they utilize roughly the same information - upper and lower parts color. Figure 6.8 shows the performance of the *PARTS-SC*, *HI* and the *Cov* approaches and their combination - the *Comb* signature. Afterwards the *Comb* approach is compared to the ELF and SDALF. *Comb* outperforms SDALF with 23% of the queries resulting in the the top ranked correct match, and 58% in the ten top ranked versus 50% by SDALF. Error bars on the *Comb* CMC curve indicate the standard deviation from the mean computed over the 10 partition sets.

Figure 6.9(a) depicts several examples of the top ranked matches returned by the *Comb* signature, and Figure 6.9(b) depicts several examples where the correct match was obtained in the first ten candidate. For viewing convenience only the top 10 candidate out of 316 were displayed.

Since PARTS-SC and EMD methods use sampled data, we checked how the sample size affects the accuracy of these methods. Figure 6.10 shows the normalized area under the CMC curve obtained using four different sample sizes. Based on these results, we have chosen the sample size of 85 for our experiments. At first glance it might seem that sample size impact on the accuracy is not very strong because of relatively small differences (0.01 order of magnitude). Note though that 0.01 difference in the normalized area under the CMC curve may result in addition of 316*0.01=3.16 units to the effective area, significantly improving the performance.

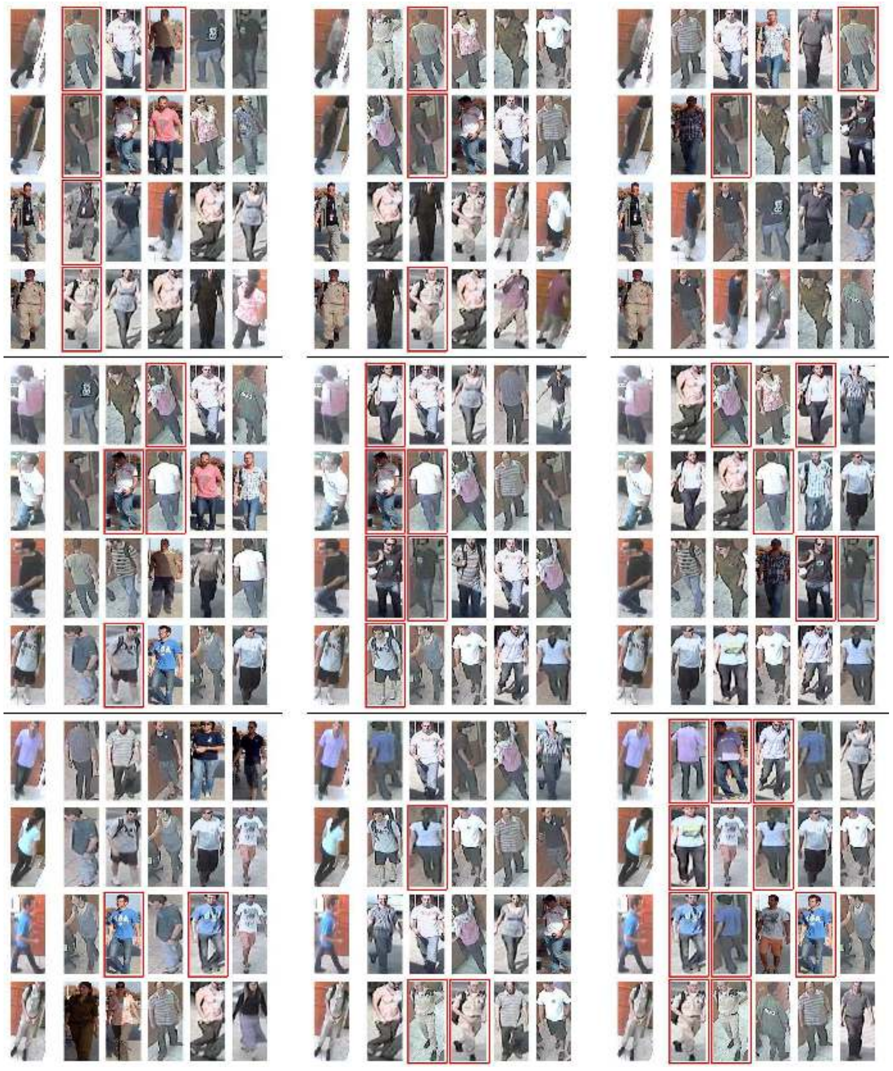
[2] Evaluation partition sets are available at http://www.lorisbazzani.info/code-datasets/sdalf-descriptor/

6.2.2 BOSS Dataset

The BOSS project dataset[3] contains video sequences shot by onboard cameras in a suburban train. We used these video sequences for capturing images of 35 passengers viewed from two cameras. Indoors and outdoors illumination changes, caused by a moving train, induce high variations on passengers appearance between the cameras. Figure 6.11 shows several examples.

Truong Cong *et al*. [53] have used the BOSS dataset for evaluation of three color signatures, differently exploiting the color and spatial information. The evaluated signatures were - RGB histogram, RGB-path-length joint histogram and a spatiogram [7]. Each signature was combined with several illumination normalization techniques. We have evaluated our signatures using the same methodology as in [53], Figure 6.12 shows our results. The *Comb* signature returns a correct top ranked match in 77% of the queries and 97% of the queries result in the correct match appearing in two top ranked results. According to experiments reported in [53], the spatiogram signature outperformed the other two, thus we used it for comparison with our Comb signature. Figure 6.13 shows the comparison of *Comb* with the spatiogram signature. The best method slightly outperforms *Comb*, but because of a small number of persons involved (35), it is hard to draw a reliable conclusion based on this experiment.

[3] BOSS project dataset is available at http://www.celtic-boss.org/



(a) (b) (c)

Figure 6.4: Experiment 2. Five top ranked candidate matches returned by *PARTS-SC* (fig. a) *HI* (fig. b) and *Cov* (fig. c) methods for twelve different queries. For each one of the figures (a)-(c), the leftmost column shows the query image and to the right of each query, five top ranked matches are displayed. Highlighted images represent a correct match.

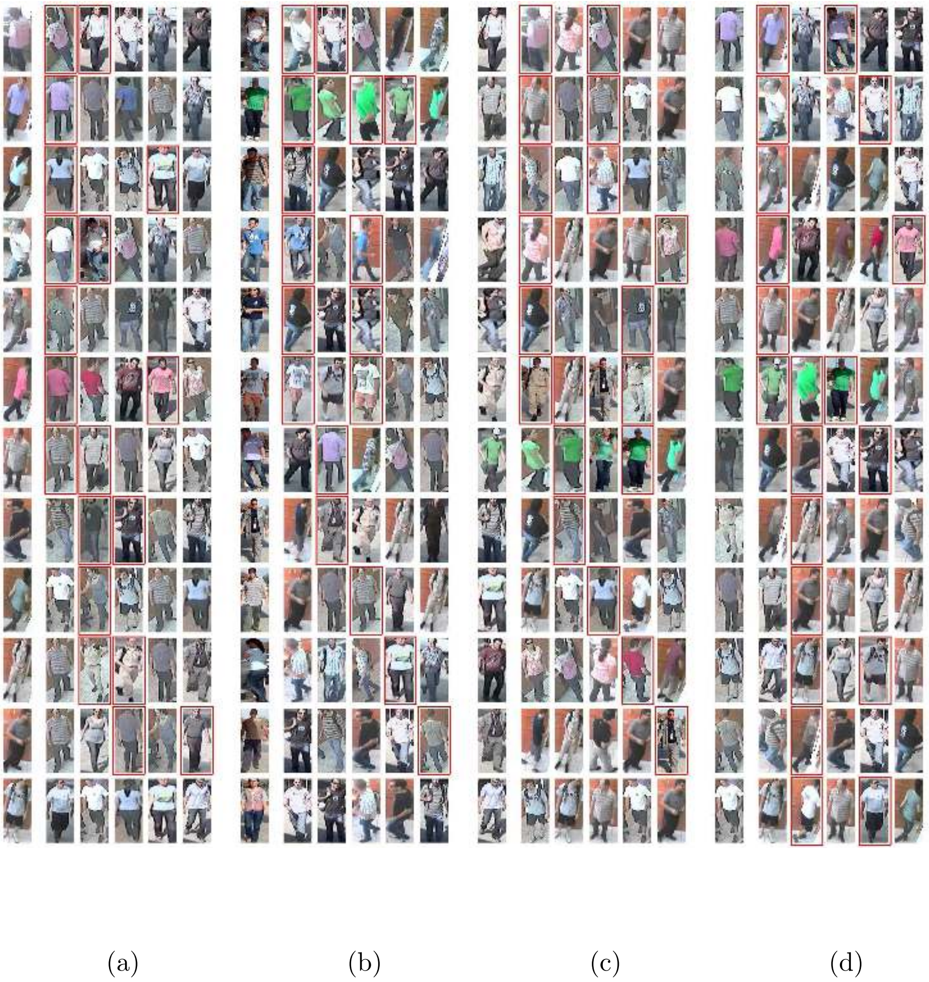


Figure 6.5: Experiment 2. Five top ranked candidate matches returned by *Comb* method for twelve queries from each one of the four cameras (a)-(d) . For each one of the figures (a)-(d), the leftmost column shows the query image and to the right of each query, five top ranked matches are displayed. Highlighted images represent a correct match. In the bottom rows of figures (a)-(c) there are examples of a queries which do not have any correct match in the five top ranked candidates.

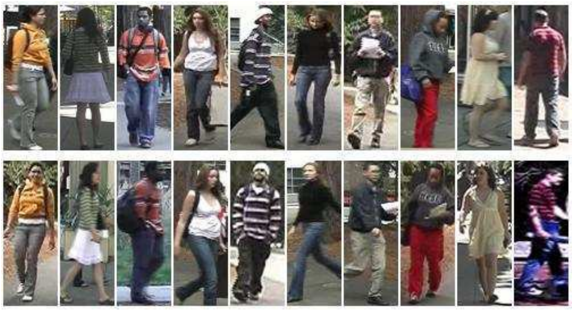
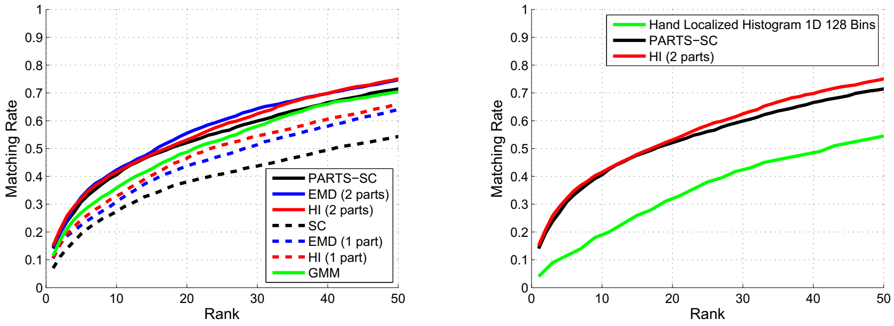


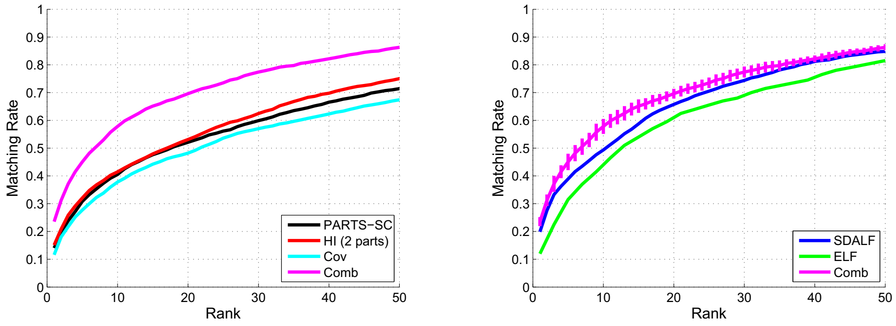
Figure 6.6: Examples of pedestrian image pairs from the VIPeR dataset. Note the strong variations in illumination and pose.



(a)

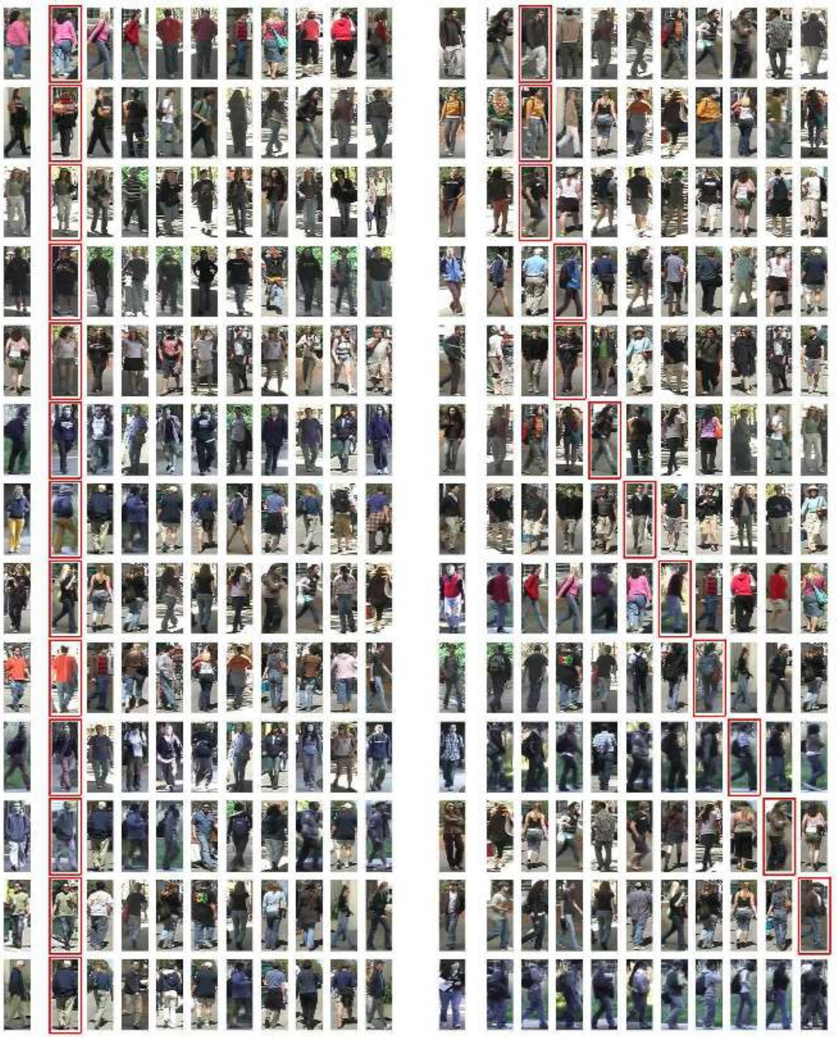
(b)

Figure 6.7: VIPeR dataset. (a) - Comparison of the three approaches for re-identification with one/two parts variation and the *GMM* approach. (b) - Comparison to the best baseline approach reported in [26].



(a) (b)

Figure 6.8: VIPeR dataset. (a) - Comparison of two parts variation of *PARTS-SC* and *HI* methods with *Cov* method and their combination, the *Comb* method. (b) - Comparison of the *Comb* method with the ELF and SDALF methods.



(a) (b)

Figure 6.9: VIPeR dataset. Examples of matches using the *Comb* signature. For each one of the figures (a) and (b), the leftmost column shows the probe image and to the right of it the ten top ranked matches out of the 316 gallery images are displayed. Highlighted images represent a correct match. The last row in column (b) is an example of a query which does not have any correct match in the ten top ranked.

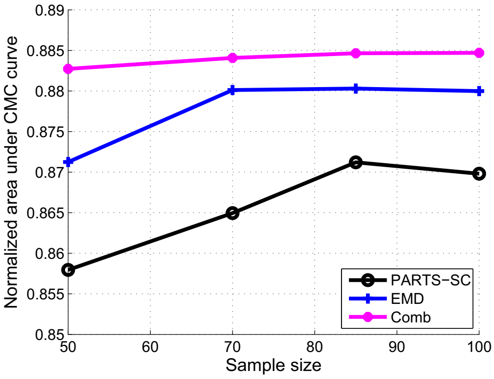


Figure 6.10: VIPeR dataset. Accuracy as a function of the sample size. Four different sample sizes were used - 50,70,85 and 100. The best results are obtained for sample size of 85.

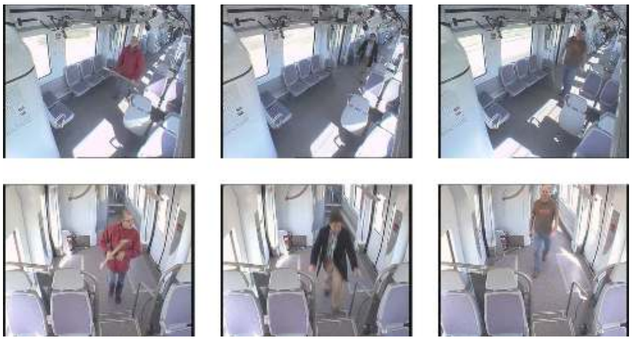
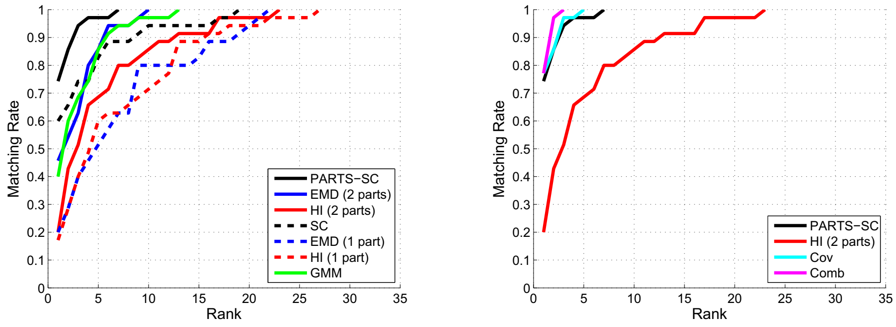


Figure 6.11: Examples of passenger images in the BOSS dataset. Each row shows images viewed by different camera. Note the difference in the viewing directions of the cameras and the different illumination conditions.



(a) (b)

Figure 6.12: BOSS dataset. (a) - Comparison of the three approaches for re-identification with one/two parts variation and the *GMM* approach. (b) - Comparison of two parts variation of *PARTS-SC* and *HI* methods with *Cov* method and their combination, the *Comb* method

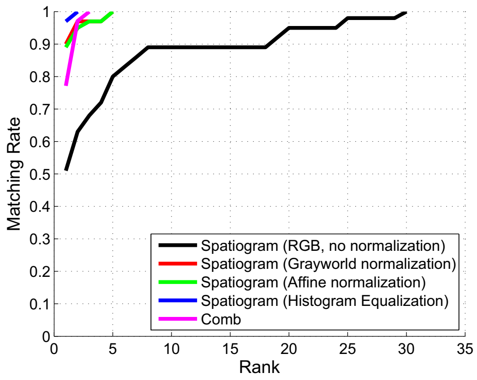


Figure 6.13: BOSS dataset. Performance evaluation of spatiogram using several illumination normalization methods reported by [53] compared with Comb signature.

Chapter 7

Discussion

7.1 Major Contributions

In this work we have evaluated and compared several methods for person re-identification using color. We conclude that color is a powerful cue for person re-identification, when used properly. Adding quite limited information regarding the color observation spatial source (upper/lower part) contributes significantly to the performance. Our main contribution is an invariant signature exploiting a structure of color distributions, using different parts of the object. We have proved the signature's invariant properties under the assumption of a diagonal model for illumination change and demonstrated its discriminative nature in various experiments. We have introduced *Comb*, a state-of-the-art signature for person re-identification, and evaluated its performance on publicly available datasets.

7.2 Future Work

We have shown that the PARTS-SC signature is perfectly invariant to illumination change under the assumption of a diagonal model for illumination change. In practice the diagonal model does not hold perfectly, but preprocessing methods such as sensor sharpening (see chapter 8.1) may improve the accuracy of the model for a specific camera. One possible way to improve the *PARTS-SC* signature invariance would be to allow camera calibration in form of deriving its sharpening transform and working with the transformed sensors.

Our *Comb* signature is actually built of three different signatures each capturing a different aspect of the target. The distance measure between two *Comb* signatures, Eq. 6.1, equally weights the distances between the component signatures. In our experiments we did not try to optimize the distance measure over the weighting parameters. Thus, their learning may lead to accuracy improvement.

Throughout this work we have treated the person re-identification problem as a ranking problem. We did not address the confidence measure returned by each one of the signatures. A possible extension of our work would be analyzing the distance measures returned by each signature and converting them into a measure of confidence. Given a query, the returned measure of confidence for each candidate could be used by a human operator as an indicator whether or not it is worthwhile searching the candidates list further or not.

Generally speaking, the problem of person re-identification is still far from being solved and offers much space for improvements either on the description side or the classification schemes side.

Chapter 8

Appendix

8.1 Diagonal Model Validation

The purpose of this section is to analyze the power of the diagonal model used for modeling the illumination changes. Finlayson *et al.* [15] have shown that under the assumption of narrow-band sensors, *i.e.* the spectral response function of imaging device sensors are close enough to the Dirac delta function, the illumination change can be modeled almost perfectly using the diagonal transformation. The authors have proved that when sensors spectral distribution is not narrow-band, it is still possible to make it more narrow-band by applying a linear transformation T on the sensor responses, a process they call *spectral sharpening*. The matrix T transforms the sensor responses to a new basis in which the diagonal model holds more precisely. Several methods for computing T were suggested in [15]. The most suitable method for practical purposes is the *database sharpening* method. The idea behind this approach is to estimate T, using a database of reflectance observations obtained under various illuminations.

Let us assume that the sensor responses $\rho^o = (R^o, G^o, B^o)$ and $\rho^c = (R^c, G^c, B^c)$, corresponding to a surface imaged under two different illuminations o and c respectively, are related by a general linear transformation M, *i.e.* $\rho^c = \rho^o M$. Let A_i and A_j denote $n \times 3$ matrices of sensor responses to n surfaces under illuminants i and j respectively. Thus the optimal M_{ij} is obtained through a least-square solution for $A_i M_{ij} = A_j$. Given a database of N images, each depicting n colored surfaces, the average mapping error of the image pairs in the database is obtained by:

$$\frac{1}{N(N-1)} \sum_i \sum_{j \neq i} \|A_i M_{ij} - A_j\|_F \tag{8.1}$$

where $\|\cdot\|_F$ is a Frobenius matrix norm. This is the lowest possible mapping error, under the assumption of a linear transformation between the sensor responses. The linear transformation T_{ij} diagonalizing M_{ij} ($M_{ij} = T_{ij}D_{ij}T_{ij}^{-1}$), is in fact the sensors sharpening transformation [15] learned from a pair of illuminations i and j. Speaking practically, sharpening transformation depends on the camera sensors and as such has to be camera dependent and not to be tailored towards a specific pair of illuminations. Several methods for obtaining such a general sharpening transformation can be found in [3]. We will not delve into these procedures, but rather assume that such a transformation T is given to us. The sensor responses of each one of the N images in the database are transformed to a new basis, $A'_i = A_iT,\ i = 1..N$. For a pair of images i, j, a diagonal mapping from A'_i to A'_j is obtained by computing the optimal, in the least-square sense, D_{ij} from $A'_i = A'_jD_{ij}$. Finally, the average mapping error for the image pairs in the database is obtained by:

$$\frac{1}{N(N-1)}\sum_i\sum_{j\neq i}\left\|A_iTD_{ij}T^{-1} - A_j\right\|_F \tag{8.2}$$

Here, for each pair of images, instead of using the ad hoc linear transformation M_{ij}, a general sharpening transformation is used together with a more limited ad hoc diagonal transformation. Thus, at best the obtained mapping error will be the same as in Eq. 8.1, but usually it will be higher. We can, therefore, use Eq. 8.1 as a lower bound on Eq. 8.2. Assuming a simple diagonal model, without sharpening whatsoever, the average mapping error for the image pairs in the database is computed using Eq. 8.2 (replacing T by the 3×3 identity matrix):

$$\frac{1}{N(N-1)}\sum_i\sum_{j\neq i}\|A_iD_{ij} - A_j\|_F \tag{8.3}$$

In our experiment we use a database of images taken under different illuminations indoors and outdoors introduced by Gehler *et al.* [20]. A Macbeth colorcheckers board (color calibration board) is shot in each one of these images, figure 8.2 shows some examples. We use the top three rows of the calibration board tiles as a set of surfaces and randomly split them into two subsets. Using the first subset of tiles we obtain the diagonal mapping D_{ij} and the linear mapping M_{ij} for each image pair (i, j). We evaluate the average diagonal mapping error and the average linear mapping error using Eq. 8.3 and Eq. 8.1 respectively, and call them error measures. We would like to estimate how significant the average error is in comparison with the average distance between different surfaces. To this end, we use the second subset of tiles for evaluating the average distance between different surfaces mapped using the diagonal and linear transformations obtained from the first subset, and call them distance measures.

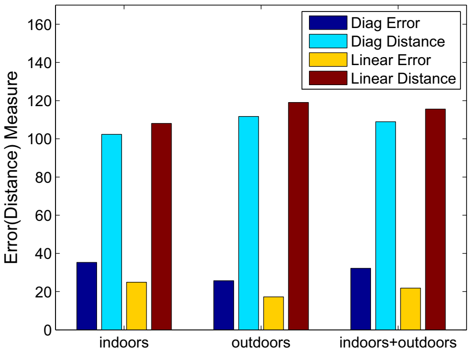


Figure 8.1: Error and distance measures of the diagonal and the linear mappings. We differentiate between three cases: 1. Both images are taken indoors. 2. Both images are taken outdoors. 3. No constraints on the scene type.

The lower the ratio between the error measure and the distance measure, the better discriminating abilities are. We differentiate between three cases in our experiments - using images taken only indoors, only outdoors or all of the images without any constraints. Figure 8.1 depicts error and distance measures for the three cases. As expected, the error measures corresponding to the diagonal mapping are higher than those corresponding to the linear mapping. The difference in the error measures is not that high, thus had we used sharpening transformation the error would not be reduced significantly. Similarly, the distance measures corresponding to the diagonal mapping are lower than those corresponding to the linear mapping. Looking at the results we can conclude that the discriminating power of the linear model is higher than that of the diagonal model, based on a smaller ratio between the error and the distance measures. But the difference is not very significant, suggesting that the diagonal model is indeed a very good compromise between simplicity and accuracy. Another conclusion is that the diagonal model is more suitable for modeling the illumination variations outdoors than indoors since the ratio between error and distance measures for the outdoor images is smaller than that for the indoor images.

As shown in section 3.2.3, *log* color space is invariant to changes in the illumination conditions assuming the diagonal model up to translation. Translation vector $(log\frac{\alpha}{\beta}, log\frac{\gamma}{\beta})$ depends only on the parameters of the diagonal model. We have extracted the diagonal model parameters (α, β, γ) for pairs of different illuminations, while we make a distinction between four cases:

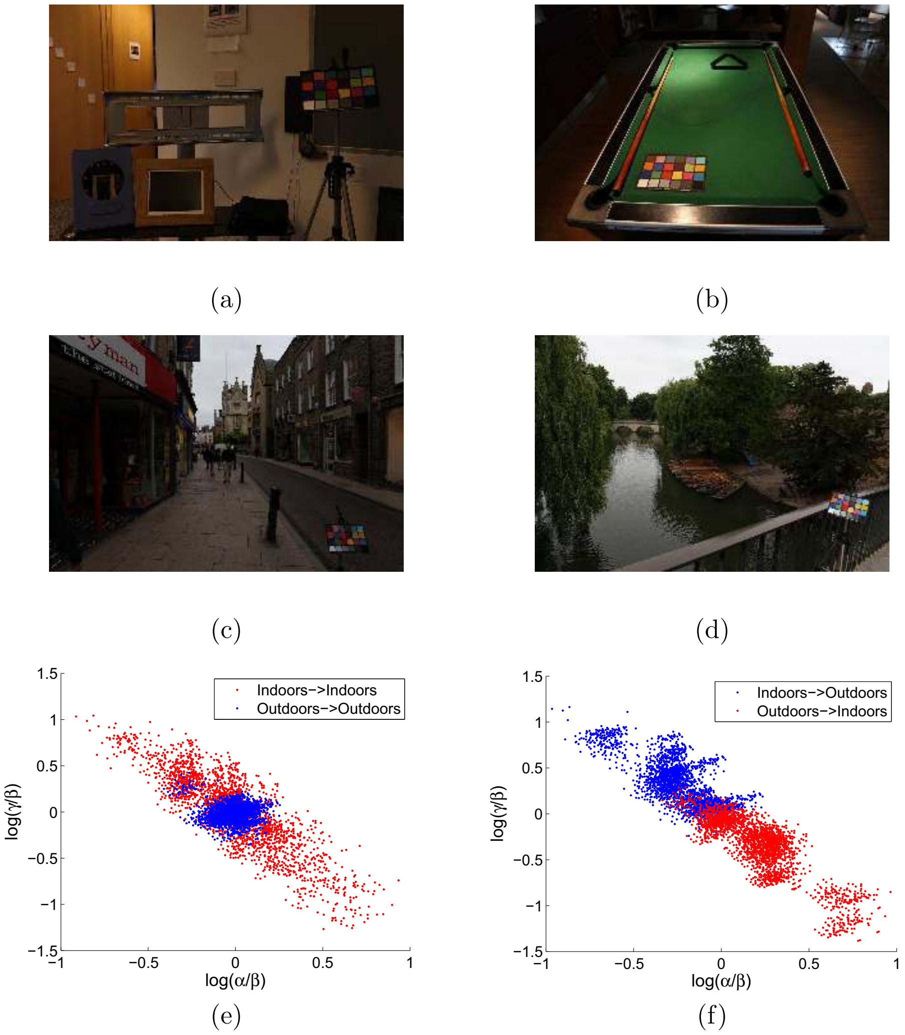


Figure 8.2: (a-d) Examples of images from database by Gehler *et al*. [20]. (e-f) Distribution of translation vectors due to illumination change. (a),(b) Images of indoors scenes. (c),(d) Images of outdoors scenes. (e) - Both scenes are imaged under either outdoor or indoor illumination. (f) - One scene is imaged under outdoor illumination and one is imaged under indoor illumination. Note the large differences in illumination conditions.

- Both illuminations are outdoors.
- Both illuminations are indoors.

- The first illumination is indoors and the second is outdoors.
- The first illumination is outdoors and the second is indoors.

Figure 8.2 depicts the distribution of translation vectors for these cases. Notice that in the first case the vectors are distributed more densely around the zero vector, while in the second case the distribution has higher variance, indicating on a higher variance in illumination indoors than outdoors. As expected, in the third and fourth case vector distributions reveal the tendency of colors to become colder (more blue) and warmer (more red) respectively.

Bibliography

[1] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person reidentification using spatial covariance regions of human body parts. In *AVSS10*, pages 435–440, 2010.

[2] K. Barnard. Practical color constancy. In *Ph.D. thesis*, 1999.

[3] K. Barnard, F. Ciurea, and B.V. Funt. Sensor sharpening for computational color constancy. 18(11):2728–2743, November 2001.

[4] H. Bay, T. Tuytelaars, and L.J. Van Gool. Surf: Speeded up robust features. In *Proc. Eurpoean Conf. on Computer Vision (ECCV)*, pages I: 404–417, 2006.

[5] S.J. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[6] D. Berwick and S.W. Lee. A chromaticity space for specularity-, illumination color- and illumination pose-invariant 3-d object recognition. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 165–170, 1998.

[7] S. Birchfield and S. Rangarajan. Spatiograms vs. histograms for region based tracking. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[8] M.C. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *Proc. Eurpoean Conf. on Computer Vision (ECCV)*, page II: 628, 1998.

[9] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages I: 886–893, 2005.

[11] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[12] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages II: 264–271, 2003.

[13] G.D. Finlayson, S.S. Chatterjee, and B.V. Funt. Color angular indexing. In *Proc. Eurpoean Conf. on Computer Vision (ECCV)*, pages II:16–27, 1996.

[14] G.D. Finlayson, M.S. Drew, and B.V. Funt. Diagonal transforms suffice for color constancy. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 164–171, 1993.

[15] G.D. Finlayson, M.S. Drew, and B.V. Funt. Spectral sharpening: Sensor transformations for improved color constancy. *Journal of the Optical Society of America A (JOSA-A)*, 11(5):1553–1563, May 1994.

[16] G.D. Finlayson, S. Hordley, G. Schaefer, and G.Y. Tian. Illuminant and device invariant colour using histogram equalisation. *The Journal of the Pattern Recognition Society (PR)*, 38(2):179–190, February 2005.

[17] G.D. Finlayson and S.D. Hordley. Color constancy at a pixel. *J. Opt. Soc. Am. A (JOSA-A)*, 18(2):253–264, February 2001.

[18] W. Forstner and Moonen B. A metric for covariance matrices. Technical report, Dept. of Geodesy and Geoinformation, Stuttgart University, 1999.

[19] B.V. Funt and G.D. Finlayson. Color constant color indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):522–529, 1995.

[20] P. V. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp. Bayesian color constancy revisited. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[21] T. Gevers. Robust segmentation and tracking of colored objects in video. *IEEE Trans. Circuits and Systems for Video Technology (CSVT)*, 14(6):776–781, 2004.

[22] T. Gevers and A.W.M. Smeulders. Color-based object recognition. *The Journal of the Pattern Recognition Society (PR)*, 32(3):453–464, 1999.

[23] T. Gevers and A.W.M. Smeulders. Pictoseek: Combining color and shape invariant features for image retrieval. *IEEE Trans. Image Processing*, 9(1):102–119, 2000.

[24] N. Gheissari, T.B. Sebastian, and R.I. Hartley. Person reidentification using spatiotemporal appearance. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages II: 1528–1535, 2006.

[25] K. Grauman and T.J. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages II: 1458–1465, 2005.

[26] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. 2007.

[27] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. Eurpoean Conf. on Computer Vision (ECCV)*, pages I: 262–275, 2008.

[28] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person reidentification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *International Conference on Distributed Smart Cameras (ICDSC)*, pages 1–6, 2008.

[29] S. Hordley. Scene illuminant estimation: Past, present and future. In *Color Research and Application 31(4)*, page 303–314, 2006.

[30] D.P. Huttenlocher and S. Ullman. Object recognition using alignment. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 102–111, 1987.

[31] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 26–33, 2005.

[32] K. Jeong and C. Jaynes. Object matching in disjoint cameras using a color transfer approach. *Machine Vision and Applications (MVA)*, 19(5-6):443–455, October 2008.

[33] M. Kliot and E. Rivlin. Invariant-based shape retrieval in pictorial databases. *Computer Vision and Image Understanding (CVIU)*, 71(2):182–197, August 1998.

[34] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, March 1951.

[35] Y. Lamdan and H.J. Wolfson. Geometric hashing: A general and efficient model-based recognition scheme. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 238–249, 1988.

[36] Zhe Lin and Larry S. Davis. Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In *International Symposium on Visual Computing (ISVC)*, pages 23–34, 2008.

[37] D.G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395, March 1987.

[38] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[39] C. Madden, E.D. Cheng, and M. Piccardi. Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Machine Vision and Applications (MVA)*, 18(3-4):233–247, August 2007.

[40] C. Madden, M. Piccardi, and S. Zuffi. Comparison of techniques for mitigating the effects of illumination variations on the appearance of human targets. In *International Symposium on Visual Computing (ISVC)*, pages II: 116–127, 2007.

[41] J. Matas. Colour-based object recognition. In *Ph.D. thesis*, 1996.

[42] J. Matas, D. Koubaroulis, and J.V. Kittler. Colour image retrieval and object recognition using the multimodal neighbourhood signature. In *Proc. Eurpoean Conf. on Computer Vision (ECCV)*, pages I: 48–64, 2000.

[43] M. J. Metternich, M. Worring, and A. W. M. Smeulders. Color based tracing in real-life surveillance data. *Transactions on Data Hiding and Multimedia Security*, (5):18–33, 2010.

[44] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, October 2005.

[45] H.J. Moon and P.J. Phillips. Computational and performance aspects of pca-based face recognition algorithms. *Perception*, 30(3):303–321, 2001.

[46] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. *The Journal of the Pattern Recognition Society (PR)*, 36(9):1997–2006, September 2003.

[47] U. Park, A.K. Jain, I. Kitahara, K. Kogure, and N. Hagita. Vise: Visual search engine using multiple networked cameras. In *Proc. Int. Conf. on Pattern Recognition (ICPR)*, pages III: 1204–1207, 2006.

[48] F. Porikli. Inter-camera color calibration by correlation model function. In *International Conference on Image Processing(ICIP)*, pages II: 133–136, 2003.

[49] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):91–121, 2000.

[50] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 1470–1477, 2003.

[51] H.M.G. Stokman and T. Gevers. Selection and fusion of color models for image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):371–381, 2007.

[52] M.J. Swain and D.H. Ballard. Indexing via color histograms. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 390–393, 1990.

[53] D.N. Truong Cong, C. Achard, L. Khoudour, and L. Douadi. Video sequences association for people re-identification across multiple non-overlapping cameras. In *Image Analysis and Processing (ICIAP)*, pages 179–189, 2009.

[54] O. Tuzel, F.M. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. Eurpoean Conf. on Computer Vision (ECCV)*, pages II: 589–600, 2006.

[55] S. Ullman. *High-Level Vision: Object Recognition and Visual Cognition*. MIT Press, 1996.

[56] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006, October 1991.

[57] P. Viola and M. Jones. Robust real time object detection. In *IEEE ICCV Workshop on Statistical and Computational Theories of Vision*, 2001.

[58] X.G. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 1–8, 2007.

[59] Y. Yu, D. Harwood, K. Yoon, and L.S. Davis. Human appearance modeling for matching across video sequences. *Machine Vision and Applications (MVA)*, 18(3-4):139–149, 2007.

[60] W.S. Zheng, S.G. Gong, and T. Xiang. Associating groups of people. In *The British Machine Vision Conference (BMVC)*, pages xx–yy, 2009.

[61] Q.A. Zhu, M.C. Yeh, K.T. Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages II: 1491–1498, 2006.

האדם ולכן השילוב של שלושת החתימות נותן תוצאות טובות יותר מכל אחת מן החתימות שמרכיבות אותו בנפרד.

בחנו את ביצועי החתימות המוצעות על מספר מאגרי תמונות. שני מאגרים, VIPeR [26] ו־ BOSS [53], הם מאגרי השוואת ביצועים של אלגוריתמי ה־ person re-identification הידועים בקהילה ועליהם השוונו את השיטה שלנו מול שיטות אחרות. בנוסף לכך אספנו מאגר תמונות משלנו הכולל 31 בני אדם שצולמו בארבע מצלמות אבטחה שונות הממוקמות בשטחים פתוחים וסגורים. על מאגר ה־ VIPeR, שנחשב למאגר המאתגר מכולם, השיטה שלנו משיגה ביצועים טובים יותר משיטות אחרות שהוצעו לאחרונה [27, 11]. על מאגר ה־ BOSS אנו משיגים ביצועים הדומים לשיטות האחרות. בניסויים בדקנו והשוונו את הביצועים של כל החתימות לחוד וגם את החתימה המשולבת כפי שתארנו ומצאנו שהחתימה המשולבת משיגה את הביצועים הטובים ביותר.

לסיכום, תרומותנו העיקריות הן:

1. הצגת חתימת צבעים חדשה המבוססת על חלקים והוכחת התכונות האינווריאנטיות שלה תחת הנחת המודל האלכסוני.
2. שילוב החתימה החדשה עם חתימות נוספות להשגת ביצועים שהם הטובים ביותר כיום בתחום ה־ person re-identification.
3. הראנו שחתימות המבוססות על צבע ומשתמשות במידע מרחבי יחסית מוגבל משיגות ביצועים מרשימים כאשר עושים שימוש נכון במרחבי הצבע האינווריאנטיים.

צירים הקרטזית הדו־מימדית. בחנו מספר מרחבי צבע והגענו למסקנה שהמרחב המתאים ביותר לתיאור אינווריאנטי של התפלגויות הצבעים באובייקט הוא ה־ *log-chromaticity*. ב־ [6] הוכחה האינווריאנטיות של המרחב לשינויי התאורה, תחת הנחת המודל האלכסוני, עד כדי הזזה קבועה במרחב הצבע הדו מימדי שתלויה אך ורק בפרמטרי המודל.

בעבודה זו אנו מציעים שיטה לתיאור המראה של האדם. השיטה מסתמכת על צורת התפלגות הצבעים של בגדיו של האדם ועמידה לשינויי התאורה במעבר בין המצלמות תחת הנחת המודל האלכסוני. בלב השיטה עומדת ההבחנה שצורת התפלגות הצבעים היחסית של חלקי הלבוש העליון והתחתון של המטרה לא משתנה כתוצאה משינויי התאורה. אנו מתארים את צורת ההתפלגות באמצעות Shape Context [5] , ייצוג לא פרמטרי שבדרך כלל שימש לתיאור קונטור של צורות. SC מתאר את התפלגות הנקודות שנדגמו על קונטור של צורה, יחסית לנקודת יחוס כלשהי, באמצעות היסטוגרמה פולארית כך שהמרחקים הרדיאליים בין מחלקות ההיסטוגרמה מפולגים אחיד במרחב הלוגריתמי. השיטה אותה אנו מציגים מבוססת על SC אבל עם שינוי קטן. למען השמירה על העקביות של תאור המטרה היינו רוצים לשמר את מקור הדגימות, כלומר האם נדגמו מחלק לבוש העליון או התחתון. לשם כך אנו מתארים את התפלגות הדגימות מחלק העליון יחסית לנקודות יחוס שמקורן בחלק התחתון. אנו קוראים למזהה החדש PARTS-SC היות והוא מתחשב בחלקים של האובייקט המתואר. חתימת המטרה היא בעצם אוסף של מזהי ה־ PARTS-SC. כך, למעשה, מתוארות צורות ההתפלגות של שני החלקים והיחס ביניהם. חשוב להדגיש שבניגוד ל־ SC המקורי שמשתמש בקואורדינטות מרחביות, ה־ PARTS-SC מתאר התפלגות נקודות במרחב הצבע, בפרט, במרחב ה־ *log-chromaticity* שתיארנו. בהינתן שתי מטרות המיוצגת באמצעות N מזהי PARTS-SC כל אחת, נגדיר את המרחק ביניהן להיות מחיר השידוך המושלם בעל המחיר המינימאלי בין מזהי המטרות.

בניסויים שביצענו ראינו שלחתימת ה־ PARTS-SC יכולת הבדלה גבוהה, בנוסף לתכונותיה האינווריאנטיות. אך היות והיא מתארת את היחס בין צבעי חלקי המטרה, יתכן ולמטרות בעלות מבנה צבעים שונה לגמרי יהיו חתימות דומות אחת לשניה. אי לכך הוספת המידע הודות הצבעים המוחלטים של המטרה צפויה להביא לשיפור ניכר. בדקנו מספר שיטות של תאור הצבעים האבסולוטיים ביניהן היסטוגרמה, EMD, ו־ GMM והגענו למסקנה שההיסטוגרמה במרחב צבע ה־ *log-chromaticity* נותנת את הביצועים הטובים ביותר הן מבחינת הדיוק והן מבחינת היעילות החישובית. השתמשנו בהיסטוגרמה לתאור שני חלקי הלבוש של האדם, החלק העליון והתחתון, והראנו שהשימוש בחלקים משפר את הדיוק בהשוואה להיסטוגרמה הכוללנית.

מאפיין חשוב נוסף של פריטי הלבוש היא טקסטורה, היות והיא מתארת תבניות כמו פסים ומשבצות. חתימת ה־ PARTS-SC וגם ההיסטוגרמה משתמשות במידע מרחבי מוגבל (בינארי) ולכן אינן מסוגלות לקודד טקסטורה. השתמשנו ב־ covariance descriptor [54] לקידוד הטקסטורה ושינויים עדינים של צבעים בבגדים. מרחב המאפיינים שנבחר הוא ה־ RGB המקורי והקואורדינטה המרחבית האנכית של הצללית. בניסויים שערכנו ראינו שהוספת הקואורדינטה המרחבית האופקית פוגעת בביצועים, כנראה בגלל הסימטריה האופקית של צורת הלבוש האופיינית. כל אחת משלושת החתימות שהצגנו, PARTS-SC , היסטוגרמה ו־ covariance descriptor מתארת אספקט שונה של מראה בגדיו של

ב־ [24] , שהיא אחת העבודות הראשונות בתחום נעשה שימוש בכמה מהטכניקות שתוארו. לאחר ביצוע הסגמנטציה לכל פריים בנפרד, מתבצעת סגמנטציה של פריטי הלבוש בעזרת הקשר הסיבתי בין הפריימים. המטרה מתוארת כאוסף תיאורים מקומיים המבוססים על צבע וטקסטורה מסביב לנקודות עניין יציבות על גבי המטרה. המחברים הראו שעצם חלוקת המטרה לחלקים ותיאור כל חלק בנפרד משפר את הביצועים בצורה ניכרת. דוגמא לעבודה שמשתמשת בגישה שונה לגמרי ניתן למצוא ב־ [27]. בניגוד ל־ [24] שם המאפיינים האינווריאנטיים הוגדרו מראש, בעבודה זו משתמשים בתהליך ממוכן (feature selection) לבחירת המאפיינים החזקים מתוך מאגר. המאגר כולל מאפיינים מבוססי צבע וטקסטורה מסוגים שונים ובעלי פרמטרים שונים. תוצאה מעניינת שהתקבלו מעבודה זו היא שיותר מ־75 אחוז מהמאפיינים שנבחרו בתהליך הממוכן היו מבוססי צבע. אבחנה זו מדגישה את החשיבות הרבה של הצבע בתהליך הזיהוי.

ההיסטוגרמה היא המזהה הנפוץ ביותר לתיאור מטרות המבוסס על צבע [52]. יתרונותיה בכך שהיא פשוטה למימוש, יעילה חישובית ובעלת יכולת הבחנה, ואילו חסרונותיה הם הרגישות לשינויי תאורה והיעדר מידע מרחבי על הפיקסלים. כדי להתגבר על השינויים בתאורה נהוג להשתמש באלגוריתמי פיצוי לצבע (color constancy) [29] או לחילופין למפות את ערכי הפיקסלים למרחב אינווריאנטי להשפעות תאורה [22]. כמו כן הוצעו מספר שיטות להוספת המידע מרחבי להיסטוגרמה, לדוגמה [59, 7].

תמונה של אובייקט צבעוני מיוצגת ע״י ערכי החיישנים של המצלמה. ברוב המצלמות קיימים שלוש סוגי חיישנים המוכרים לנו כ־ R, G ו־ B (אדום, ירוק וכחול), הרגישים לאורכי גל גבוהים, בינוניים ונמוכים בהתאמה. על פי המודל הלמברטי, תגובת החיישן לאור המגיע אליו תלויה בשלושה גורמים עיקריים ־ רגישות החיישן, התכונות הספקטרליות של מקור האור והתכונות הספקטרליות של החזרת האור מהחומר שמרכיב את המשטח המצולם. חילוץ מאפייני המשטח מתוך המשוואות של המודל הלמברטי איננו אפשרי בגלל שמספר האילוצים על המשוואות קטן ממספר הנעלמים. ולמרות זאת, ברוב האפליקציות, היכולת לשוות תמונות אובייקטים בתנאי תאורה שונים היא יותר חשובה מלמצוא את מאפייני המשטח במפורש. אי לכך, לרוב מתארים את השינויים שעוברים צבעי תמונת האובייקט תחת תנאי התאורה השונים באמצעות טרנספורמציה לינארית על ערכי החיישנים. לרוב מסתפקים בטרנספורמציה מסוג מסוים ־ הטרנספורמציה האלכסונית ־ לפיה ערכו של כל חיישן מוכפל בקבוע ואינו תלוי בערכי החיישנים האחרים. המודל האלכסוני נחקר במשך שנים ונמצא מתאים לתאר שינויים בתאורה טבעיית ומלאכותיית [14]. עובדה זאת וגם פשטות המודל הובילו לכך שהרבה עבודות מניחות אותו כהנחת עבודה [42], וגם אנחנו אימצנו מודל זה בעבודה זו.

מרחב הצבע הסטנדרטי, RGB, אינו אינווריאנטי תחת הנחת המודל האלכסוני. לעומת זאת ישנם מרחבי צבע לגביהם ניתן להוכיח אינווריאנטיות מלאה או חלקית. לדוגמה, מרחב הצבע המנורמל *rgb* אינווריאנטי לעוצמת התאורה אך אינו אינווריאנטי לצבע התאורה. המעבר מ־ RGB למרחב צבע אחר היא המרה מערכי הפיקסל המקוריים לחדשים תוך שימוש בטרנספורמציה אלגברית. כל פיקסל עובר את אותה הטרנספורמציה באופן בלתי תלוי בפיקסלים האחרים. לדוגמה, מרחב הצבע ה־ *log-chromaticity* [6] הוא דו־מימדי, כלומר כל פיקסל המיוצג בקוואורדינטות תלת ממדיות במרחב ה־ RGB ממופה לנקודה במערכת

תקציר

בשנים האחרונות חלה עליה משמעותית בהיקפי פריסת מצלמות אבטחה. בדרך כלל רשת מצלמות אבטחה במתחם כלשהו כוללת מספר סוגים של מצלמות הפרוסות באיזורים פתוחים וסגורים. לעתים קרובות שדות הראיה (field of view) של המצלמות אינם נחתכים. סרטי אבטחה המצולמים עוברים עיבוד באמצעות אלגוריתמים לגילוי צלליות של אנשים כמו אלגוריתמי חיסור מרקע (background subtraction) ואלגוריתמי גילוי הולכי רגל (pedestrian detector). צלליות אלה מאוחסנות במסד נתונים. בעיית הזיהוי החוזר של אנשים (person re-identification) היא לזהות אדם על סמך תמונתו שצולמה באחת המצלמות ברשת בהינתן תמונה שצולמה במצלמה אחרת ברשת. הבעיה נחשבת למאתגרת במיוחד מפני שהתמונות שצולמו ממצלמות שונות ומתארות את אותו האדם נראות מאוד שונה אחת מהשנייה. הסיבות לכך הן שינויים בתנאיי התאורה, המיקום ומאפייני המצלמות והתנוחה של האדם ביחס למצלמה. כמו כן לסרטי אבטחה רזולוציה נמוכה, מה שמונע מלהשתמש בטכניקות המבוססות על מאפיינים ביומטריים כמו זיהוי פנים. נוסף על כך, ייתכן שהמידע שעל פיו יש לבצע את הזיהוי הוא מאוד מצומצם ־ מספר תמונות בודד או וידאו קצר.

על מנת לפתור את בעיית הזיהוי יש לתאר את מראה האדם בצורה שמצד אחד תשמר את זהותו למרות השינויים בתנאי הצילום ומצד שני תאפשר להבדיל בין מספר רב של אנשים. ישנן מספר גישות לפתרון הבעיה הנבדלות במספר היבטים. חלק מהשיטות מנצלות את הקשר בציר הזמן בין הפריימים וחלקן מתייחסות אליהם כאל אוסף של תמונות. שיטות שונות משתמשות במאפיינים שונים לתיאור המראה של האדם. רוב המאפיינים מסתמכים על הצבע והטקסטורה של הבגדים ולעתים גם משתמשים בצורת הצללית. יש עבודות בהן מרחב המאפיינים מוגדר מראש, ואילו באחרות משתמשים בבחירת מאפיינים (feature selection) לבחירת המאפיינים הטובים ביותר מתוך אוסף גדול שהוגדר מראש. ייצוג מראה האדם יכול להיות כוללני או מבוסס על חלקים, לדוגמא, תיאור נפרד לפלג גוף העליון ולרגליים. לעתים גם נעשה שימוש בגישת ה־ bag of features לתיאור המראה. לפי גישה זו, בתהליך מקדים נלמד מילון של מאפיינים מורכבים המתארים את המראה של האנשים וכל אדם ספציפי מתואר כהתפלגות מעל מילות המילון שנלמד. בחלק מהעבודות משתמשים במציאת נקודות עניין יציבות (SIFT, SURF) ומתארים את מראה האדם כאוסף של תיאורים מקומיים. העבודות גם נבדלות באופן בו הן מבצעות את הזיהוי עצמו. לרוב משתמשים בשילוב של מטריקת מרחק בין תיאורי המראה ומסווג פשוט, כדוגמת השכן הקרוב, אך לפעמים משתמשים בסכמות למידה מורכבות יותר הדורשות תהליך לימוד מקדים כדוגמת SVM.

המחקר נעשה בהנחיית דר' עמית אדם, פרופ' אהוד ריבלין ופרופ' מיכאל היימן בפקולטה למדעי המחשב.

ברצוני להודות לדר' עמית אדם, פרופ' אהוד ריבלין ופרופ' מיכאל היימן על הנחייתם ועזרתם המסורה. ברצוני גם להודות למשפחתי וחברי על תמיכתם.

אני מודה לטכניון על התמיכה הכספית הנדיבה בהשתלמותי.

אינווריאנטים בצבע לזיהוי חוזר של אנשים

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים במדעי המחשב

איגור קביאטקובסקי

הוגש לסנט הטכניון – מכון טכנולוגי לישראל
שבט תשע"ב חיפה פברואר 2012

אינווריאנטים בצבע לזיהוי חוזר של אנשים

איגור קביאטקובסקי