A LEARNING-BASED METHOD FOR PERSON RE-IDENTIFICATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS INSTITUTE
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BURÇİN BUKET OĞUL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE
IN THE DEPARTMENT OF INFORMATION SYSTEMS

APRIL 2013

A LEARNING-BASED METRIC FOR PERSON RE-IDENTIFICATION

Submitted by **Burçin Buket Oğul** in partial fulfillment of the requirements for the degree of **Master of Science in Information Systems, Middle East Technical University by**,

Prof. Dr. Nazife Baykal _____
Director, Informatics Institute

Prof. Dr. Yasemin Yardımcı Çetin _____
Head of Department, Information Systems

Assist.Prof. Dr. Alptekin Temizel _____
Supervisor, Work Based Learning, METU

Examining Committee Members

Prof.Dr. Yasemin Yardımcı Çetin _____
IS, METU

Assist.Prof. Dr. Alptekin Temizel _____
WBL, METU

Assist. Prof. Dr. Erhan Eren _____
IS, METU

Assist. Prof. Dr. Banu Günel _____
IS, METU

Assist. Prof. Dr. Sinan Kalkan _____
CENG, METU

Date: 16.04.2013

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Burçin Buket Oğul

Signature :

ABSTRACT

A LEARNING-BASED METRIC FOR PERSON RE-IDENTIFICATION

Oğul, Burçin Buket

MSc, Department of Information Systems

Supervisor: Assist.Prof. Dr. Alptekin Temizel

April 2013, 58 pages

Matching pedestrian images captured from different cameras is called person re-identification problem. The problem is challenging due to the low resolution of images, differences in illumination, the positional variance and possible appearance of carried objects, such as a bag, at different viewpoints. In this thesis, we investigate the discriminative ability of different features extracted from image in a binary classification framework. We finally propose a learning based method to combine different feature sets, Hue, Saturation, Value (HSV) histogram, Maximally Stable Color Regions (MSCR) and Speeded up Robust Features (SURF) matches, in a single framework. The experiments on widely used benchmark sets have shown that the best accuracy is obtained with weighted and localized histogram features. We also argue that further division of pedestrian body along the horizontal axis has the potential to increase the reidentification performance. Final integrative framework that we built outperforms the existing state-of-the-art models in terms of prediction accuracy.

Keywords: Person re-identification, learning based method, HSV histogram, Maximally Stable Color Regions (MSCR), Speeded up Robust Features (SURF)

ÖZ

KİŞİLERİN YENİDEN SAPTANMASI İÇİN ÖĞRENME TABANLI BİR YÖNTEM

Oğul, Burçin Buket

Yüksek Lisans, Bilişim Sistemleri Bölümü

Tez Yöneticisi: Y.Doç. Dr. Alptekin Temizel

Nisan 2013, 58 sayfa

Farklı kameralardan elde edilmiş yaya görüntülerinin eşleştirilmesi, kişilerin yeniden saptanması problemidir. Düşük çözünürlüklü görüntüler, aydınlatmadaki değişiklikler, konumsal değişimler ve çanta gibi taşınan bazı objelerin değişik açılardan görünür olup olmaması bu problemi zorlaştırmaktadır. Bu tezde görüntüden çıkarılmış değişik özniteliklerin ayırt edebilirlik yeteneği, bir ikili sınıflandırma altyapısında incelenmiştir. Sonuçta, değişik öznitelik kümelerini (HSV histogramı, Maximally Stable Color Regions (MSCR) ve Speeded up Robust Features (SURF)), tek bir çatı üzerinde birleştirebilen öğrenme tabanlı bir yöntem önerilmiştir. Bazı kıyaslama kümeleri üzerinde yapılan deneyler göstermiştir ki, en iyi doğruluk değerleri, ağırlıklandırılmış ve yerelleştirilmiş histogram özniteliklerinden elde edilmiştir. Yayaların vücut görüntülerinin yatay eksende daha da bölünmesinin, kişilerin yeniden saptanmasındaki performansı arttırdığını savunuyoruz. Gerçekleştirdiğimiz nihai entegre altyapı, doğruluk anlamında en gelişkin modellerden daha iyi sonuçlar üretmiştir.

Anahtar Kelimeler: Kişilerin Yeniden Saptanması, öğrenme tabanlı yöntem, HSV histogramı, Maximally Stable Color Regions (MSCR), Speeded up Robust Features (SURF)

To my family, my husband and my son to be born…

ACKNOWLEDGEMENTS

I express sincere appreciation to Assist.Prof. Dr. Alptekin TEMİZEL for his guidance and insight through the development of this thesis. I would also like to thank my beloved family, Mafiret, Hikmet and Pelin URAL, for their support, motivation and belief in me. Lastly, grateful thanks to my husband, Hasan OĞUL, for his understanding, endless love, valuable ideas and advices throughout my graduate studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MSER: Maximally Stable Extremal Regions

MSCR: Maximally Stable Color Regions

RHSP: Recurrent High-Structured Patches

HSV: Hue Saturation Value

SURF: Speeded Up Robust Features

SIFT: Scale-Invariant Feature Transform

ELF: Ensemble of Localized Features

CMC: Cumulative Match Characteristic

VIPeR: Viewpoint Invariant Pedestrian Recognition

FLD: Fisher Linear Discriminant

SVM: Support Vector Machine

AUC: Area Under Curve

RHSP: Recurrent High-Structured Patches

CHAPTER 1

INTRODUCTION

1.1 Motivation

We are witnessing a ubiquitous use of surveillance cameras both in outdoor and indoor environments for different purposes such as security, traffic monitoring and employee management. One of the main problems in these video surveillance systems is the matching of a moving person over multiple cameras. If a pedestrian in a camera view is seen in another one, matching their equality could simplify the rest of the recognition process. This problem is called as the person re-identification problem. Although the problem has received a great attention of researchers in the field of computer vision, the desired level of recognition accuracy could not be achieved so far. Several obstacles exist to hinder a significant improvement in the prediction performance. First, it is difficult to observe spatial continuity between two disjoint camera views especially when the camera views do not overlap. Second, the illumination conditions may differ in different views and different time points, which make it hard to calibrate. Third problem is the potential variation in poses and occlusions across time and camera. For example, while a backpack belonging to a person is visible when the image of the person is captured from the side, it may not be visible in another camera view captured from the front. The similarities between dressing habits make also the problem harder. Many people dress jeans, or white and black are quite common choices in upward clothes. Finally, the current methods suffer from the low quality of input images. Since the cameras are usually located to monitor a wide area images of individual persons have low resolution. Therefore, viewpoint and scale invariant models are needed to solve the problem.

1.2 Contributions

Tough great effort has been spent, the problem of person re-identification is still far from being effectively solved and faces several challenges for further improvement. In this study, we attempt to overcome the limitations in the literature by a discriminative framework that combines several feature sets over a learning-based metric. We also propose to divide the body into a number of horizontal segments and compute distinct histograms for each of these segments. In this respect, the body is first extracted from the background and automatically divided into two sub-parts: torso and leg. Then, each sub-part is divided into further horizontal segments with equal vertical lengths. These features are then combined with other features like SURF and stable segment content over the proposed supervised

distance metric. The experiments have shown that this approach can produce more accurate results in comparison with other feature sets and existing models.

The contribution of this thesis is three-fold:

- First, we propose to combine three major image feature sets in a single model; color features by histograms, texture features by MSCR, interest point features by SURF matches. We have shown that SURF matching can be useful when it is used in association with other features, while it is not so successful when used alone.

- Secondly, we argue that a better identification performance can be achieved when the weights of these features are guided by a supervised learning metric for final decision instead of using fixed participant weights for each.

- Finally, we report that dividing semantically segmented body parts (torso and leg) into further horizontal segments can improve the prediction accuracy. We have demonstrated that our final model can improve upon the state-of-the-art on a common challenging dataset.

1.3 Organization of the Thesis

The remainder of this thesis is as follows:

In Chapter 2, related works are presented. In Chapter 3, the proposed methods are discussed. First, the general concepts of learning based metrics are described. Then feature sets that are used in this thesis are explained in detail. Finally, preprocessing methods and localized histograms on horizontal segments are presented. In Chapter 4, firstly, the dataset and validation methods used in thesis are described and then, the obtained results are shown. The summary of the thesis, the conclusion and the future work are given In Chapter 5.

CHAPTER 2

RELATED WORK

The task of determining an object which appears in the field of view of one camera and recognizing the same object again in the same or another camera is called as “object re-identification” (Hamdoun, 2010). The commonly used technique in object re-identification is object histogram matching used in Gandi and Trivedi (2007), Pham et al. (2007), Orazio et al. (2009). For object re-identification, using object texture characteristics (Lantagne et al., 2003) and interest points are applied in some other works (Arth et al., 2007). Javed et al. (2008) combines object motion parameters with object appearance models. Two most common forms of object re-identification problem are vehicle and person re-identification. The former one is simpler (Gandi and Trivedi, 2007) because of the rigidity of the vehicles, paths they move on and the uniform color they have. In vehicle re-identification the most common features used are: size, velocity, lane position (Huang and Russel, 1998), color information (Kogut and Trivedi, 2007) and time of observation (Trivedi et al., 2005).

Various models have been introduced for person re-identification in the literature to address the challenges described in previous chapter. They usually differ in the feature sets used to represent images and the strategy used to make the final decision of pedestrian matching. A summary of previous studies can be found in Table 1.

In one of the earliest studies, Gheissari et al. (2006) defined an invariant signature based on a combination of normalized color and salient edgel histograms. They introduced a novel spatiotemporal segmentation algorithm to generate salient edgels that are robust to changes in appearance of clothing. The color information was captured by histograms based on hue and saturation. Final re-identification was achieved by evaluating pairwise histogram distances.

Hamdoun et al. (2008) proposed another signature based on interest point descriptors obtained from a set of consequent images. They built a multi-view invariant model of each pedestrian by accumulating time-series interest points using an efficient variant of SURF [11]. To match pedestrians, they implemented the Best Bin First search in a KD-tree containing all models. The major limitation of this work is the fact that it requires multiple-shot of each person across more than two cameras to effectively exploit time-series signatures. Another problem is that, as we also validated in our experiments, a high quality image set is required to detect valuable interest points, which is rarely the case in surveillance camera records.

While several color and texture features had been exploited in image distance calculations for person matching, Gray and Tao (2008) introduced an automated selection scheme to

identify most representative feature sets. They released an accompanying dataset, called ViPeR, with their method Gray et al. (2007). In this challenging dataset, they showed that a viewpoint invariant feature set can be selected by an effective learning approach based on AdaBoost. They prove that most valuable feature for person recognition is localized Hue histogram. Their method is usually referred as ELF (Ensemble of Local Features) in the literature.

Prosser et al. (2010) conducted a similar research, but they used Support Vector Machines instead of AdaBoost and retrieved top ranked results by redefining the person reidentification task as a ranking problem. They reported similar results to Gray and Tao (2008); emphasizing the importance of localized color features.

Oliveira and Luiz (2009) developed a model based on the matching of interest points collected in a query image with those collected in each video sequence used for each previously seen person. They used hue information as feature descriptor. The interest points were detected in two steps: finding the image Hessian matrix and searching the points that are significant, i.e. maximum and minimum values in Hessian. For comparison, a Haar-wavelet is used as an invariant signature, calculated for a set of pixels in a circle centered at an interest point.

Farenzena et al. (2010) reported the best results obtained to date on VIPeR dataset, currently the most difficult single-shot data set available. With a rigorous preprocessing of images, they extracted the whole body from the background and automatically divided it into three semantic parts: head, torso and leg. On vertically partitioned images, they accumulated horizontally symmetry-driven features to overcome the variance due to a different viewpoint. They also introduced a novel feature representation scheme based on the presence of Recurrent Highly Structured Patches. Overall model is called as SDALF (Symmetry Driven Accumulation of Local Features).

A comparison of interest-point-based features for person reidentification was presented by Bauml et al. (2011). Similarly, a comparison of color histogram features can be found in Gray et al. (2007). According to experimental results, color features outperform the others in terms of cumulative matching characteristic (CMC) curve, a common experimental measure of person reidentification accuracy.

Bak et al. (2010a) and Bak et al. (2010b) are two other examples of descriptive techniques which employ a set of local features extracted from image and then use a distance measure to match people. Former approach combined Haar-like texture features with dominant color descriptor. The latter one evaluated the performance of spatial covariance regions by segmenting the image into body parts.

Cai and Pietikainen (2010) proposed a novel approach, inspired from topic models in document matching, which represented each image by the counts of a set of colorwords previously constructed using color features. The colorwords were created by applying a simple k-means algorithm on 3x3 windows in a set of training images. They reported that hue and opponent histograms could achieve fairly well accuracy in person reidentification.

Hirzer et al. (2011) proposed a hybrid model that combines the descriptive and discriminative approaches. The image pairs were first evaluated by a region covariance descriptor and then fed into a discriminative model if a reasonably high rank was obtained.

The result of discriminative step was considered as a validation of first prediction, which in turn inhibited the number of false positives. They have shown that the hybrid solution could achieve better accuracy than single models.

Brun et al. (2011) represented an image by a graph whose nodes refer to segmented regions in the body based on Statistical Region Matching (SRM). Each node comprises average RGB values and the number of pixels for the corresponding region. Similarity of two graphs then refers to the similarity of two persons.

Table 1. Previous models for person re-identification

Reference	**Features**		**Decision**	
	Color	**Texture**	**Feature Selection and Use**	**Association**
Gheissari et al. (2006)	HS	Salient edgels	Histograms	Euclid distance
Hamdoun et al. (2008)	None	SURF-like	Interest points	Sum of absolute differences by Best Bin First search
Gray and Tao (2008)	RGB, HSV, YCbCr	Gabor, Schmid filters	Learning based (AdaBoost)	AdaBoost result
Oliveira and Luiz (2009)	Hue	SURF-like	Interest points	Sum of quadratic distances
Prosser et al. (2010)	RGB, HSV, YCbCr	Gabor, Schmid filters	Learning based (SVM)	SVM-rank
Farenzena et al. (2010)	HSV, stable color regions	RHSP	Symmetry-driven	Euclid distance
Bak et al. (2010a)	Dominant color descriptor	Haar-like	Binned descriptors	An ad-hoc similarity measure
Bak et al. (2010b)	Oriented gradient	Spatial covariance	Binned descriptors	Pyramid matching
Cai and Pietikainen (2010)	Hue, Opponent	None	Codeword composition	Chi-square distance
Bauml et al. (2011)	None	SIFT-like	Interest points	Point matches
Hirzer et al. (2011)	None	Haar-like, region covariance	Learning based (Boosting)	Boosting rank
Zheng et al. (2011)	RGB, HSV, YCbCr	Gabor, Schmid filters	Learning-based (Probabilistic)	Probabilistic Relative Distance Comparison
Brun et al. (2011)	RGB of segmented regions	Area of segmented regions	Graph of segmented regions	Graph comparison

CHAPTER 3

METHODS

The overview of the proposed person reidentification system is illustrated in Figure 1. The study consists of three main steps: preprocessing, feature extraction and classification. First of all, in the preprocessing step, HSV color space is used with two operators, chromatic bilateral operator and spatial covering operator, to extract background information and body parts of each pedestrian. Feature extraction step is a combination of 3 feature sets, Weighted Histogram, Maximally Stable Colour Regions (MSCR) and Speeded up robust features (SURF) detector. In the final step, classification, we use FLDA to solve the binary classification problem. Further sections give the details of preprocessing, feature extraction and classification steps.

3.1 Preprocessing

Before the feature extraction process, a 3-step preprocessing phase is applied to eliminate unnecessary background information, to extract horizontal body parts and to find vertical symmetry axis for feature extraction. For background extraction structuring element component analysis, which is known to be a successful method for foreground/background separation, is customized for pedestrian images as suggested by Farenzena et al. (2010). The other parts of preprocessing will be described in the following subsections in detail.

3.1.1 Body Part Detector (Horizontal Segments)

In this thesis, the body part extraction and symmetric partitioning steps are performed as suggested by Farenzena et al. (2010). The body is first automatically divided into three meaningful parts: head, torso and leg. This extraction process starts with separating torso and leg, the two largest body parts characterized by different color distributions (e.g. the regions comprising t-shirt/pants or suit/legs). The process goes on with searching asymmetrical axis between head and shoulders. The operators used for symmetry-based silhouette partitions are:

- *Chromatic bilateral operator:* For each horizontal axis, say i, in HSV image, the Euclidian distance of foreground pixel values at p_i and $\widehat{p_i}$ which locates symmetrically with respect to the i^{th} axis is calculated and summed over $[i - \delta, i + \delta]$. δ equals to $I/4$ to achieve scale independency, where I is the image height.

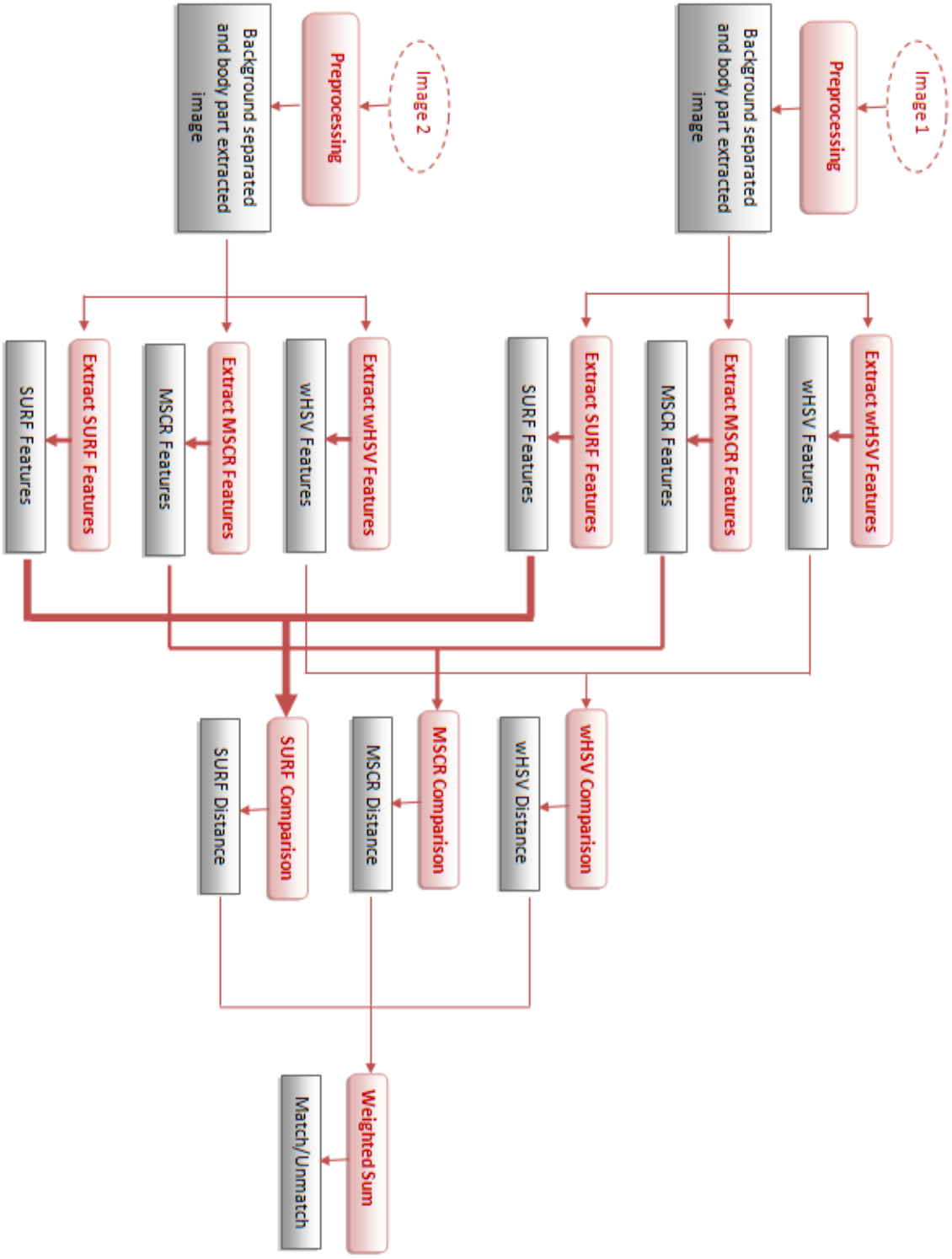


Figure 1. System Overview

Being a color operator, chromatic bilateral operator value is optimized with spatial covering operator at a point i_{TL}, where two regions are most distinguishable by colors such as skirt-T-shirt or jean-shirt. Calculation of chromatic bilateral operator is given in Equation 3.1.

$$C(i,\delta) = \sum_{B[i-\delta,i+\delta]} d^2\,(p_i, \widehat{p_i}), \qquad \text{(Equation 3.1)}$$

where $d(.,.)$ is Euclidian distance, B is the subregion with respect to the i^{th} horizontal axis with a height δ and finally, p_i and $\widehat{p_i}$ are the pixels which locate symmetrically with respect to the i^{th} horizontal axis.

- *Spatial covering operator*: Similar to chromatic bilateral operator, spatial covering operator calculates the difference of foreground areas for two regions which are symmetric to the i^{th} horizontal line. It is given in the Equation 3.2:

$$S(i,\delta) = \frac{1}{J\delta}\left|A\left(B_{[i-\delta,i]}\right) - A(B_{[i,i+\delta]})\right|, \qquad \text{(Equation 3.2)}$$

where in $A\left(B_{[i-\delta,i]}\right)$, A shows the foreground area with a width of J and vertical extensions $[i-\delta, i]$.

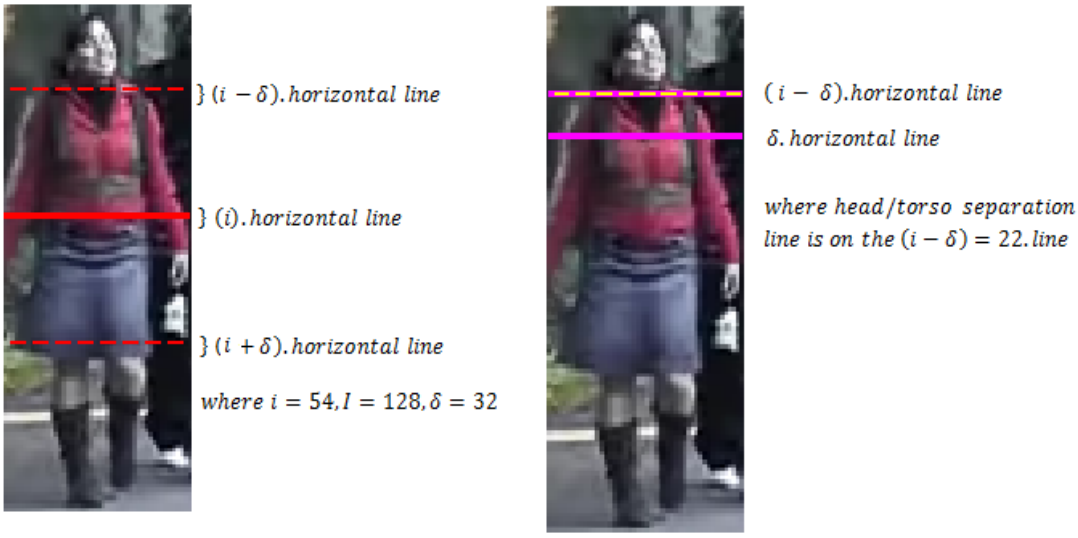


Figure 2. (a) Torso-leg separation, (b) Head-torso separation

i_{TL} has an interval between [δ, (I - δ)] = [I/4, 3I/4]]. In Figure 2 (a), the regions with a height of δ with respect to the i^{th} horizontal line is shown. The horizontal line which separates the torso and the legs, i, is found at the 54th slice where the height of the image is, I, 128 and the range is between δ =32, I - δ =96. The torso-leg separation axis is given by Equation 3.3. This equation gives the separating region with strongly different appearance and similar area.

$$i_{TL} = argmin_i \left(1 - C(i,\delta)\right) + S(i,\delta) \quad \text{(Equation 3.3)}$$

Once i_{TL} found, i_{HT} is searched between the range depending on i_{TL}, [δ, i_{TL}– δ], which can be between [0, $I/2$]. The magenta lines in (b) shows the head and torso separation line interval. Then, δ becomes 32 and i_{TL}- δ becomes 22. The yellow dashed line shows the i_{HT} line at 22nd slice. i_{HT} is given by Equation 3.4. The formula gives the separating regions that strongly differ in area.

$$i_{HT} = argmin_i \left(-S(i,\delta)\right) \quad \text{(Equation 3.4)}$$

3.1.2 Symmetric Partitioning (Vertical segments)

To find vertical symmetry axis, chromatic bilateral operator and spatial covering operator are used in a similar manner. Since head partition consists of very few pixels, it is assumed that the head does not contain much information. Therefore, the symmetry axis is found for only torso and legs. For symmetry search, chromatic bilateral operator and spatial covering operator are calculated as described above, not for the regions that locates symmetrically with respect to the i^{th} horizontal axis but for j^{th} vertical axis. Since we want to locate symmetric vertical axis, we now look for the minimum distance, i.e. the maximum similarity between the appearance and area of these two regions. Hence, both for torso and leg the vertical axis is given by the Equation 3.5.

$$j_{LRk} = argmin_j \, C(j,\delta) + S(j,\delta) \quad \text{(Equation 3.5)}$$

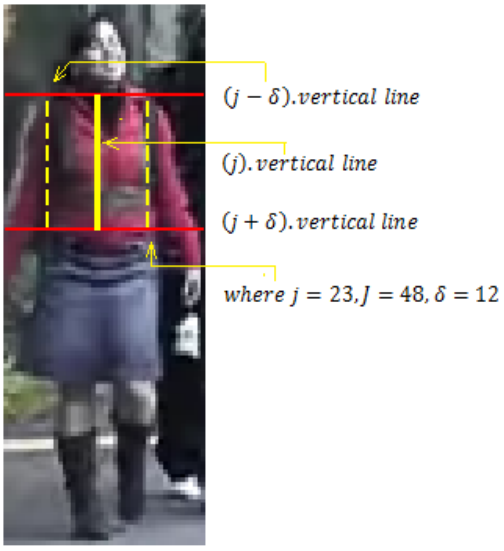


Figure 3. Symmetric Partitioning sample

Figure 3 shows the vertical axis, j=23, where the torso symmetry is found corresponding image. As described above, while calculating C and S, the regions are δ-width regions which are symmetric to the j^{th} vertical axis, where the width of image, J = 48, and δ = $J/4$, 12.

3.2 Feature sets

A feature is a numerical attribute that represents a local or global property of a given object. The selection of features usually depends on the problem under consideration. The discriminative abilities of some attributes usually impose their direct use in the inference model studied. In some cases, using a single set of features may not be successful to obtain a satisfactory inference performance. This usually leads to a decision of data integration to obtain higher accuracy. When the images are our concern, several feature sets have come into prominence in computer vision applications. In this thesis, we consider some of them that potentially provide more valuable discriminative information in recognition of pedestrian images. In addition to their single use, we also consider integrating them in a single framework using the model described in Section 3.3. The feature sets that are considered in the thesis are (1) color features described by weighted histograms, (2) texture features described by maximally stable color regions and (3) interest points matches with two common practical approaches, called SIFT (Scale-Invariant Feature Transform) and SURF (Speeded Up Robust Features). The details of these features are explained in the following subsections.

3.2.1 Weighted Color Histograms

Most common representation of a pixel is three numeric values that refer three color channels; red, green and blue. A color histogram is a frequency representation of the distribution of these colors in an image. An example color histogram is shown in Figure 4 for each RGB channel of a pedestrian image (the image numbered by 0221001 in VIPeR data set, see Section 4.1.1):

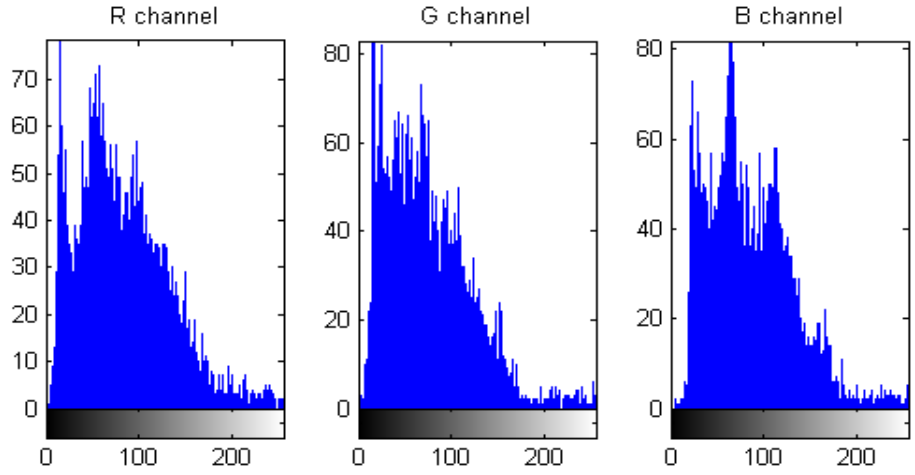


Figure 4: Histogram of each RGB channel of the image numbered by 0221001 in VIPeR dataset (Gray et al., 2007).

Since RGB *(Red, Green, Blue)* color representation is related with the amount of the light that hits the object, it is easily affected by illumination changes. Due to this problem of RGB representation, HSV *(Hue, Saturation, Value)* which is also called HSB *(Hue, Saturation, Brightness),* became the most widely used color space in person re-identification problem.

While in some cases value of each channel is used as a direct feature (Gheissari, 2006; Gray-Tao, 2008; Prosser, 2010; Zheng, 2011), remaining studies use color histograms (Cai, 2010; Farenzena, 2010). Colors are described in RGB color space as a combination of primary colors while in HSV outside of color, saturation and brightness terms are also used. HSV simulates the perception of color by human since we interpret the colors based on their hue, saturation and brightness.

- Hue represents the observed dominant color of an object (Plataniotis, 2000). It is represented on an angular dimension in which the start point at 0° is shown with Red. RGB to Hue conversion formula is shown in Equation 3.6:

$$MAX = max\{R, G, B\}, \quad MIN = min\{R, G, B\},$$

$$H = \begin{cases} undefined, & if\ MAX = MIN \\ 60\dfrac{G-B}{MAX-MIN} + 0, & if\ MAX = R\ and\ G \geq B \\ 60\dfrac{G-B}{MAX-MIN} + 360, & if\ MAX = R\ and\ G < B \\ 60\dfrac{B-R}{MAX-MIN} + 120, & if\ MAX = G \\ 60\dfrac{R-G}{MAX-MIN} + 240, & if\ MAX = B \end{cases} \quad \text{(Equation 3.6)}$$

- Saturation represents the amount of white light on the color. The more pureness of the color (such as red, violet) the less white appears in the color. However colors such as pink and lavender which becomes integration of red+white and violet+white respectively are less saturated (Plataniotis, 2000). RGB to Saturation conversion is done:

$$S = \begin{cases} 0, & if\ MAX = 0 \\ 1 - \dfrac{MIN}{MAX}, & otherwise \end{cases} \quad \text{(Equation 3.7)}$$

- Value is the brightness of the color (Equation 3.8). It has a range from 0 to 255; the former is dark and the latter is fully bright.

$$V = MAX \quad \text{(Equation 3.8)}$$

In this thesis, for each horizontal segment, except head part of the body, we use 16 bins for each of the channels which convey color information (H and S), but for brightness value, we use bigger intervals between bins by using 4 bins to keep effects of different pose and illumination conditions in minimum. The head part is not used in any of the feature extraction steps as suggested by Farenzena et al. (2010). It is mentioned that this part carries very low information in discriminating two images since the color content does not change significantly between two people. Since it is proven to be more effective in this problem, we use **Weighted Histogram** approach proposed by Farenzena et al. (2010). Weighted histogram is a more specific color histogram in which a vertical axis of

appearance symmetry is used to weigh histograms. To count the pixels near the symmetry axis more in the histogram, a one-dimensional Gaussian kernel is used to weight each pixel based on its position with respect to the symmetry axis found. In Figure 5 (a) pedestrian numbered by 0221001 in VIPeR data set, (b) its HSV image (for illustration, HSV is directly mapped into RGB) and (c) its Gaussian kernel used in Weighted Color Histogram calculation is shown.

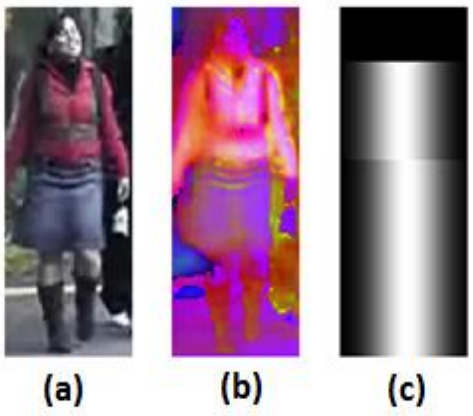


Figure 5: (a) Same pedestrian in Figure 4, (b) HSV image of this pedestrian, (c) Gaussian kernel

A distinct weighted histogram is extracted for each automatically extracted parts of the body; torso and leg. Further division of these parts is also considered obtaining several set of histograms for a single pedestrian image. Automated body part detection and the division of the image into lower segments are further elaborated in Sections 3.1 and 3.4.

3.2.2 Maximally Stable Color Regions (MSCR)

MSCR is an extension of Maximally Stable Extremal Regions (MSER) to color. The concepts of MSER are defined by Matas et al. (2002). A MSER in a gray level image, *I*, is given by:

$$R_t(x) = \begin{cases} 1, & if\ I(x) \geq t \\ 0, & otherwise \end{cases} \qquad \text{(Equation 3.9)}$$

where t consists of all possible threshold values in the gray level image, *I*. A MSER is the connected regions in R_t with a size change below a predefined threshold value over a range of thresholds. Figure 6 depicts the MSER region in the input image (a). All the coins in the figure are detected as stable regions since the change on the size of the connected regions is relatively small from (d) to (h).

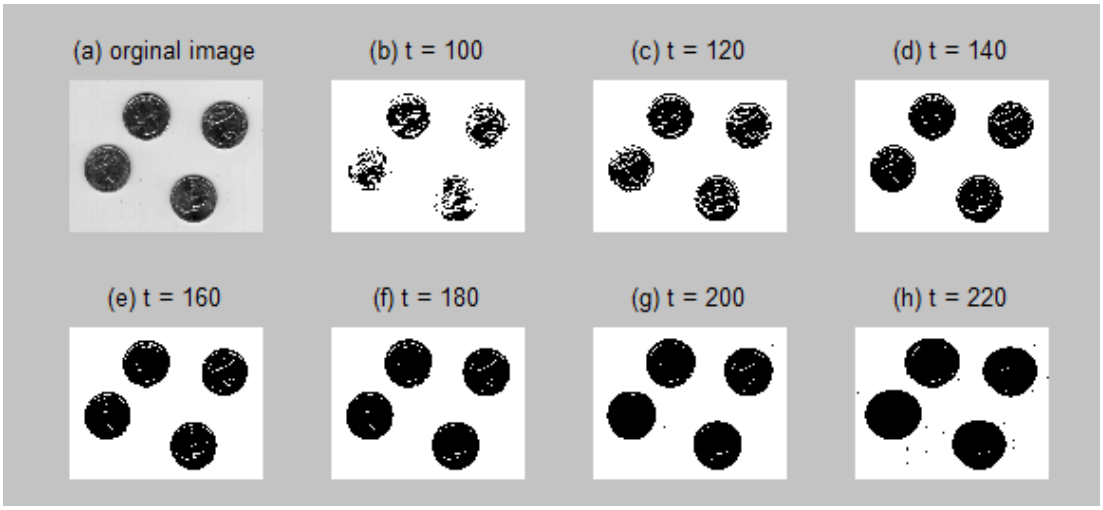


Figure 6: (b) to (h) depicts the MSER regions for the image in (a)

MSCR is an operator to detect a number of stable regions obtained by clustering of pixels based on spatial color distribution (Forssen, 2007). This agglomerative clustering process clusters neighboring pixels with similar colors in which similarity of the colors is modeled by using Chi-Squared Distribution. Figure 7 illustrates a MSCR of an image. The regions detected by this operator are abstracted by some of their properties such as area, centroid and average color. To get an attribute value for learning-based metric, the minimum distance of closest regions detected in two camera image is calculated.

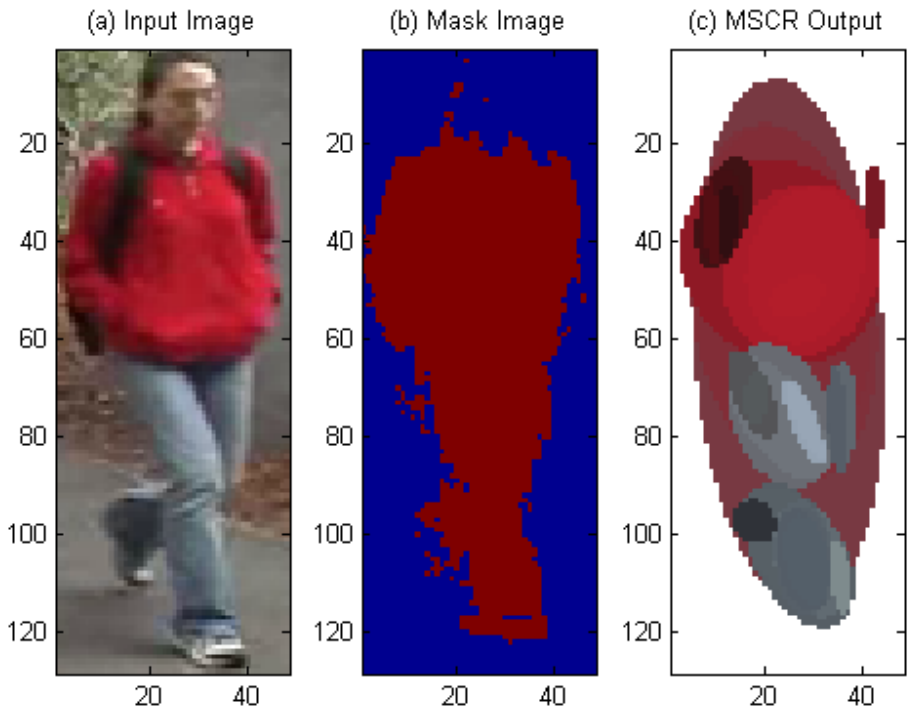


Figure 7: (a) Image numbered by 0010001 in VIPeR, (b) Mask Image, (c) MSCR image of (a)

3.2.3 Recurrent High-Structured Patches (RHSP)

RHSP, also called epitexture feature, consists of four main steps. Firstly, for each pedestrian, a random set of patches p are extracted from the foreground part. To eliminate non-informative ones, a thresholding operation which is based on the values of entropy of the patches is applied. After removing the uniformly colored patches, the transformation process is applied. To check each patch's invariance to geometric variations of the object, a set of patches $\hat{p}$ generated for each patch by considering rotations along the *y* axis of the patch. The third step shows how recurrent a patch is. For each patch in $\hat{p}$, Local Normalized Cross-Correlation is evaluated and an average map is obtained by merging and thresholding the maps. The thresholding takes place to discard small values in each map. The last step clusters the patches to avoid similar patches.

3.2.4 Interest Points

Using interest points for extracting local features has been widely studied in matching images belonging to the same object from different view-points. An interest point can be defined as a local pattern in an image that describes it in a highly distinctive way, independently from the color information. We have seen a few applications of this approach for the person re-identification problem. Several variants of interest points matching has appeared in the literature. We start the introduction with a basic model called SIFT and describe its two improved versions, SIFT-Flow and SURF, which are used in this thesis as one of the feature sets to represent the pedestrian images.

3.2.4.1 Scale-invariant feature transform (SIFT)

SIFT is an interest point operator that extracts distinctive scale and rotation invariant features which are also robust to noise, clutter, occlusion, illumination and 3D camera viewpoint changes (Lowe, 2004). After extracting interest points, the aim is to compute descriptors for these interest points. SIFT is composed of the stages below:

- **Detection of the scale-space extrema :** First stage aims to find stable keypoints in scale space using (Difference of Gaussians) DoG function. The candidate interest points are detected by finding local maxima and minima for the images computed using DoG function for different scales.

- **Keypoint localization:** First stage gives stable candidate keypoints. In this stage, some candidates are removed due to their low contrast, sensitivity to noise and close appearance to the edges.

- **Orientation assignment:** Each keypoint is characterized by the dominant gradient magnitude and the orientation calculated over designated pixels.

- **Keypoint descriptor:** For each keypoint, a 128-dimensional vector is assigned as a result of 4x4 array of histograms with 8 orientation bins for each.

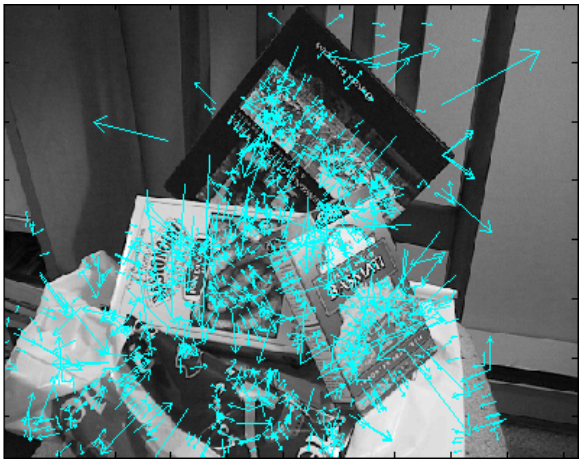


Figure 8: 1021 keypoints are found

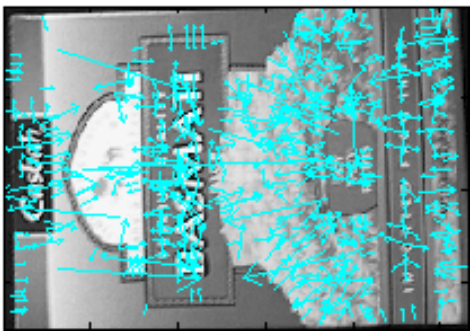


Figure 9: 579 keypoints are found

In the Figures 8 and 9, the location, the scale and orientation of keypoints are given on two different images. The length of the arrows indicate the scale in which keypoint found, orientation of the arrows shows the dominant gradient orientations assigned to each keypoint and root of the arrow shows their location. Matched features between these two images are displayed in Figure 10.

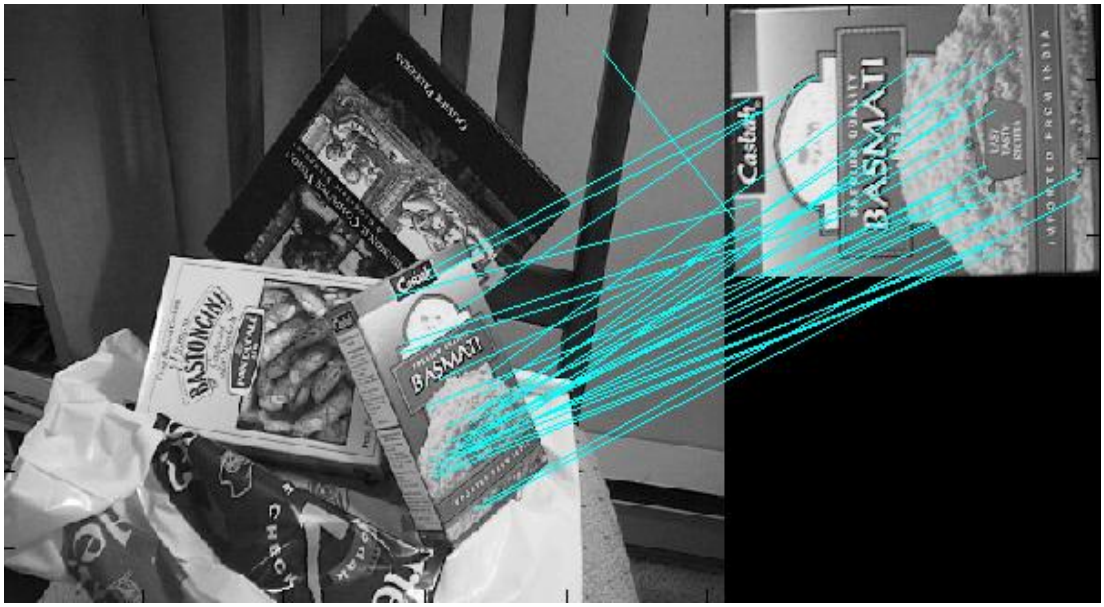


Figure 10: 34 matched keypoints found

3.2.4.2 SIFT Flow

SIFT Flow, a SIFT descriptor based approach, developed for *image alignment* (aka *image registration*) problem at scene level (Liu et al., 2011). In a large image database which consists of different scenes, SIFT Flow aligns the query image to its nearest neighbors in this database. The main idea behind the SIFT Flow approach is extracting SIFT features for all pixels in an image, not just for the keypoints. Liu et al. (2011) called this per-pixel SIFT

descriptor as *SIFT Image.* Pixel to pixel correspondences are used to find best matches and calculate warped images. Matching two SIFT images is very similar to *optical flow,* where an image is aligned to its temporally adjacent frame. The aim is to match SIFT descriptors from 2 images along the flow vectors by using the energy function below (Liu et al., 2008):

$$E(w) = \sum_{p} \left|\left| s_1(p) - s_2(p+w) \right|\right|_1 + \frac{1}{\sigma^2} \sum_{p} \left(u^2(p) + v^2(p)\right) + \sum_{(p,q)} min(\alpha |u(p) - u(q)|, d) + min(\alpha |v(p) - v(q)|, d) \quad \text{(Equation 3.10)}$$

where p is the index of pixels, $w(p) = (u(p), v(p))$ is the flow vector for every pixel. s_1 and s_2 represent the SIFT image for two frames, respectively. In image retrieval in a large database Liu et al. (2011) stated that the best matched pairs are the ones with the minimum energy.

Using SIFT Flow in this thesis, our aim is to calculate a distance matrix for all pedestrian images in a given dataset. For every image in the dataset, densely sampled SIFT features are extracted and for each combination of these *SIFT images* the energy values are calculated as described in the method above.

3.2.4.3 SURF (Speeded Up Robust Features)

Similar to SIFT, SURF is also scale and rotation-invariant interest point detector and descriptor (Bay et al., 2008). However, it is more robust and much faster than SIFT approach. This strong performance is achieved by the use of **Integral Image** and changing the methods used in **interest point detection and** extracting **descriptors**.

- **Integral Image:** The computational time and efficiency provided by Hessian matrix leads Bay et al. to use it as a detector. The Fast Hessian detector relies on *integral image* described by Bay et al. (2008) for image convolutions.

- **Detector and descriptor:** For detection, Hessian matrix is used for speed considerations, as described above. For descriptor part, the sum of the Haar wavelet response around the point of interest is used. Also, instead of using a 128 dimensional descriptor, it is reduced to 64.

In this thesis, we used SURFmex implementation[1] to find and match interest points. Number of matched interest points is counted as a similarity measure and used as an attribute in Fisher discriminant function.

[1] Available in: http://www2.maths.lth.se/matematiklth/personal/petter/surfmex.php

3.2.4 Distance Metrics

To compare histograms obtained from two images or image parts, several distance metrics can be used. Given two vectors X = (x_1, x_2, …, x_n) and Y = ($y_1, y_2, \ldots, y_n$), say that they correspond to two distinct histograms, we can use the following metrics to identify their similarity. Indeed, these metrics usually define a distance between two samples to evaluate their dissimilarity, which in turn can be used a similarity measure:

- **Euclidean Distance:** Euclidean Distance, also known as 'Pythagorean distance' is one of the most widely used distance metric. The distance between the points in X and Y is calculated as:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad \text{(Equation 3.11)}$$

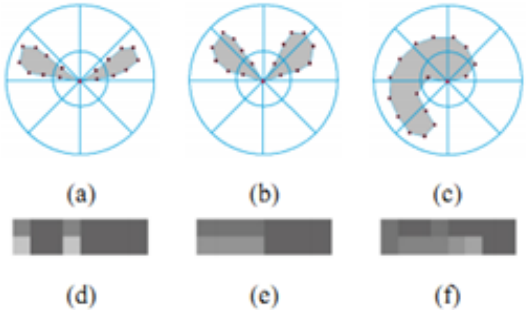
- **Chi (X^2)-Square Distance:** X^2 distance is useful when making a bin-to-bin comparison on histograms (Pele and Werman, 2010). The bin-to-bin dissimilarity measures just compare the corresponding histogram bins. Here, it is theoretically assumed that the domain of the histograms is aligned. The name of this distance metric comes from Pearson's X^2 squared test statistic, (Pearson, 1900) which is used to show likeliness of one distribution being drawn from another one. In some cases, e.g. while comparing histograms; the difference between small bins becomes more important than the difference between large bins. This metric considers this issue and reduces the difference between small bins by:

$$X^2(X,Y) = \frac{1}{2}\sum_{i} \frac{(X_i - Y_i)^2}{(X_i + Y_i)} \qquad \text{(Equation 3.12)}$$

  This metric is used in several domains such as texture and object categories classification (Cula et. Al, 2004; Zhang et al., 2007; Varma et al., 2009), local descriptors matching (Forssen et al., 2007), shape classification (Belongie et al., 2002; Ling et al. 2007) and boundary detection (Martin et al., 2004).

- **Earth Mover's Distance:** Unlike bin-to-bin dissimilarity measures which compare corresponding histogram bins, in cross-bin measures, non-corresponding bins are also compared. Main disadvantage of bin-by-bin comparison is the assumption that aligned histogram bins are not practically possible due to the lightening conditions or noise effects. Earth Mover's Distance is developed to overcome this problem (Rubner et al., 2000). This method defines the distance between two histograms as a solution of the transportation problem in which the minimal cost while transforming one histogram to another is tried to be found (Rubner et al., 1997). An example is illustrated in Figure 11, in which EMD performs superior to common bin-to-bin comparison metrics. (a), (b) and (c) shows three shapes and log-polar grid on them. (d), (e) and (f) corresponds 2D histograms of the figures on them using the same 2D grids. (g) summarizes the distances calculated by using EMD and three bin-to-bin distance metrics, L1, L2 and X^2 between the histograms of the figures (a),

(b) and (c). In spite of the huge difference between the 2D histograms of (d) and (e), their corresponding figures, (a) and (b) has a small change in the blobs on them. This large change causes all bin-to-bin distance functions incorrectly describe that the similar pairs (d(a,b) > d(a,c)). However, EMD correctly states that the similarity between (a) and (b) is approximately 3 times larger than (b) and (c).



Distance	$d(a,b)$	$d(b,c)$
L_1	1.0	0.875
L_2	0.3953	0.3644
χ^2	0.6667	0.6625
EMD	0.5	1.5625

(g)

Figure 11: An example where EMD measure performs better than bin-to-bin dissimilarity measures. Source: Ling and Okada (2007)

Considering two signatures, $P = \{(p_i, u_i)\}_{i=1}^{m}$ and $Q = \{(q_j, v_j)\}_{j=1}^{n}$ where the elements in P are supplies with a size m and located at u_i and the elements in Q are demands with a size n and located at v_j, p_i and q_j give us the amount of supply and demand respectively. The EMD between two signatures is formalized as:

$$EMD(P,Q) = \min_{F=\{f_{ij}\}} \frac{\sum_{i,j} f_{ij}\, d_{ij}}{\sum_{i,j} f_{ij}}, \qquad \text{(Equation 3.13)}$$

with following constraints:

$$\sum_j f_{ij} \le p_i, \quad \sum_i f_{ij} \le q_j, \quad \sum_{i,j} f_{ij} = \min\{\sum_i p_i, \sum_j q_j\}, \quad f_{ij} \ge 0 \qquad \text{(Equation 3.14)}$$

where $F = \{f_{ij}\}$ consists of a set of flows. f_{ij} is the flow between p_i (supplies) and q_j (demands). The aim is to find the flow f_{ij} that minimizes the amount of work. d_{ij} called the ground distance between the position u_i and v_j. This formulation can be used accordingly for the histogram vectors *X* and *Y* described above.

- **Bhattacharyya Distance:** Originally, in statistics, this distance measure is used to compare two probability distributions (Bhattacharyya, 1943). However, in our case, we use X and Y, the distributions of histogram bins, instead of probability distributions. Bhattacharyya distance is defined as,

$$d(X,Y) = \sqrt{1 - \rho(X,Y)}, \qquad \text{(Equation 3.15)}$$

where $\rho(X,Y)$ denotes Bhattacharyya coefficient (measure) (Bhattacharyya, 1943).

$$\rho(X,Y) = \sum_{i=1}^{N} \sqrt{X(i)Y(i)}. \qquad \text{(Equation 3.16)}$$

This measure is widely used in Computer Vision and Pattern Recognition, especially in surveillance systems (Sharif et al., 2010), feature extraction and selection (Xuan et al., 2006; Ke et al., 2010; Choi et al., 2003), clustering (Mak et al., 1996), recognition systems (You et al., 2010) and also in several domains such as Statistics-Theory (Chaudhuri et al., 1991), Communication Technology (Kailath, 1967).

3.3 Learning-based metric

Our model takes two images as input and decides if they belong to the same person. Each image is taken from disjoint cameras and corresponds to the whole body view of a pedestrian. Since the model is built upon a discriminative framework, the problem turns out to be a binary classification problem where an image pair is represented by a fixed-length of the feature vector and the system reports their match as a positive prediction. To settle a confidence measure for any pedestrian matching, we define the problem using a linear discriminant function which in turn will provide a rank for predicted matches:

$$y = w_1x_1 + w_2x_2 + \ldots + w_kx_k. \qquad \text{(Equation 3.17)}$$

Here, x_i denotes any attribute that represents the similarity (or distance) between two pedestrian images in some evaluation criteria, and w_i denotes the contribution of that feature in the final decision. y is a measure of potential match between two pedestrians included in the pair images. It also represents the confidence of prediction; i.e. a higher value indicates a higher probability of a correct match. Given a set of known matched/unmatched pairs, the problem is to find w which optimizes a function that discriminates best between correct and incorrect matches. In order to solve this equation we used 2 different methods:

3.3.1 Fisher Linear Discriminant Analysis (FLDA)

The main idea of the FLDA is to search for a projection line that well separates the objects from predefined classes (Fisher, 1936). The result of a linear combination of features created by LDA gives us the largest mean differences between the desired classes (Martinez and Kak, 2001).

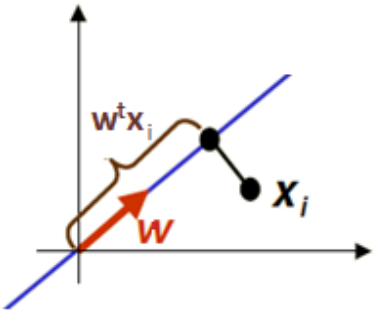


Figure 12: Projection of sample

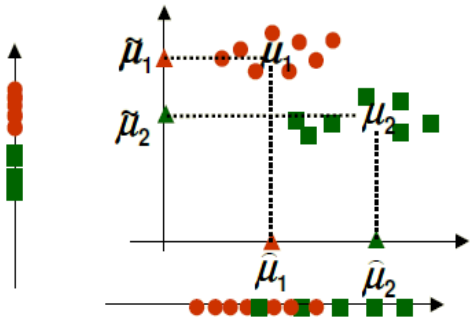


Figure 13: Means of classes 1 and 2

In Figure 12, the distance of projection of sample x_i is given by $w^t x_i$ where w^t is a unit vector which shows the line direction. Figure 13 shows the means of classes 1 and 2 ($\mu_1$and μ_2) and the means of the projection of classes 1 and 2 (horizontal projections: $\widetilde{\mu_1}$ and $\widetilde{\mu_2}$ vertical projections: $\widehat{\mu_1}$ and $\widehat{\mu_2}$) where $\widetilde{\mu_1} = w^t\mu_1$ and similarly $\widetilde{\mu_2} = w^t\mu_2$. From the definition of FLD above, the larger the distance of projections means the better is the expected separation. From the Figure 13, it is seen that the vertical axes is a better line than the horizontal axes to project to for class separability. However the distance in horizontal axes is bigger than vertical one. In order to eliminate such problems, the variance of the classes must be considered. Therefore, the means must be normalized by a factor which is proportional to variance called scatter. Scatter is the spread of data around the mean. If we define projected samples as y_i, given by:

$$y_i = w^t x_i \qquad \text{(Equation 3.18)}$$

then the scatter for projected samples of class k becomes:

$$\tilde{s}_k^2 = \sum_{y_i \in Class\ k} (y_i - \widetilde{\mu_k})^2 = w^t S_k w, \qquad k = 1, 2 \qquad \text{(Equation 3.19)}$$

The objective function which creates a linear combination of the classes becomes:

$$J(w) = \frac{(\widetilde{\mu_1} - \widetilde{\mu_2})^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \qquad \text{(Equation 3.20)}$$

To maximize this objective function, projected means must be far from each other while scatter values in each class must be as small as possible which means each class must be clustered around their projected means. To this end, two measures should be defined:

- **Within-class scatter matrix**: Define the separate class scatter matrices S_1 and S_2 for classes 1 and 2. They measure the scatter of original samples x_i (before projection):

$$S_1 = \sum_{x_i \in Class\ 1} (x_i - \mu_1)(x_i - \mu_1)^t \qquad \text{(Equation 3.21)}$$

$$S_2 = \sum_{x_i \in Class\ 2} (x_i - \mu_2)(x_i - \mu_2)^t \qquad \text{(Equation 3.22)}$$

now within-class scatter matrix can be defined as:

$$S_w = S_1 + S_2 \quad \text{(Equation 3.23)}$$

- **Between-class scatter matrix:** measures separation between the means of two classes (before projection) and is given by

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t \quad \text{(Equation 3.24)}$$

FLD tries to minimize within-class measure while maximizing the between-class measure. Rewriting the objective function gives the final criterion to maximize for FLDA:

$$J(w) = \frac{(\widetilde{\mu_1} - \widetilde{\mu_2})^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{w^t S_B w}{w^t S_w w} \quad \text{(Equation 3.25)}$$

This function is proven to be optimized when;

$$w = {S_w}^{-1}(\mu_1 - \mu_2) \quad \text{(Equation 3.26)}$$

The computation of w refers to the training phase of FLDA. Each sample is then predicted by its projection into separating line by the formulation that we introduce in our learning-based metric.

3.3.2 Support Vector Machines (SVM)

SVM classifier aims to find the optimal hyperplane that separates the data into two categories in which one side of the plane consists of the first category of the target variable and the other side consists of the second category. The N-dimensional hyperplane constructed by SVM analysis maximizes the margin between support vectors, the vectors near the hyperplane.

SVM must deal with:

(a) more than two predictor variables,
(b) handling the cases where clusters cannot be completely separated,
(c) nonlinear dividing lines ,
(d) handling classifications with more than two categories.

These issues are clarified in more details below:

a. The separating hyperplane: As described above, finding a separating hyperplane on a 2-dimensional plane geometrically means that inferring a rule that corresponds to drawing a line between the two clusters. For the data which are in N-dimensional space, there exists a separating hyperplane in N-1 dimension.

b. Soft-margin: In many cases, the real data sets cannot be linearly separable; instead, some samples will appear that breaks the linear separability. In such cases, the amount of overlap between two categories is controlled by a cost parameter, to allow some error while separating them. In this way, a so-called soft-margin can be

created, which enables the seperating hyperplane to keep some samples inside the margin without changing its final position, and thus not affecting the training result. A high cost value C forces the SVM to create a more accurate model that may not generalize well and increase the cost of misclassification while a lower cost parameter leads to a simpler prediction function.

c. **Nonlinear dividing line:** If the points are separated by a nonlinear region a nonlinear dividing line must be needed. However, to find the optimal hyperplane, nonlinear curves are not tried to fit the data. The data is mapped into upper spaces using a kernel function. The aim of the kernel function is to perform optimal separation in higher dimensional space even in complex boundaries.

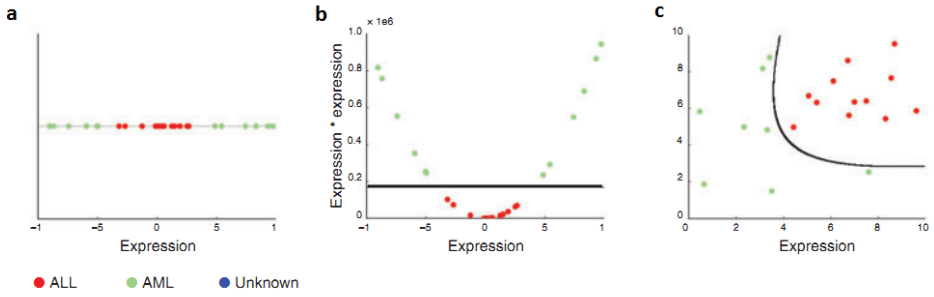


Figure 14: (a) A non-separable 1D dataset, (b) Separation of (a), (c) A linearly non-separable 2D dataset

Figure 14 (a) illustrates a nonseparable data distribution. While all AML values have large absolute values, ALL examples are grouped near zero. Since there is not any single point that can separate the two classes, the values are squared to get a new dimension. As seen in the Figure 14 (b), in this dimension, a straight line can separate the examples easily. In Figure 14 (c) the two-dimensional data cannot be separated linearly so by calculating the products of all pairs of features two-dimensional data, it is projected to the four-dimensional space in which a kernel can be found. The data cannot be drawn in a four-dimensional space, but it can be projected the SVM hyperplane in that space back down to the original two-dimensional space. The result is shown as the curved line in Figure 14 (c).

Some common kernel functions used in SVM applications are as follows, where x_i and x_j are vectors in the input space:

1. **Linear:** In Linear kernel no mapping is done. Linear discrimination (or regression) is done in the original feature space. Linear kernel is defined as:

$$K(x_i, x_j) = (x_i{}^T x_j) \qquad \text{(Equation 3.27)}$$

2. **Polynomial:** The Polynomial kernel is considered to be a non-stationary kernel. This kernel is especially convenient for the problems in which the training data is normalized before. For degree-d polynomials, it is defined as:

$$K(x_i, x_j) = (\gamma x_i{}^T x_j + r)^d \qquad \text{(Equation 3.28)}$$

Scalar parameter γ, constant term r and polynomial degree d are adjustable parameters.

3. **Radial basis function (RBF):** The RBF kernel is the most common kernel, which defined over a Gaussian assumption to map two data points using the following function (Vert et al., 2004):

$$K(x_i, x_j) = \exp(-\frac{d(x_i, x_j)^2}{2\sigma^2}) \quad \text{(Equation 3.29)}$$

where σ denotes a parameter and d the Euclidian distance between two vectors.

Because of its popularity and reported success in computer vision applications such as object category classification and detection (Sreekanth et al., 2010), in this thesis, RBF kernel is used to separate between the samples that correspond to matched and unmatched image pairs.

3.4 Localized histograms on horizontal segments

While two body regions, torso and leg, have reasonably different characteristics due to different clothes worn in those parts, some internal differentiations can appear within each region. This issue has not been considered in previous studies including Farenzena et al. (2010). Lower part of the leg could be dominated by the color of shoes, which is not necessarily the same as trousers or skirt. For a person wearing a skirt and a boot, three horizontal segments (corresponding to the skirt, leg and boots) may appear with significantly different color characteristics in the same camera view, whilst the general characteristic is conserved in other camera view. Therefore, we argue that these semantic body parts can be separated into further horizontal segments for better discrimination. Since further divisions cannot be guided by a general semantic rule applicable to all images, we propose to create lower segments in each body part with vertically uniform lengths. Figure 15 (a), (c) shows an image pair representing the same pedestrian. The body is first segmented into torso and leg. Then each region is further divided into three sub horizontal segments with equal length (b), (d). Note the color difference between distinct segments in one leg and the similarity between two lower leg segments in different images. In following sections, we use the following notation to explain the use of this technique in histogram calculation: An attribute describing the whole body distance using HSV histogram is denoted with *hist*. When the body is divided into two semantic parts, this attribute is denoted with *histPart*, and if the body is divided into *K* total segments with further divisions, then the attribute is referred to as *histKseg*.

(a) (b)

(c) (d)

Figure 15: (a) Image from the first camera, (b) Horizontal segments for the first camera image, (c) Image from the second camera, (d) Horizontal segments for the second camera image.

CHAPTER 4

RESULTS

4.1 Experimental setup

We use a common experimental setup to assess the performance of our model. This section will introduce the datasets used and the procedures to evaluate the results.

4.1.1 Dataset

The experimental evaluation of the methods developed for the person re-identification problem requires a dataset composed of a set of pedestrian image pairs taken from two disjoint camera views. The images taken from one camera is usually called as the gallery set, while the ones taken from other camera, for which the reidentification is desired, is called as the probe set. In this thesis, it is expected to match single appearances of two pedestrians from different cameras, while some other approaches use more than one image belonging to same person in one camera view. Therefore, a single-shot data set is needed to evaluate the introduced methods. When a multiple-shot set is available, each case should be considered as a distinct sample. In this thesis, two challenging public benchmark datasets, VIPeR (Gray et al., 2007) and ETHZ (Schwartz and Davis, 2009), are used with their single-shot versions.

Provided by Gray et al. (2007), the most challenging benchmark dataset currently available in person re-identification area is the VIPeR (viewpoint invariant pedestrian recognition) dataset. In many of the previous studies, this dataset is used to evaluate the identification performance of the proposed methods. It involves 632 pedestrian image pairs taken from two disjoint cameras. The images are captured from different locations over the course of several weeks. There are four main viewpoint angle changes, 45, 90, 135 and 180. In order to create a viewpoint invariant model, using 45 degree segments, $\binom{8}{2}$ = 28 different viewpoint pairs should be used. But using symmetry, Gray et al. (2007) reduce 28 to 10. The viewpoint angles for two different images belonging to one person may vary: 45-0, 90-0, 90-45, 135-0, 135-45, 135-90, 180-0, 180-45, 180-90, 180-135. Images are cropped and normalized to 128x48 pixels. Each pair shows the same person with a different pose, illumination and viewpoint. These significant changes in pose, illumination and viewpoint make re-identification a very challenging task. Some challenging cases are shown in Figure 16. Each column shows the same person from different camera views.

Figure 16: Sample images from VIPeR dataset (Gray et al., 2007)

In the second dataset, ETHZ, which was originally proposed to be used for the evaluation of pedestrian detection performance (Ess et al., 2007), but adopted by Schwartz and Davis (2009) for person re-identification task, each person was recorded using moving cameras as a video sequence. Images are recorded at a resolution of 640x480 pixels and at 15 FPS using a stereo pair of cameras mounted on a children's stroller. The pose variations are not as much as VIPeR but there are higher illumination changes and occlusions. The dataset consists of 3 sequences: 1st sequence contains a total of 4857 images of 83 pedestrians and is taken under similar weather conditions, 2nd sequence contains a total of 1961 images of 35 pedestrians including people moving in all directions and the last one contains a total of 1762 images of 28 pedestrians and is taken on a sunny day on a sidewalk. A few example images are shown in Figure 17.

Figure 17: Sample images from 1st sequence of ETHZ dataset (Schwartz and Davis, 2009)

4.1.2 Procedure of evaluation

In this thesis, all experiments are performed as a leave-one-out cross-validation. For training purposes, for each person from the first camera his/her image from the other camera is used as positive sample and remaining pairs are used as negative samples. However, a subset of this training set is used for each testing iteration. While comparing and calculating the distance values between the first image, p1, and all remaining images, such as (p1, p2), (p1, p3), …, (p1, p1264), firstly, all sample pairs including p1, either matched or unmatched, are removed from training set and the remaining subset of the data is used to train a classifier. Secondly the model given by the classifier is used to test all combinations between p1 and the remaining values.

For validation, two image sets, so-called gallery and probe sets, are determined. In the VIPeR dataset, the probe set consists of images from the first camera and images taken from the other camera are used as gallery set. In previous studies, the best results on VIPeR dataset was reported by Farenzena et al. (2010). To compare our results with theirs fairly, we employ the same splitting strategy in our validation. We run our method on 10 random subsets containing 316 pedestrians in the gallery set and take the average of the results. In ETHZ dataset, each person has a different number of images. To have only a single-shot case, the experiments are done on the first sequence of ETHZ dataset, as suggested by Farenzena et al. (2010). To build the gallery set, a random image for each person is selected and remaining images are used to construct the probe set. Then, the model evaluates the potential match between probe and gallery set. For every image in the probe set, the rank of the correct match is found. This entire procedure is repeated 10 times.

All experiments are conducted on an Intel Core i7-2600, 3.4 GHz CPU with 8 GB of RAM running on Windows 7 operating system. The implementation is based on MATLAB. Creating positive and negative pairs and training the FLD classifier takes 12.16 seconds for each person on VIPeR dataset on average.

To discern the ability of person reidentification methods, a common metric used is CMC (Cumulative Matching Characteristics) curve suggested by Gray et al. (2007). CMC curve shows the expectation of finding the correct match in the top *n* matches. A rank *n* matching rate indicates that the percentage of the images in the probe set correctly found in gallery set in the top *n* ranks. Moreover, as Farenzena et al. (2010) does, we also consider another useful measure, Area Under Curve (AUC). For a perfect identifier, AUC becomes %100. While it is common to compute AUC for the curve covering all hits until the target match is found, we argue that a good identification system should perform well in some early hits. Hence, we also use a new criterion called AUCk, to compute the area under the curve where the targets are found in first k hits. Since the ultimate goal of these surveillance systems is to detect the person in earliest hits, we believe that this measure is a better representation of the success of tested method.

We assess the following issues in our experiments in the common benchmark setup defined above:

- Effect of body part division and further segmentation of body parts on final accuracy,
- Comparison of individual use of different features for person re-identification,

- The performance of integrative model based on learning-based metric,
- Comparison of different learning methods,
- Robustness of the best method by evaluating its performance on different datasets.

Finally, we compare our best result with the method proposed by Farenzena et al. (2010).

4.2 Empirical results

Several experiments are conducted over the benchmark datasets described. For each experiment, detailed CMC curves and AUC tables are depicted. The results are reported rigorously to explain the effect of all contributors in the methodology applied.

4.2.1 HSV Histogram

In the previous studies, it was already proven that HSV color space has better performance than the others (such as RGB and YCbCr) in person reidentification (Gray and Tao, 2008). Therefore in this thesis, we adopt HSV histogram representation. First decision issue regarding the application of HSV histograms is whether to extract a histogram for the whole body or separate histograms for the body parts such as torso and leg. The head part is excluded in our experimental evaluation since it was already shown not to provide any positive contribution (Farenzena et al., 2010). Our second aim in this experiment is to observe the effect of weighted histogram in which the pixels near the vertical symmetry axis count more than others in the final histogram. The Table 2 and the Figure 18 shows the results of the previously mentioned items.

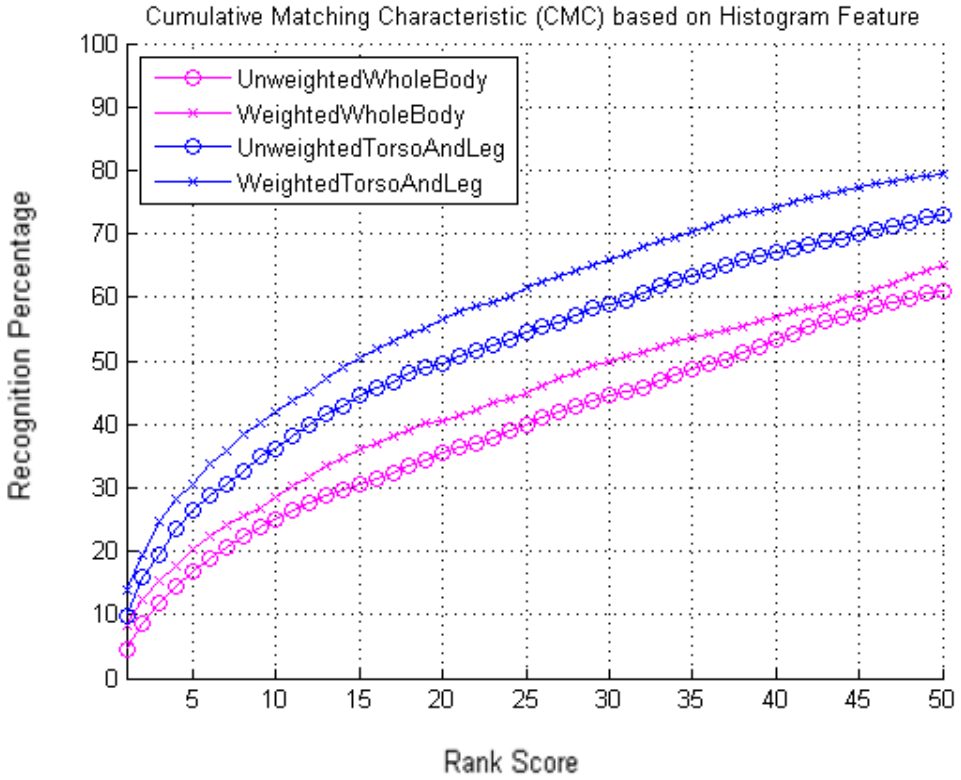


Figure 18: CMC Results based on Histogram Feature in the VIPeR dataset

From the Figure 18, it is seen that, regardless of the fact that the body is partitioned or not, the weighted histograms show superior performances than the unweighted ones. We also observe that dividing the body into torso and leg part can improve the discriminative performance of the histogram feature by %5 in overall AUC100. Upon this conclusion, the weighted HSV histograms obtained from both torso and leg are fed separately in remaining part of the thesis.

Table 2. AUC results based on Histogram Feature

histogram used	AUC 100(%)	AUC 50(%)	AUC 20(%)	AUC 10(%)	AUC 1(%)
UnweightedWholeBody	83.45	39.05	23.88	16.71	4.72
WeightedWholeBody	84.70	43.35	28.11	20.11	8.30
UnweightedTorsoAndLeg	88.52	51.87	35.25	25.84	9.83
WeightedTorsoAndLeg	**90.03**	**58.97**	**41.27**	**30.72**	**14.02**

4.2.2 Comparison of distance metrics

To compare HSV histogram in the previous experiments, the Bhattacharyya distance is used. To evaluate the performance of other distance measures, we repeat the experiments with three other distance metrics: Euclidian distance, EMD distance and chi-square distance. The results using the weighted HSV histograms obtained from both torso and leg are shown in the Figure 19:

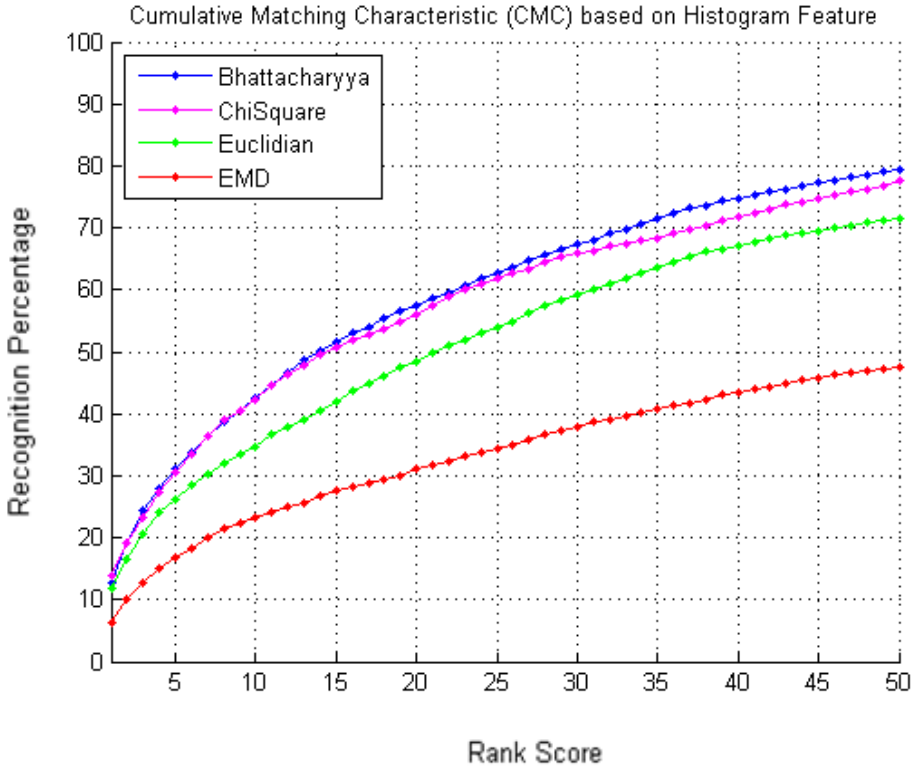


Figure 19: Different distance metrics used on Histogram comparison in VIPeR dataset

As the CMC curve given in Figure 19 and AUC records (including whole AUC and its all partial calculations) in Table 3 consistently demonstrate, the Bhattacharyya distance gives the best results in comparison of HSV histograms. In chi square distance, which is useful when making a bin-to-bin comparison on histograms, the AUC values are higher than the others except the Bhattacharyya measure. However, surprisingly, the other histogram comparison metric, EMD distance, provides worse results than Euclidian distance. This is probably due to the fact that the illuminations changes, which are usually promised to be recognized by EMD measure, are already considered by HSV histograms in an implicit way.

Table 3. AUC results obtained using different distance metrics

Metric used	AUC 100(%)	AUC 50(%)	AUC 20(%)	AUC 10(%)	AUC 1(%)
wHSVEmd	69.74	33.21	22.16	16.64	6.40
wHSVEuclidean	87.58	51.35	34.26	25.85	11.93
wHSVChiSq	89.49	57.50	40.71	30.58	**13.86**
wHSVBhattacharyya	**90.03**	**58.97**	**41.27**	**30.72**	12.75

4.2.3 Feature Sets

The common feature representations that are used in this domain are described in the Section 3.2 in detail. In this subsection, the effect of the widely used color, texture and interest point features are to be analyzed.

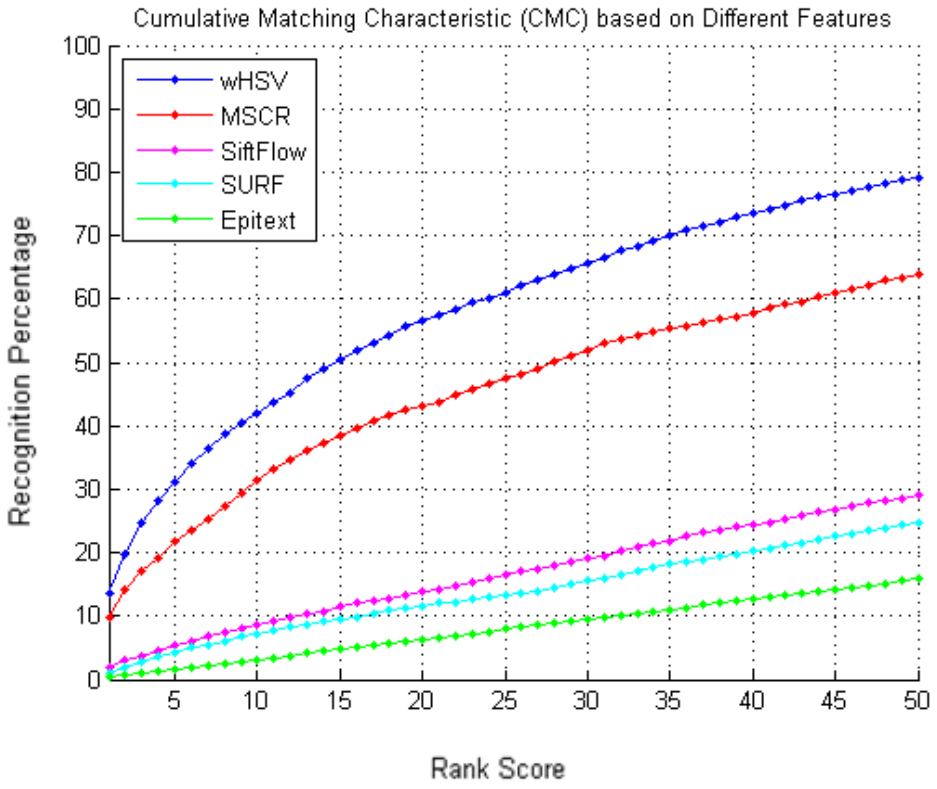


Figure 20: Comparison of different feature sets as single attributes for person reidentification in VIPeR dataset

From the Figure 20, it is clear that the color based features show superior performances than the texture based features. The most distinctive feature is weighted histogram whereas maximally stable color regions comes after that. Because of the low resolution of the images, the features based on keypoints matching cannot be representative enough. Subsequently, they do not show good performances on this dataset. However, we will revisit both SiftFlow and SURF features to see their contribution in overall performance when combined with other features.

Table 4. AUC results obtained using different feature sets as single attributes

Metric used	AUC 100(%)	AUC 50(%)	AUC 20(%)	AUC 10(%)	AUC 1(%)
EpitextDistance	50.15	8.07	3.32	1.74	0.32
SURFDistance	57.05	13.75	7.08	4.42	1.04
SiftFlowDistance	61.45	16.59	8.56	5.51	1.84
MSCRDistance	83.29	45.06	30.30	21.87	9.72
wHSVDistance	**89.79**	**58.05**	**40.81**	**30.91**	**13.7**

4.2.4 Combining Features

In this subsection, the results of combining different features described above are given. While combining features the same matching distance formula is used as proposed by Farenzena et al. (2010).

$$\begin{aligned} d(I_A, I_B) \quad = \quad & \beta_{WH}.\, d_{WH}\big(WH(I_A), WH(I_B)\big) \\ & + \beta_{MSCR}.\, d_{MSCR}\big(MSCR(I_A), MSCR(I_B)\big) \\ & + \beta_{RHSP}.\, d_{RHSP}\big(RHSP(I_A), RHSP(I_B)\big) \end{aligned}$$

Farenzena used this formula while comparing two pedestrians, I_A and I_B.

The distance d_{WH} evaluates the weighted color histograms extracted from the images of two pairs on different disjoint cameras. It is calculated via Bhattacharyya distance metric. d_{MSCR} shows the minimum distance of each MSCR element b in I_B to each element a in I_A. d_y^{ab} , that compares the y component of the MSCR centroids, and d_c^{ab} , that compares their mean color are the two components that are used to calculate the distance d_{MSCR}. Euclidian distance is used to calculate for both d_y^{ab} and d_c^{ab} . d_{MSCR} formula is:

$$d_{MSCR} = \sum_{b \in I_B} \min_{a \in I_A} \gamma.\, d_y^{ab} + (1-\gamma).\, d_c^{ab}$$

where γ takes values between 0 and 1.

Recurrent High-Structured Patches (RHSP), epitexture feature, is a texture feature that shows how much a patch is recurring. This step involves the accumulation of potential patches based on some color changes and ranking them in terms of the continuity of detected motifs inside the patches and the recurrence in several parts of the image. d_{RHSP} is obtained by selecting the best pair of RHSP, one in I_A and one in I_B. The minimum Bhattacharyya distance is evaluated among the RHSP's HSV histograms. This calculation is done independently for each body part. The final distance is then computed by normalizing the sum of these independent values.

In the experiments, Farenzena et al. (2010) fixed all the β parameters to 0.4, 0.4 and 0.2 respectively. We use their parameter set in this experiment without any change. Other distance measures are SiftFlow and SURF distances which use the matched keypoint count the generate distance matrix.

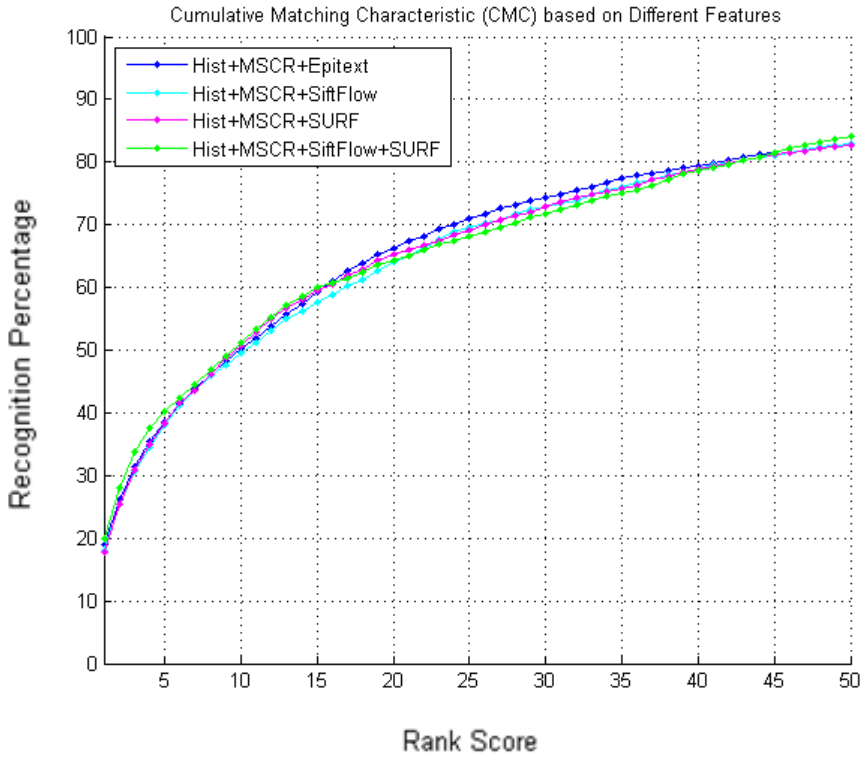


Figure 21: Combination of features in VIPeR dataset

Here, the first title in CMC curve and the AUC table, ***Hist+MSCR+Epitext,*** is the combination used by Farenzena et al. (2010). It should be noted that, Epitexture feature is not useful at all since the distance measure based on this feature produces the same value for all images. From the Figure 21 there is not any combination with an apparent advantage over the others. It is seen that the last combination, ***Hist+MSCR+SiftFlow+SURF,*** performs a slightly better performance on earlier hits such as 10, 15, 20 in comparison with the remaining combinations. Table 5 also confirms that the best performance is obtained using 4 features, weighted Histogram, MSCR, Siftflow, SURF on small hits. However, since Siftflow and SURF are in the same feature family, both used as an interest point detector and descriptor, and using two of them at the same time does not show any significant change. Hence, we decide not to use both of them. Because the results obtained using SURF with Histogram

and MSCR are better than using Siftflow, we select best feature combination set as ***Histogram, MSCR and SURF***. These 3 features are used in remaining experiments.

Table 5. AUC results obtained from different combinations of features

Combination used	AUC 100(%)	AUC 50(%)	AUC 20(%)	AUC 10(%)	AUC 1(%)
Hist+MSCR+Epitext	***91.89***	64.86	47.90	37.37	19.21
Hist+MSCR+SiftFlow	91.45	64.51	47.73	37.45	18.10
Hist+MSCR+SURF	91.84	64.88	48.75	37.81	17.69
Hist+MSCR+SiftFlow+SURF	*91.79*	***64.94***	***49.50***	***39.33***	***19.84***

4.2.5 Effect of Learning Based Metric

In this subsection, we discuss the effect of learning based metric on this problem. In the previous subsection while comparing two pedestrians, we used matching distance formula proposed by Farenzena et al. (2010) in which the β parameters, the coefficients of the features, are fixed**.** Our aim is to use a learning based metric to find optimal values for each β coefficient. Using the same features used by Farenzena, Histogram, MSCR and Epitexture, we got better results using FLD (Figure 22).

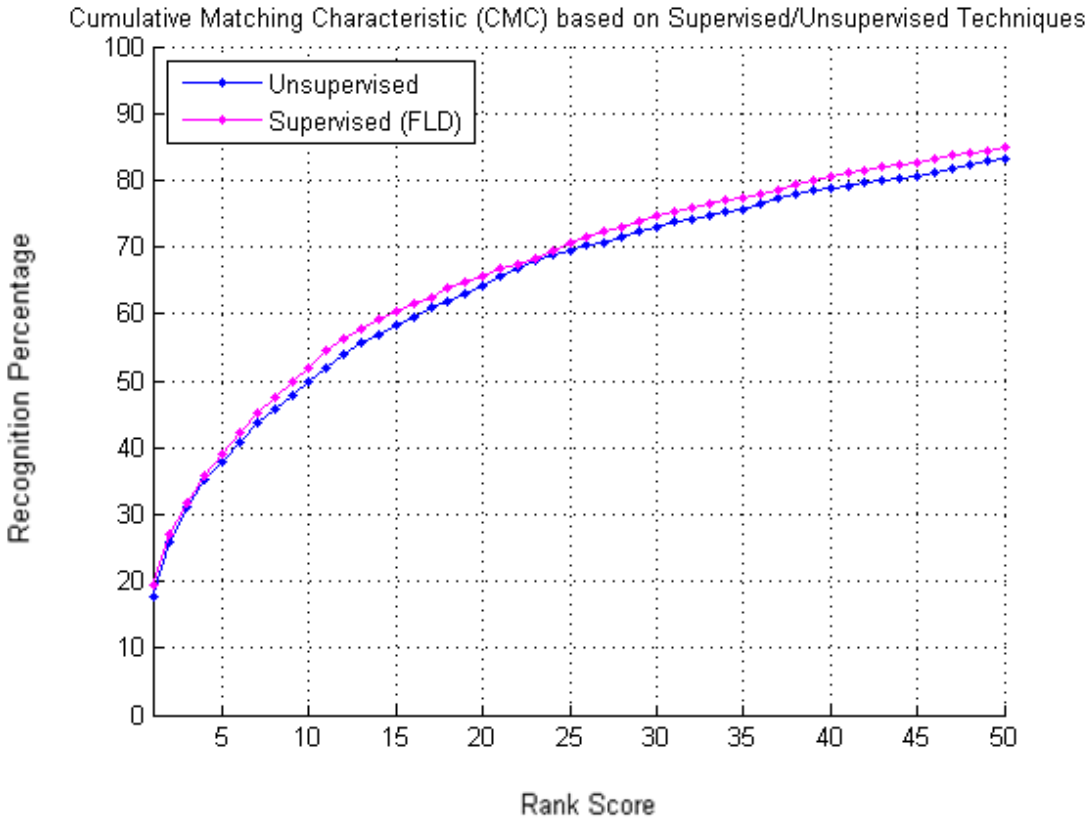


Figure 22: Effect of learning based technique in VIPeR dataset

The unsupervised result shows the Farenzena's original result, where the distance between the images calculated using 3 features with manually curated weights. In supervised result, since we want to find the coefficients of each distance matrix, FLD is trained with 2 different distance matrix, weighted Histogram and MSCR. We do not include Epitexture feature in FLD because it has no discriminative ability. The calculated weights of Histogram and MSCR are not the same as Farenzena et al. (2010) claim. They set both the weights of Histogram and MSCR to %40. From FLD, the weight of Histogram becomes %53 and MSCR becomes %47 which proves the results we got at the section 4.2.3 Feature Sets. From the Figure 22 and Table 6, we can say that learning based method improves the performance of the overall system which uses fixed parameters.

Table 6. AUC results obtained from Farenzena's original results versus FLD

Metric Used	AUC 100(%)	AUC 50(%)	AUC 20(%)	AUC 10(%)	AUC 1(%)
Unsupervised	91.89	64.86	47.90	37.37	19.21
Supervised (FLD)	**92.47**	**66.24**	**49.80**	**39.01**	**19.46**

4.2.6 Effects of localized histograms on horizontal segments

In section 4.2.1, we showed that dividing the body into parts improve systems ability to discriminate different people and correlate same people. In the Section 3.4, we argue that for better discrimination, we can separate the semantic body parts (torso and leg) into further horizontal segments. The Figure 23 shows the effect of localized histogram on further horizontal segments. When the body is divided into two semantic parts, this attribute is denoted with *histPart*, and if the body is divided into *K* total segments with further divisions, then the attribute is referred to as *histKseg.*

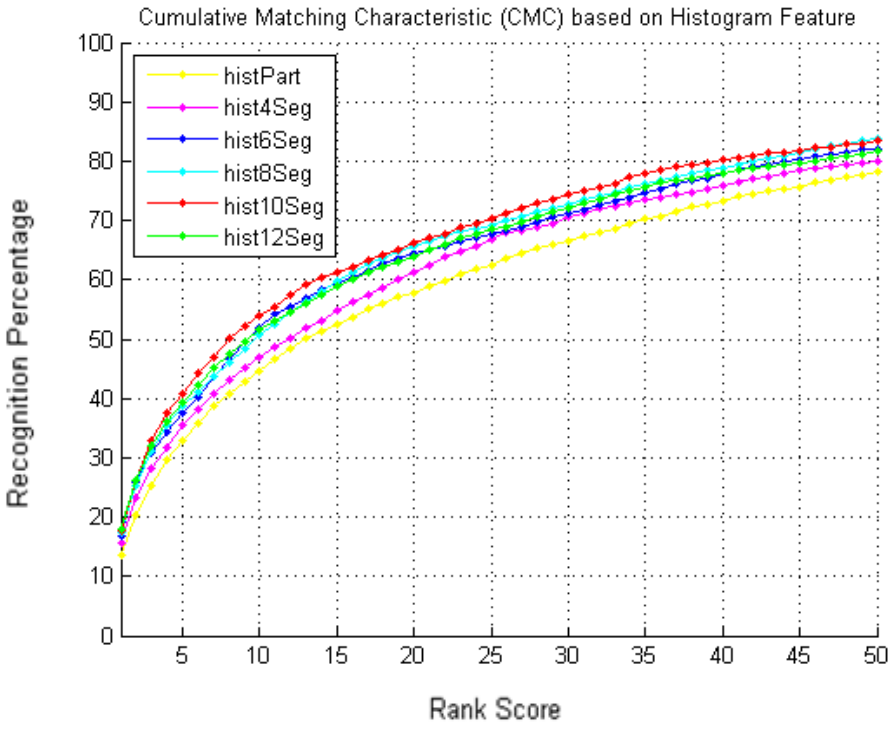


Figure 23: Effects of localized histograms on horizontal segments in VIPeR dataset

Figure shows that dividing torso and leg into further horizontal segments and using their histogram features to train FLD brings us better results than using only torso and leg histogram information. Best results are obtained when dividing torso and leg into 5 horizontal segments (hist10Seg). The curves show an increasing performance until we divide both torso and leg into 5 segments. After that, in hist12Seg, the results are getting worse, and further division cannot provide better performance except AUC %1. AUC records given in Table 7 consistently support our argument here.

Table 7. AUC results obtained from using different number of horizontal segments

histogram used	AUC 100(%)	AUC 50(%)	AUC 20(%)	AUC 10(%)	AUC 1(%)
histPart	89.79	58.96	42.67	32.51	13.51
*hist**4**Seg*	90.68	61.81	45.06	34.87	15.7
*hist**6**Seg*	91.50	64.17	48.72	37.74	16.87
*hist**8**Seg*	91.93	65.08	48.85	37.82	17.44
His10Seg	**92.11**	**66.46**	**50.85**	**40.24**	17.56
*His**12**Seg*	91.61	64.44	48.93	38.82	**18.07**

4.2.7 Comparing various machine learning techniques

Using FLD, we show that learning-based selection of distance metric parameters can improve the performance of person reidentification without changing the feature sets and their individual comparison metric. In this subsection, we compare FLD with a popular machine learning technique, SVM. In this experiment, the SVM is run with widely-used training parameters: an RBF kernel, a capacity value, C, of 0.1 and gamma value, G, of 0.1. The negative set is selected in a way that the numbers of positive and negative examples are balanced. In the next experiment, it will be shown that this a better choice for training SVM as opposed to training FLD, where all possible negative samples are fed into training stage. Indeed, we want to obtain a fair comparison setup by creating best possible environment for each machine learning algorithm. The details of these dataset partitioning strategies will be discussed in following sections together with experimental results supporting our arguments.

Figure 24 shows overall performances of FLD and SVM machine learning techniques on VIPeR using best feature sets. The CMC curve depicts that FLD outperforms SVM in this setup.

The AUC records (Table 8) indicates that SVM still performs better than the Farenzena's unguided version but does not provide as good results as FLD does. This superiority is more apparent in early hit performance (AUC10).

Table 8. AUC results obtained from FLD versus SVM

Method used	AUC 100(%)	AUC 50(%)	AUC 20(%)	AUC 10(%)	AUC 1(%)
Unsupervised	91.89	64.86	47.90	37.37	19.21
SVM	92.80	68.26	52.86	42.52	21.99
FLD	**93.78**	**70.47**	**54.45**	**43.73**	**22.31**

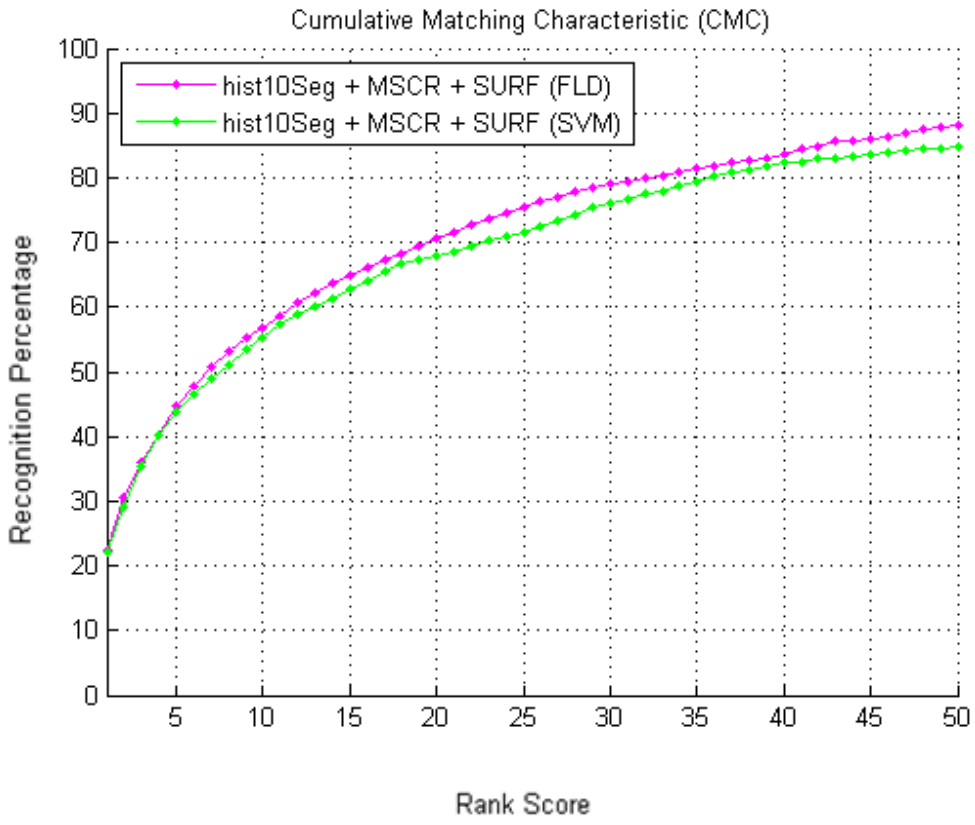


Figure 24: Comparison of FLD and SVM in VIPeR dataset

4.2.8 Effect of training set partitioning

In here, we show the effect of training set partitioning on two learning methods that we introduce in previous subsections, FLD and SVM.

Since the idea of FLD is based on the mean and scatter of the samples, we assume that increasing the size of training set even if the positive and negative sets are being unbalanced would be beneficial. Therefore, we anticipate that using all possible negative and positive samples will provide us the best results. The effects of other positive/negative partitioning ratios, 1/2, 1/10, 1/50, 1/100, 1/200 and 1/AllNegatives are shown in the Figure 25.

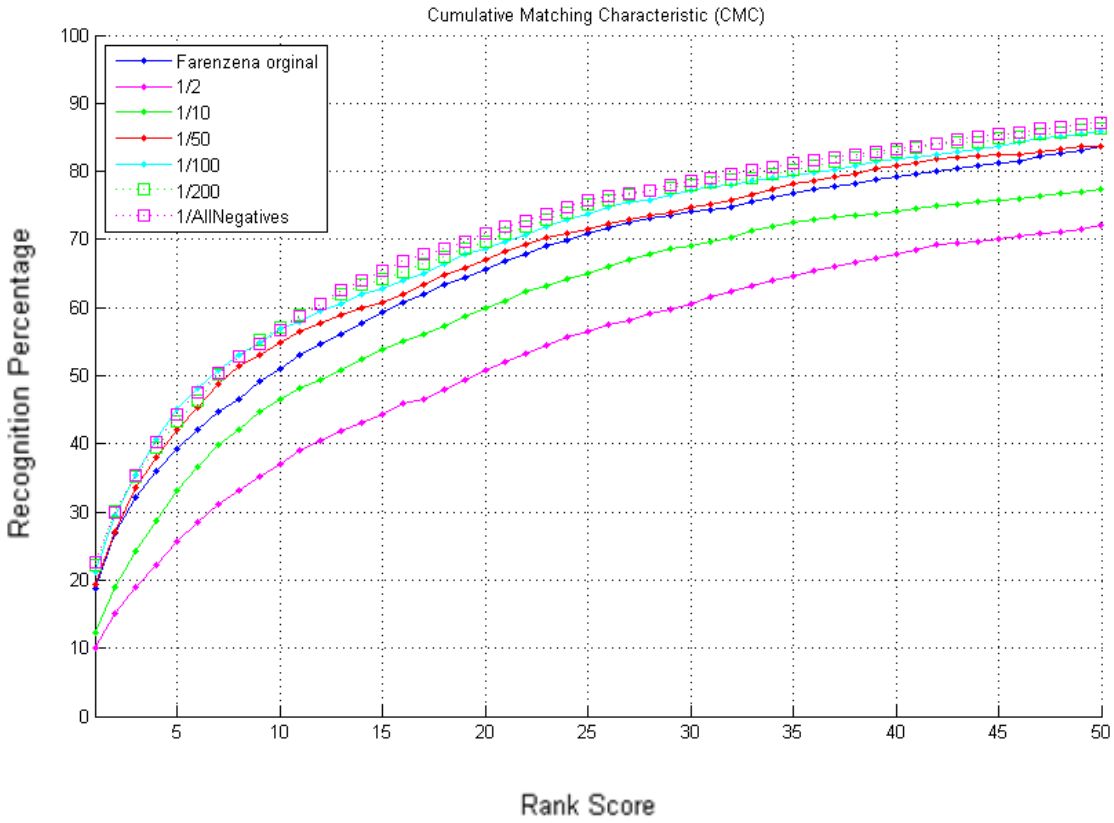


Figure 25: Effect of training set partitioning on FLD in VIPeR dataset

The results confirm our argument. The fewer negative samples we use the poorer results we get. The increase in the size of negative samples has a significant effect on the CMC curve (such as 1/2, 1/10, 1/50), when the larger ratios are used, the curve slightly converges to one that all negatives are used. This result obviously indicates the benefit of using as many as available negative samples in FLD training. Upon this remark, our final model is designed to comprise all negative samples when FLD is used.

Table 9. AUC results using different p/n training set ratio in FLD

p/n ratio FLD	AUC 100(%)	AUC 50(%)	AUC 20(%)	AUC 10(%)	AUC 1(%)
1/2	86.68	52.50	35.32	25.70	10.13
1/10	89.42	60.11	43.45	32.73	12.18
Farenzena	91.89	64.86	47.90	37.37	19.21
1/50	92.26	67.09	51.51	41.34	19.37
1/100	93.04	68.92	53.50	43.54	21.17
1/200	93.26	69.66	53.90	43.18	22.18
1/AllNegatives	**93.78**	**70.47**	**54.45**	**43.73**	**22.31**

In SVM, we do not use as many negatives as we do in FLD. As we discuss earlier (3.3.2), the imbalance in the number of positive and negative samples significantly changes the occurrence and position of support vectors, which can cause our model to be affected

adversely in terms of classification accuracy. Therefore, we just consider four cases of positive to negative ratios; 1/2, 1/4, 1/10 and 1/20.

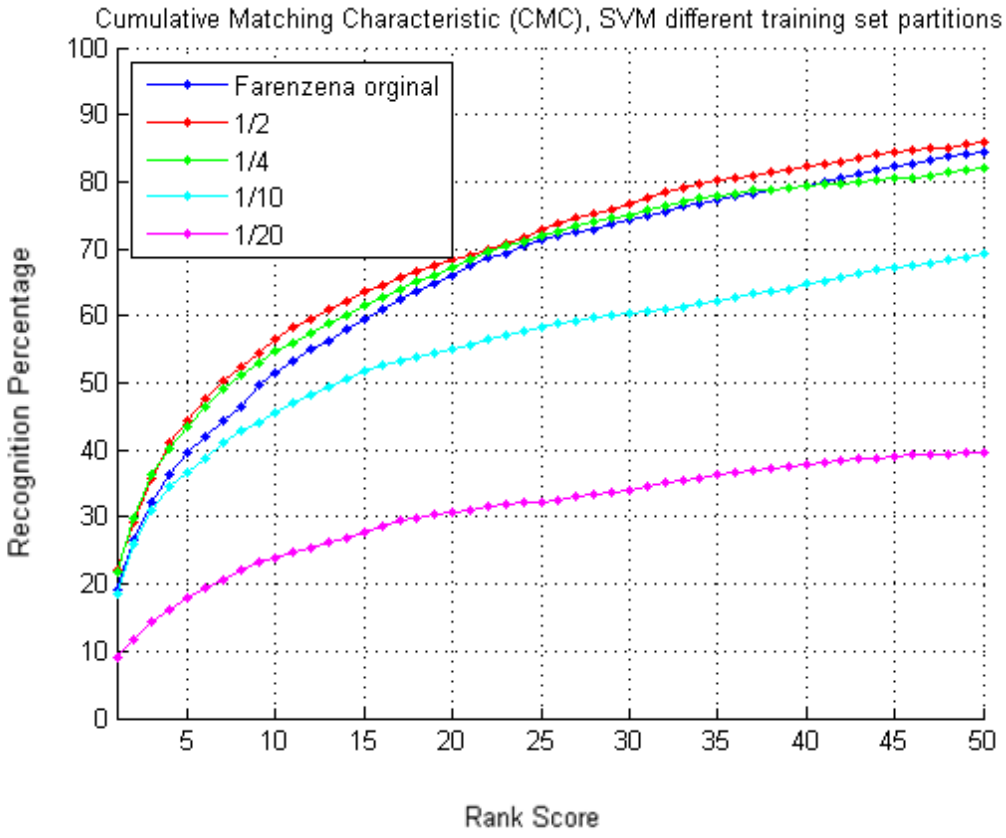


Figure 26: Effect of training set partitioning on SVM in VIPeR dataset

The CMC curve in Figure 26 and AUC records in Table 10 confirm our expectations about SVM accuracy. The best results with SVM is obtained when a 1/2 ratio is used between the sizes of positive and negative samples. The AUC values are adversely affected by the increase in the number of negative samples in comparison to positive samples. This result suggests that balanced number of positive and negative samples is needed in training SVM. But in any case, FLD performs better than SVM in this application.

Table 10. AUC results using different p/n training set ratio in SVM

p/n ratio SVM	AUC 100(%)	AUC 50(%)	AUC 20(%)	AUC 10(%)	AUC 1(%)
1/20	64.96	30.76	22.93	17.86	9.15
1/10	82.82	55.14	43.75	35.89	18.42
Farenzena	91.89	64.86	47.90	37.37	19.21
1/4	90.79	67.02	52.28	42.62	21.68
1/2	**92.96**	**68.91**	**53.52**	**43.32**	**22.03**

4.2.9 Comparison with previous methods

Based on the experiments until now, what we infer from the results can be summarized as follows:

- Weighted histograms can provide a more discriminative information than the unweighted ones.
- Best distance measure is the Bhattacharyya metric while comparing histograms
- Best feature combination is Histogram + MSCR + SURF
- Learning based metric can remarkably improve the results
- Five horizontal segments in torso and five in legs give the best discriminative ability.

An integrative model can then be introduced based on these observations. In this subsection, we experiment the integration of all these ideas on VIPeR and ETHZ datasets and compare our results with the best results previously reported in the literature on these datasets.

4.2.9.1 Results on VIPeR dataset

When compared to the best result reported in Farenzena et al. (2010), which was known as the current state-of-the-art, our method can result with a better CMC curve as shown in Figure 27. This result is consistent for all stages of the curve.

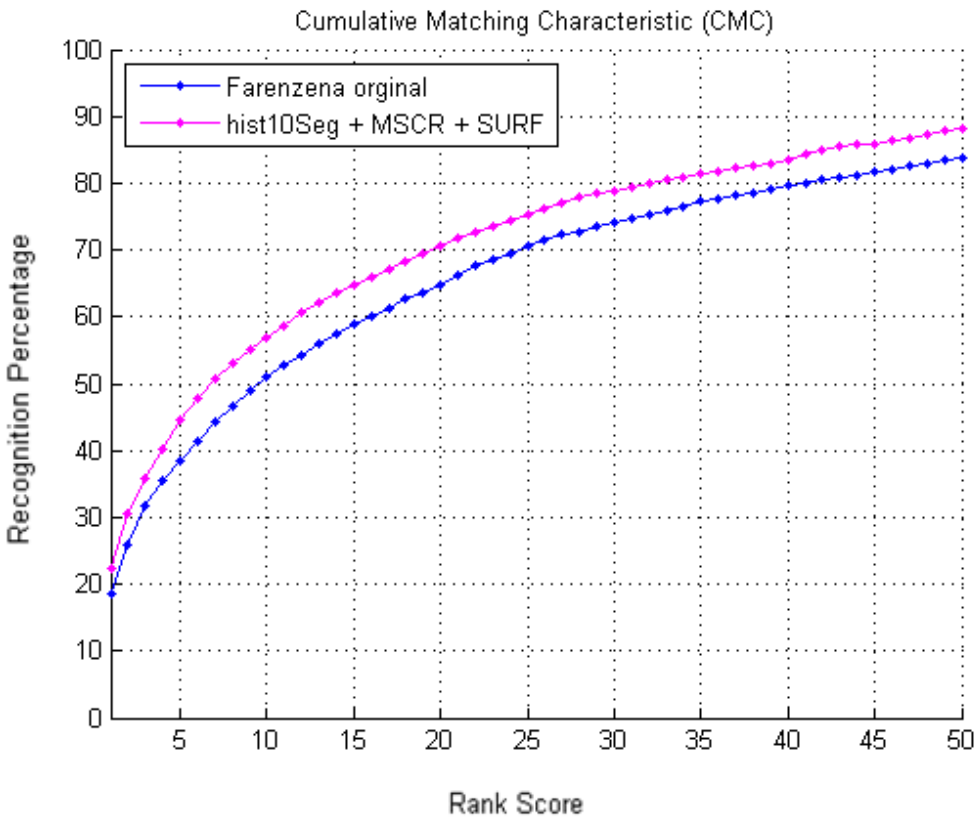


Figure 27: Results on VIPeR dataset

The superiority of the present method can also be observed in AUC analysis. Table 11 compares two methods in terms of AUC, AUC50, AUC20 and AUC10. It demonstrates that the present method is evidently more successful in identifying pedestrians in earlier hits (Note the difference in AUC10).

Table 11. AUC results from Farenzena's original version versus this study on VIPeR

Method	AUC 100(%)	AUC 50(%)	AUC 20(%)	AUC 10(%)	AUC 1(%)
Farenzena et al. (2010)	91.89	64.86	47.90	37.37	19.21
This study	**93.78**	**70.47**	**54.45**	**43.73**	**22.31**

Each row in Figure 28 shows first 10 matches of the input images found in VIPeR. From (a) to (c) the second camera view of the input images are found at first match. The retrieved results show the consistency between query and resulting images. Most of the matched images from (a) to (c) has the similar color distribution of their query image. For example in (a) the upper parts of the results mostly have orange/light brown and in leg parts the dominant color is grayish tones. In (b) the results with red torso and blue legs are found and in (c) the results are mostly the ones who wear sportswear. This tendency continues also in other queries not only in the images which found before 10^{th} match but also in images which cannot be found at earlier hits for example in (g) and (h). The top 10 ranking images are not the second camera view of the inputs in (g) and (h) but the first query returns the results who wears jeans and white shirt while the second query returns who has dark tones on their upper part and jeans. Some other matched and unmatched pairs can be found in Appendix A.

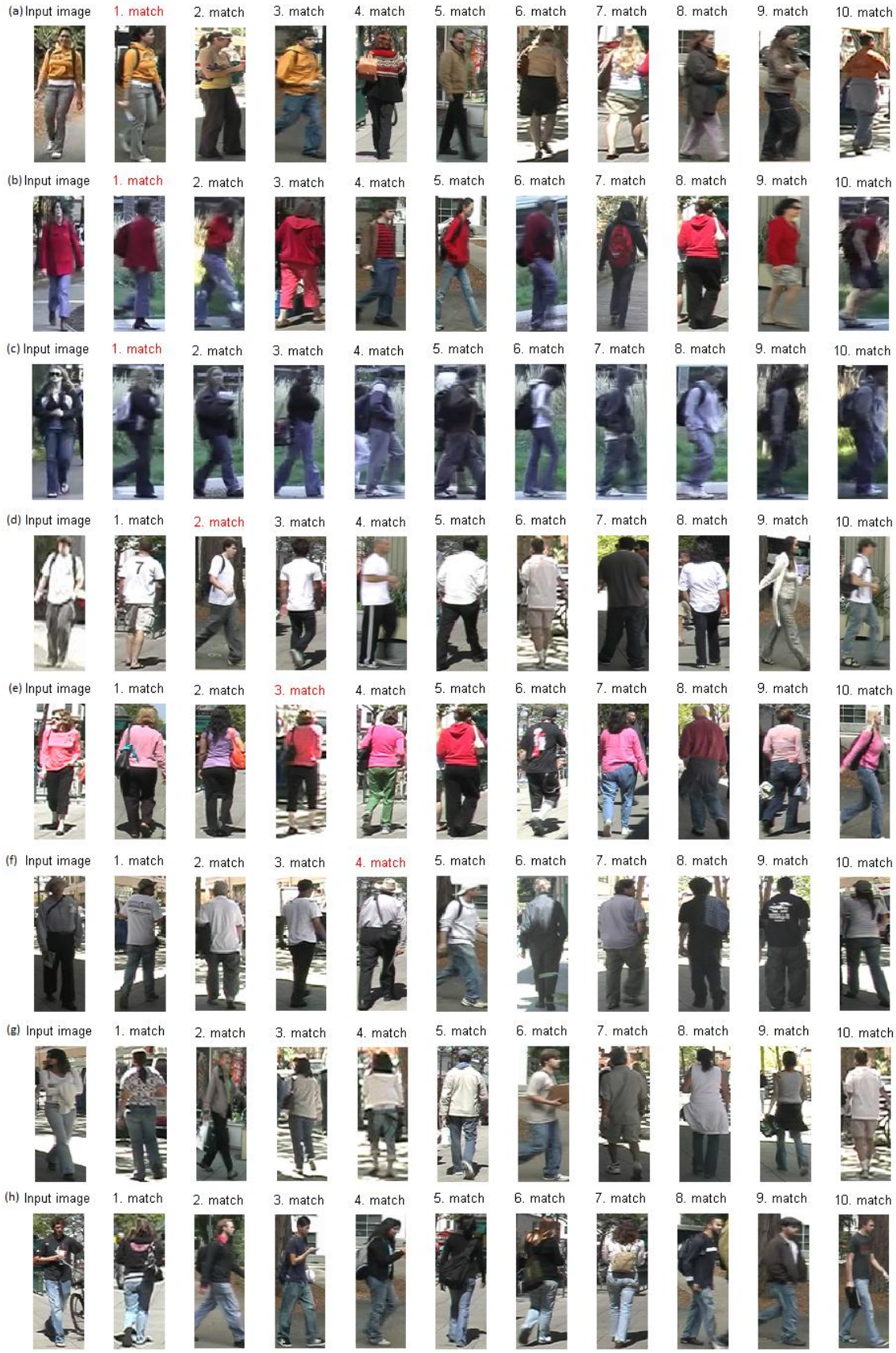


Figure 28. First 10 matches found in VIPeR

4.2.9.2 Results on ETHZ dataset

The results for both our implementation and single shot case of Farenzena et al. (2010) for sequence 1 on ETHZ are shown in the Figure 29. Since first sequence of ETHZ dataset contains 83 pedestrians, we just show first 10 hit rates of CMC curve in the figure.

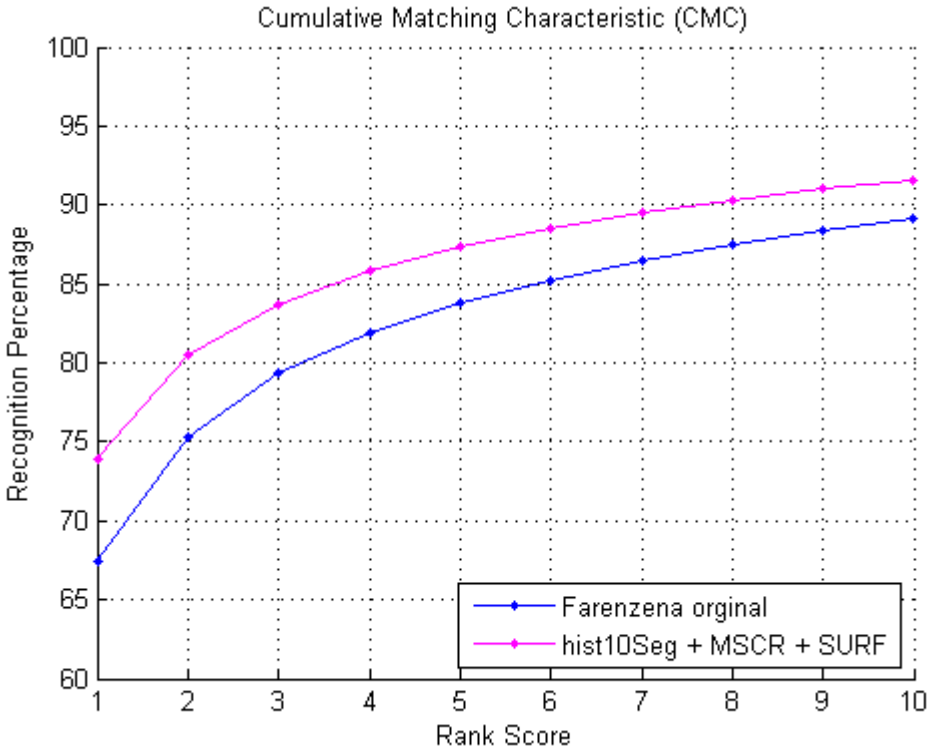


Figure 29: Results on ETHZ dataset

From the Figure 29 and Table 12, we can say that, our model gives not only better results on VIPeR but also in ETHZ dataset.

Table 12. AUC results from Farenzena's original version versus this study on ETHZ

Method	AUC 83(%)	AUC 10(%)	AUC 5(%)	AUC 1(%)
Farenzena et al. (2010)	95.61	82.45	77.55	67.46
This study	**96.48**	**86.21**	**82.23**	**73.90**

The best 10 matches for 4 input images in 1st sequence of ETHZ are shown in Figure 30. Similar to the results in Figure 28, in here the system also brings the persons who have similar clothing tendencies (in (b) the results are the ones who wear dark clothes while in (d) the results are most likely the ones who wear jean or dark pants and light coats). In 4.1.1 we said that the biggest problem in ETHZ is occlusion. In Figure 30 (c) we see that even though the occlusion appears in the input image, the system finds her match at first hit. Other matched samples can be found in Appendix B.

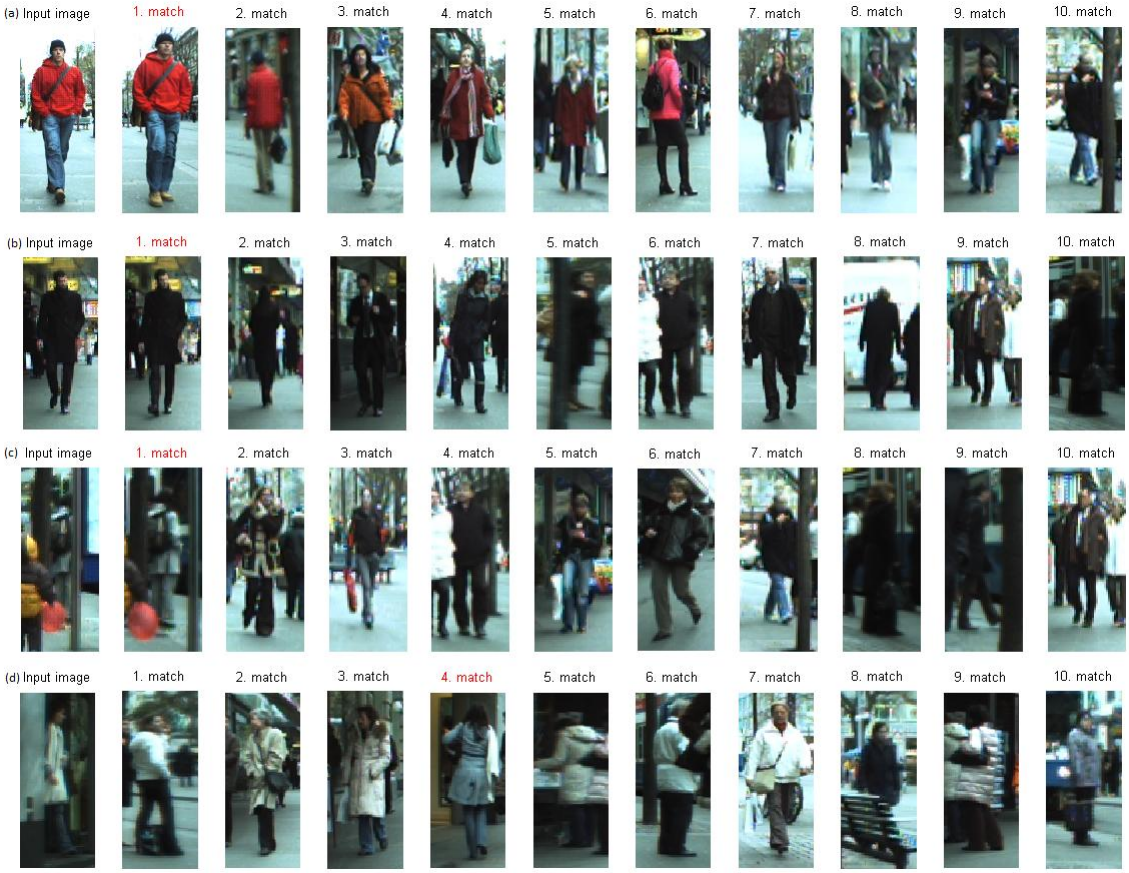


Figure 30. First 10 matches found in ETHZ 1st Seq.

In the training part of the results given above, all the images for each person, except the test figure of each person, are used. Because of the two reasons, the large number of the training examples and the small number of the pedestrians, we get better results than we did on VIPeR dataset. Therefore, we conduct other experiment on ETHZ dataset as if we have two images of each person taken from two disjoint cameras. To do so in a fair manner, we take each person's first and last frames, because these are the frames which have the most difference. The results are shown in Figure 31 and Table 13.

Table 13. AUC results on ETHZ using two images of each pedestrian

Method	AUC 83(%)	AUC 50(%)	AUC 20(%)	AUC 10(%)	AUC 1(%)
This study	84.85	75.99	57.62	45.56	23.27

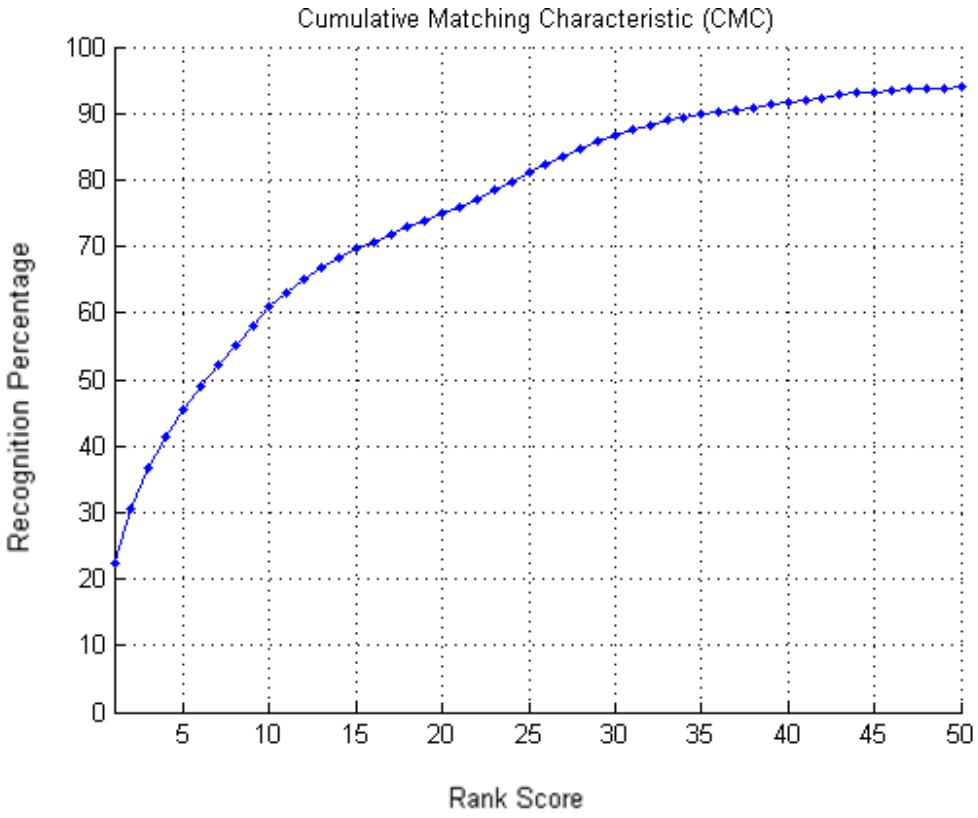


Figure 31. Results on ETHZ dataset using two images of each pedestrian

CHAPTER 5

CONCLUSION

In a multi-camera network system, if a pedestrian in a camera view is seen, finding his/her image in other camera views is called as person re-identification. In this thesis, we deal with this problem where only a single shot image of a pedestrian is available from each camera view. To overcome the limitations of existing methods, we here introduce an integrative and learning-based model for person reidentification.

The central points of our system lie in four main parts: background extraction, body part division, feature extraction and a learning based metric. For background extraction structuring element component analysis, which is known to be a successful method for foreground/background separation, is customized for pedestrian images as suggested by Farenzena et al. (2010).

Several observations based on the rigorous experiments on two common benchmark sets are reported to facilitate the future research in the field. We show that the color is the most significant information in detecting pedestrians in different cameras. Among several color representation schemes, HSV color space was already shown to be the most discriminative in different illumination conditions and view-invariant recognition. Here, we also demonstrate that using semantically localized color information can contribute to the result more than using a global color context. When we investigate the effect of body part division using HSV histogram, it is revealed that extracting the histogram values both on whole body and on torso and leg brings us to see the positive effect of the division. Additionally, using the vertical symmetry axis, we find out that the pixels near the symmetry axis are more important than others.

To compare the pedestrians, we need to calculate the distance between their HSV histograms. We compare Euclidian, Bhattacharyya, a well-known bin-to-bin histogram comparison metric, X^2, and finally a cross-bin metric, EMD. The experimental results reveal that the best comparison metric which can be used in HSV histogram match is Bhattacharyya metric.

We made a comparison of different feature sets as single attributes for person reidentification. The representatives of three common feature sets in computer vision, color features, texture features and interest point matches, are utilized to see their individual and combined effects on the prediction performance. According to the results, while the best distinguishable feature is weighted HSV histogram, MSCR, SiftFlow and SURF comes after that in a descending order.

We have shown that SURF matching can be useful when it is used in an integrative manner with other features, while it is not so successful when used alone. SURF is slightly better effect than the SIFT-Flow in integrated model, but their joint use does not enhance the final performance.

We also report that dividing semantically segmented body parts (torso and leg) into further horizontal segments can improve the prediction accuracy. Despite its simplicity, this idea is used for the first time in comparison of local color content for person reidentification, to our knowledge. We use this idea in our final integrative model and observe a significant improvement in the identification performance.

One of the contributions of this thesis is the introduction of a learning based metric which provides the integration of distinct feature sets to compare two pedestrian images. We have shown that a better identification performance can be achieved when the participation of these feature sets are guided by a supervised learning metric for final decision instead of using fixed participant weights for each.

In the aim of designing a powerful learning framework, we explore the performances of two widely used machine learning techniques; FLD and SVM. The experiments show that the learning based metric with either of these methods show superior performances to the unguided one. When compared to each other, FLD performs better than SVM. This may have two reasons. First, the mathematical formulation of FLD approach is inherently closer to the distance measure that we actually attempt to model for person matching in two images. In our application, we do not use directly the local image features but instead, we indirectly feed the learning system with the distances computed from these features. In mapping feature vectors into higher dimensions, SVM might be losing the actual information contained in the distance values. Second reason is due to the training set partitioning. Due to the nature of our datasets, the number of negative samples is higher than positive samples. It is well-known that the SVM becomes worse when the positive and negative sets are unbalanced. On the other hand, FLD algorithm is not affected by this imbalance since its learning parameters are not based on the individual samples but on the mean and scatter of the data matrix. Therefore, an increase in the number of samples enhances the performance of the FLD algorithm as opposed to the SVM approach. To sum up, the FLD algorithm can benefit from the larger training set in this application.

To assess the robustness and general applicability of the model, it has been tested on two benchmark datasets with different properties. Its high performance in both datasets encourages that the model can be successfully applied in distinct environments. We have demonstrated that our final model can improve upon the state-of-the-art on two challenging datasets.

Although the research has successfully been completed, there were some limitations. First, the head part contains small number of pixels. Farenzena et al. (2010) mentioned that this part carries very low information in discriminating two images since the color content does not change significantly between two people. Therefore the head is not used in our experiments. Secondly, due to the challenging characteristics of the datasets some feature sets do not show their best performances, for instance, low resolution is a drawback for interest point detectors.

It is anticipated that the problem of person reidentification will continue to receive the attention of researchers in the field. One of the major concerns will be the identification in practical applications. A potential problem in this respect is the occlusion of a person by another one in crowded areas. In this case, some other preprocessing techniques should be used to detect the boundaries of person body. Another problem is the low resolution of images when the camera is inserted in a certain distance to viewing area. This requires the use of other techniques for enhancing the images. New feature representation schemes are still needed to improve the overall performance.

REFERENCES

Arth, C., Leistner, C. and Bishof, H. (2007). Object reacquisition and tracking in large-scale smart camera networks. IEEE International Conference on Distributed Smart Cameras, Vienna, Austria.

Bauml, M., Stiefelhagen, R. (2011). Evaluation of local features for person reidentification in image sequences. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, Klagenfurt, Austria.

Bak, S., Corvee, E., Bremond, F., Thonnat, M. (2010a). Person re-identification using haar-based and dcd-based signature. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, Boston, USA.

Bak, S., Corvee, E., Bremond, F., Thonnat, M. (2010b) Person re-identification using spatial covariance regions of human body parts. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, Boston, USA.

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L. (2008). SURF: Speeded Up Robust Features. Computer Vision and Image Understanding, Vol. 110, pp. 346-359.

Belongie, S., Malik, J., Puzicha, J. (2002). Shape matching and object recognition using shape contexts. Pattern Analysis and Machine Intelligence, Vol. 24, pp. 509-522.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distribution. Bulletin of the Calcutta Mathematical Society, Vol. 35, pp. 99–110.

Brun, L., Conte, D., Foggia, P., Vento M. (2011). People reidentification by Graph Kernels methods. International Conference on Graph-based Representations in Pattern Recognition, Münster, Germany.

Cai, Y., Pietikainen, M. (2010). Person re-identification based on global color context. Int. Conf. on Computer vision, November 08-09, Queenstown, New Zealand.

Chaudhuri, G., Borwankar, J.D. and Rao, P. (1991). Bhattacharyya distance-based linear discriminant function for stationary time series. Communications In Statistics-Theory And Methods. Vol. 20, pp. 2195–2205.

Choi, E. (2003). Feature Extraction Based on the Bhattacharyya Distance. Pattern Recognition, Vol. 36, pp. 1703–1709.

Cula, O.G., Dana, K.J. (2004). 3D texture recognition using bidirectional feature histograms. International Journal of Computer Vision, Vol. 59, pp. 33-60.

Détection et ré-identification de piétons par points d'intérêt entre caméras disjointes by Hamdoun, Omar, Ph. D., l´Ecole Nationale Supérieure des Mines de Paris, 2010

Derpanis, K.G., "The Bhattacharyya Measure," available online at http://www.cse.yorku.ca/~kosta/CompVis_Notes/bhattacharyya.pdf , accessed February 2013.

Ess, A., Leibe, B. and Gool, L. V. (2007). Depth and Appearance for Mobile Scene Analysis. IEEE Int. Conf. on Computer Vision, Rio de Janeiro, Brazil.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M. (2010). Person reidentification by symmetry-driven accumulation of local features. IEEE Int. Conf. on Computer Vision and Pattern Recognition, San Francisco, USA.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, Vol. 7, pp. 179-188.

Forssen, P.E. (2007). Maximally stable colour regions for recognition and matching. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Minneapolis, USA.

Forssen, P., Lowe, D. (2007). Shape Descriptors for Maximally Stable Extremal Regions. International Conference on Computer Vision, Rio de Janeiro, Brazil.

Gandhi, T., and Trivedi, M. (2007). Person tracking and reidentification: Introducing Panoramic Appearance Map (PAM) for feature representation. Machine Vision and Applications, Vol. 18, pp. 207-220

Gheissari, N., Sebastian, T.B., Tu, P.H., Rittscher, J., Hartley, R. (2006). Person reidentification using spatiotemporal appearance. IEEE Int. Conf. on Computer Vision and Pattern Recognition, New York, USA.

Gray, D., Brennan, S., Tao, H. (2007). Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. IEEE Int. Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro, Brazil.

Gray, D., Tao, H. (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. Europ. Conf. on Computer Vision, Marseille, France.

Guggenberger, A. (2008). Another Introduction to Support Vector Machines, available online at mindthegap.googlecode.com/files/AnotherIntroductionSVM.pdf, accessed February 2013.

Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B. (2008). Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. ACM/IEEE Int. Conf. on Distributed Smart Cameras, Stanford, CA, USA.

Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H. (2011). Person re-identication by descriptive and discriminative classification. Scandinavian Conference on Image Analysis, Ystad Saltsjöbad, Sweden.

Huang, T. and Russell, S. (1998). Object identification: A Bayesian analysis with application to traffic surveillance. Artificial Intelligence, Vol. 103, pp.77 -93.

Javed, O., Shafique, K., Rasheed, Z. and Shah, M. (2008). Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. Computer Vision and Image Understanding, Vol. 109, pp. 146 –162.

Jojic, N., Perina, A., Cristani, M., Murino, V., Frey, B. (2009). Stel component analysis: Modeling spatial correlations in image class structure. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Florida, USA.

Kailath, T. (1967). “The Divergence and Bhattacharyya Distance Measures in Signal Selection,” IEEE Trans. Comm. Technology, Vol. 15, pp. 52-60.

Ke, K., Zhao, T., Li, O. (2010). Bhattacharyya distance for blind image steganalysis. International Conference on Multimedia Information Networking and Security, Nanjing, China.

Kogut, G.T., Trivedi, M. A. (2001). Maintaining the identity of multiple vehicles as they travel through a video network. IEEE Int. Conf. on Intelligent Transportation Systems, Oakland, USA

Lantagne, M., Parizeau, M. and Bergevin, R. (2003). VIP : Vision tool for comparing images of people. IEEE Conference on Vision Interface, Halifax, Canada.

Ling, H., Jacobs, D. (2007). Shape classification using the inner-distance. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 29, pp. 286-299.

Ling, H., Okada, K. (2007). An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 29, pp. 840-853.

Liu, C., Yuen, J., Torralba, A., Sivic, J., & Freeman, W. (2008). SIFT flow: dense correspondence across difference scenes. European conference on computer vision, Marseille, France.

Liu, C., Yuen, J., Torralba, A. (2011). SIFT Flow: Dense Correspondence across Scenes and Its Applications. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 33, pp. 978–994.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints, cascade filtering approach. International Journal of Computer Vision, Vol. 60, pp. 91 – 110.

Mak, B., Barnard, E. (1996). Phone clustering using the Bhattacharyya distance. Int. Conf. Spoken Language Processing, Philadelphia, USA.

Martin, D., Fowlkes, C., Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.26, pp. 530 – 549.

Martinez, A., Kak, A. (2001). PCA versus LDA. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 23, pp. 229-233.

Matas, J., Chum, O., Urban, M., Pajdla, T. (2002). Robust wide baseline stereo from maximally stable extremal regions. British Machine Vision Conference, Cardiff, UK.

Noble, W.S. (2006). What is a support vector machine? Nature Biotechnology, Vol. 24, pp.1565-1567.

Oliveira, I., Luiz, J. (2009). People re-identification in a camera network. IEEE Int. Conf. on Dependable, Autonomic and Secure Computing, Chengdu, China.

Orazio, T. D., Mazzeo, P. L. and Spagnolo, P. (2009). Color brightness transfer function evaluation for non-overlapping multi camera tracking. International Conference on Distributed Cameras, Como, Italy.

Pham, T., Worring, M. and Smeulders, A. (2007). A multi-camera visual surveillance system for tracking re-occurrences of people. International Conference on Distributed Smart Cameras, Vienna, Austria.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, Series 5 50 (302): 157–175.

Pele, O., Werman, M. (2010). The quadratic-chi histogram distance family. European Conference on Computer Vision, Crete, Greece.

Plataniotis, K. N., Venetsanopoulos A. N. (2000). Color Image Processing and Applications. Berlin, Germany: Springer-Verlag.

Prosser, B., Zheng, W.S., Gong, S., Xiang, T. (2010). Person re-identification by support vector ranking. British Machine Vision Conf., Aberystwyth, UK.

Rubner, Y., Tomasi, C., Guibas, L. (2000). The Earth Mover's Distance as a Metric for Image Retrieval. International Journal of Computer Vision, Vol: 40, pp.99–121

Rubner, Y., Guibas, L. J., Tomasi, C. (1997). The earth mover's distance, multidimensional scaling, and color-based image retrieval. DARPA Image Understanding Workshop.

Schwartz, W., Davis, L. (2009). Learning discriminative appearance-based models using partial least squares. Braz. Sym. on Comput. Graphics and Image Proc., Brazil.

Sharif, Md. H., Uyaver, S., Djeraba, C. (2010). "Crowd Behavior Surveillance Using Bhattacharyya Distance Metric", Lecture Notes in Computer Science, Springer, 311-323.

Sreekanth, V., Vedaldi, A., Jawahar, C. V., Zisserman, A. (2010). Generalized rbf feature maps for efficient detection. British Machine Vision Conference, Wales, UK.

Xuan, G., Zhu, X., Chai, P., Shi, Y., Fu, D. (2006). Feature Selection Based on the Bhattacharyya Distance. IEEE International Conference on Pattern Recognition, Hong Kong, China.

Trivedi, M. M., Gandhi, T. L. and Huang K. S. (2005). Distributed interactive video arrays for event capture and enhanced situational awareness. IEEE Intelligent Systems, Vol. 20, pp. 58-66

Varma, M., Zisserman, A. (2009). A statistical approach to material classification using image patch exemplars. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.31, pp. 2032 - 2047.

Vert, J., Tsuda, K., Scholkopf, B. (2004). A primer on kernel methods, Cambridge, Massachusetts: MIT Press.

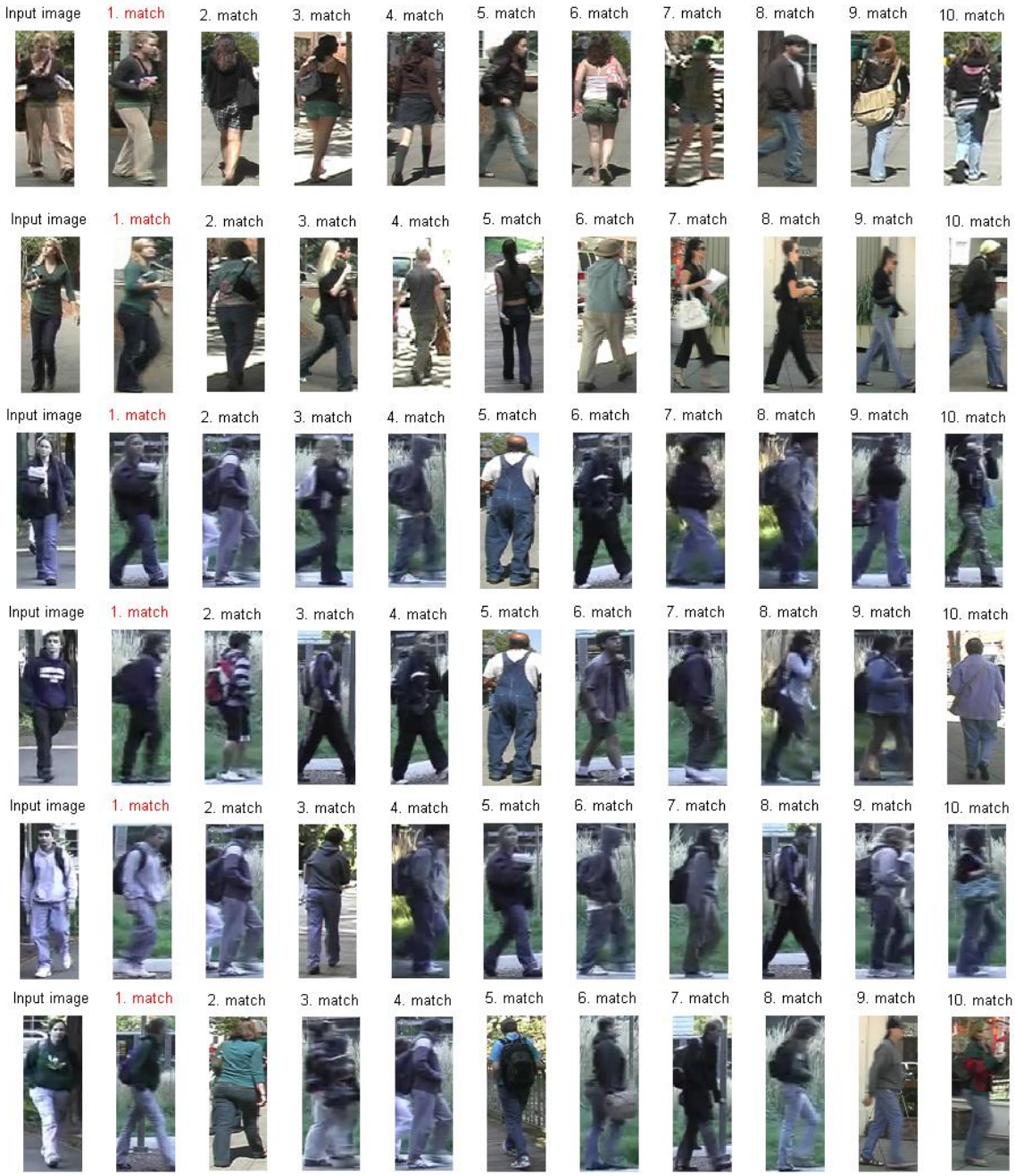
You, C.H., Lee, K.A., Li, H. (2010). GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition. IEEE Trans. Audio, Speech and Language Processing, Vol. 18, pp.1300–1312.

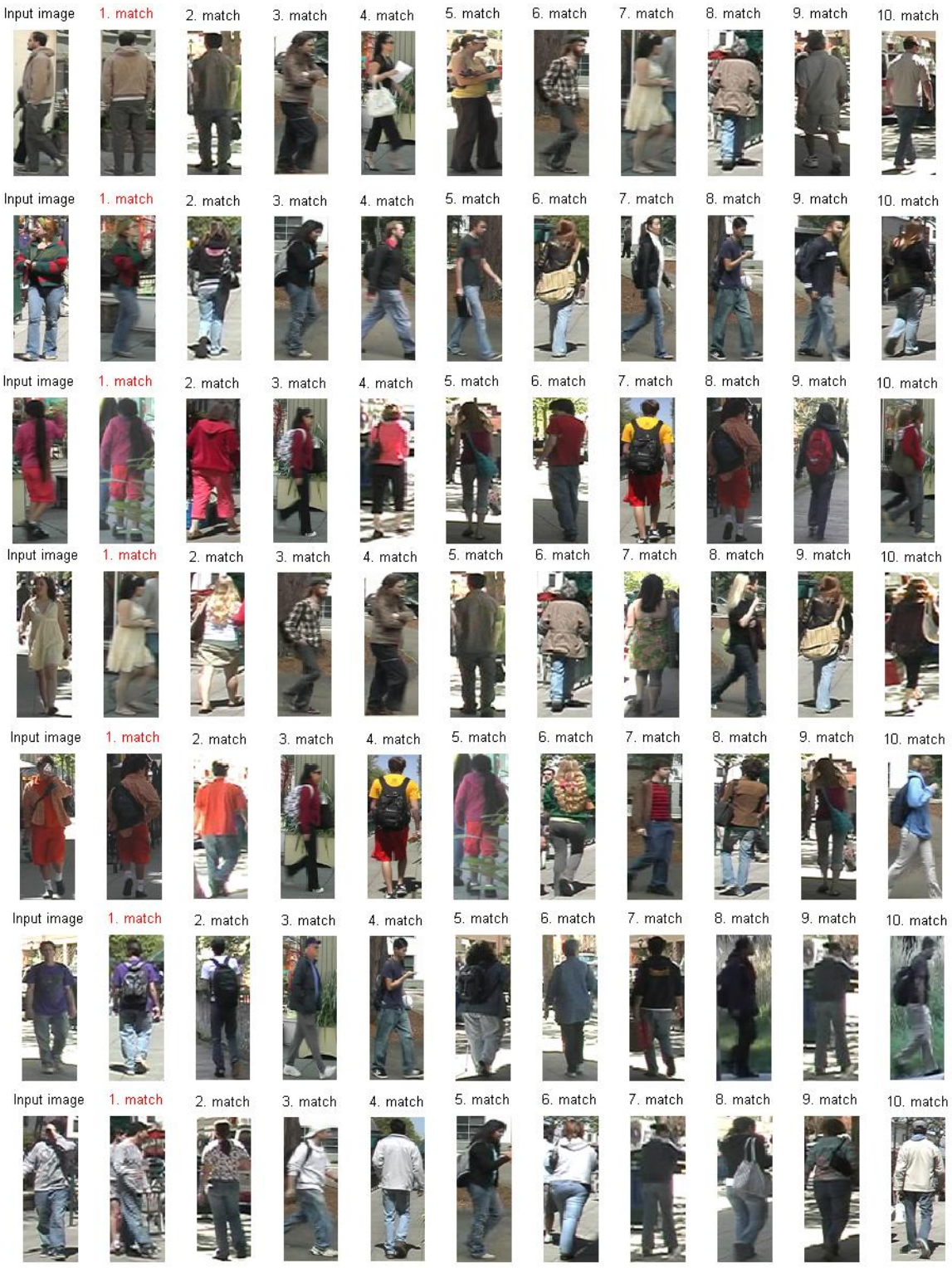
Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision, Vol. 73, pp. 213-238.

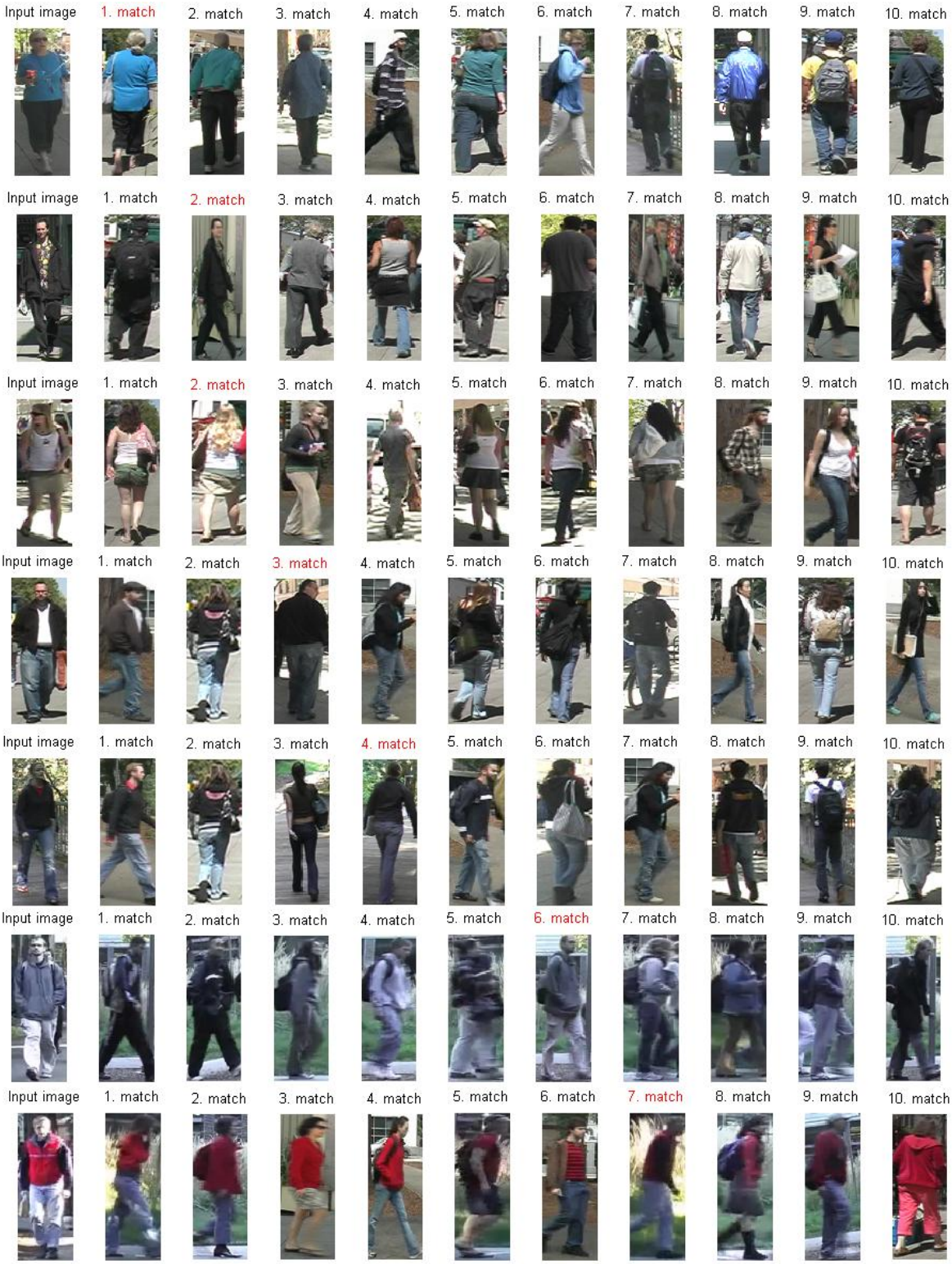
Zheng, W.S., Gong, S., Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA.

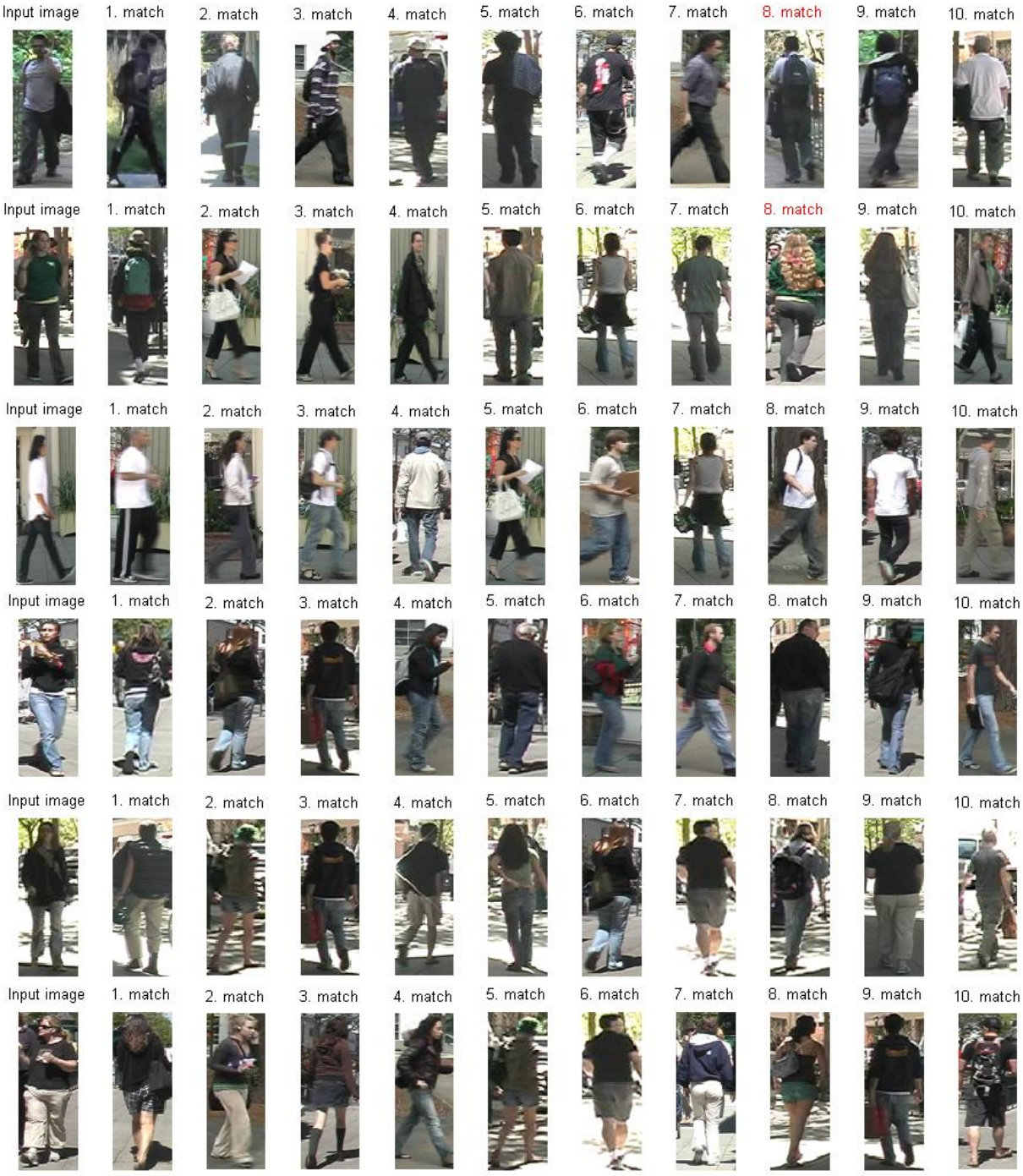
APPENDICES

Appendix A. The top 10 ranking images for some of the images in VIPeR

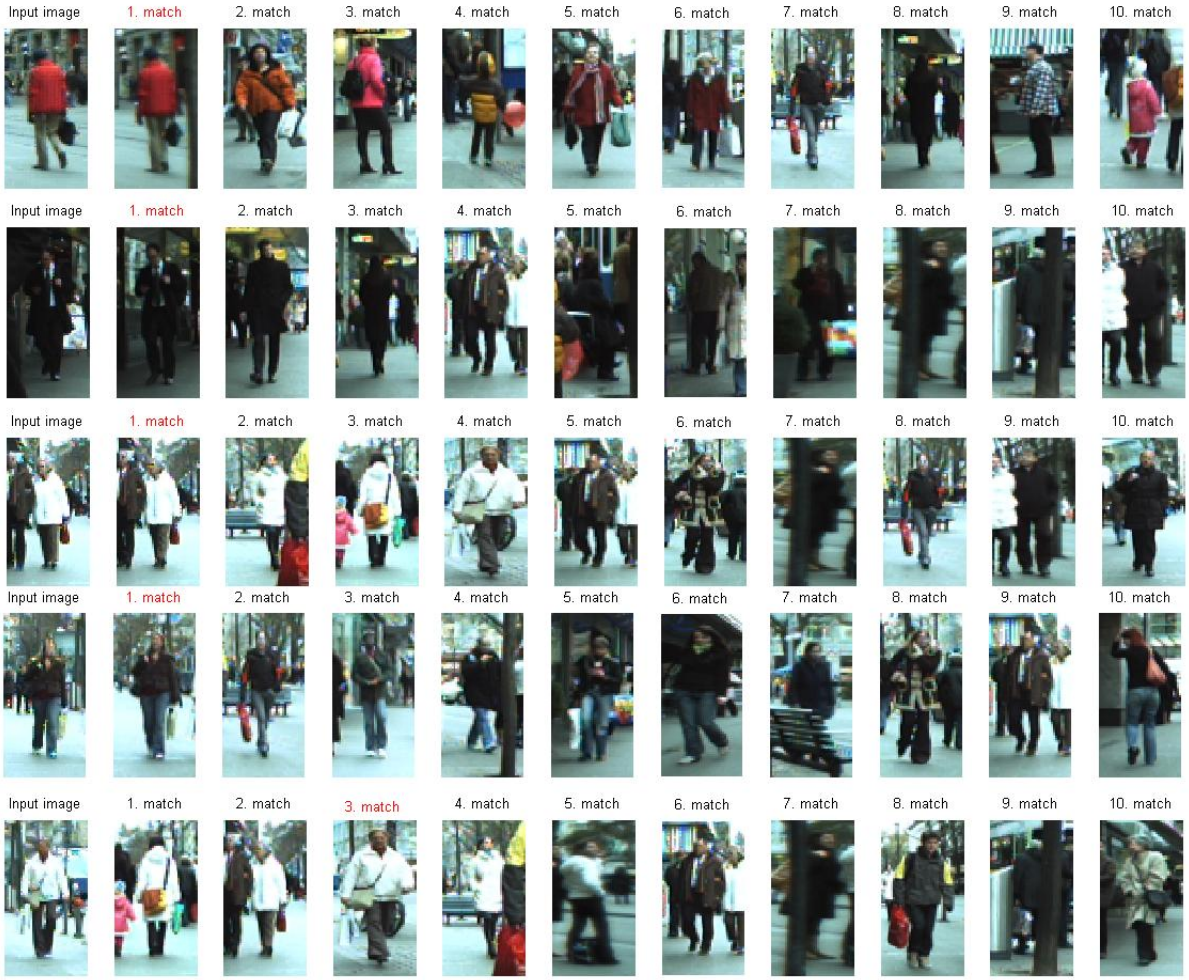



Input image
1. match
2. match
3. match
4. match
5. match
6. match
7. match
8. match
9. match
10. match
Input image
1. match
2. match
3. match
4. match
5. match
6. match
7. match
8. match
9. match
10. match
Input image
1. match
2. match
3. match
4. match
5. match
6. match
7. match
8. match
9. match
10. match
Input image
1. match
2. match
3. match
4. match
5. match
6. match
7. match
8. match
9. match
10. match
Input image
1. match
2. match
3. match
4. match
5. match
6. match
7. match
8. match
9. match
10. match
Input image
1. match
2. match
3. match
4. match
5. match
6. match
7. match
8. match
9. match
10. match
Input image
1. match
2. match
3. match
4. match
5. match
6. match
7. match
8. match
9. match
10. match


Input image 1. match 2. match 3. match 4. match 5. match 6. match 7. match 8. match 9. match 10. match
Input image 1. match 2. match 3. match 4. match 5. match 6. match 7. match 8. match 9. match 10. match
Input image 1. match 2. match 3. match 4. match 5. match 6. match 7. match 8. match 9. match 10. match
Input image 1. match 2. match 3. match 4. match 5. match 6. match 7. match 8. match 9. match 10. match
Input image 1. match 2. match 3. match 4. match 5. match 6. match 7. match 8. match 9. match 10. match
Input image 1. match 2. match 3. match 4. match 5. match 6. match 7. match 8. match 9. match 10. match
Input image 1. match 2. match 3. match 4. match 5. match 6. match 7. match 8. match 9. match 10. match


Input image
1. match
2. match
3. match
4. match
5. match
6. match
7. match
8. match
9. match
10. match
Input image
1. match
2. match
3. match
4. match
5. match
6. match
7. match
8. match
9. match
10. match
Input image
1. match
2. match
3. match
4. match
5. match
6. match
7. match
8. match
9. match
10. match
Input image
1. match
2. match
3. match
4. match
5. match
6. match
7. match
8. match
9. match
10. match
Input image
1. match
2. match
3. match
4. match
5. match
6. match
7. match
8. match
9. match
10. match
Input image
1. match
2. match
3. match
4. match
5. match
6. match
7. match
8. match
9. match
10. match

Appendix B. The top 10 ranking images for some of the images in 1st sequence of ETHZ



TEZ FOTOKOPİSİ İZİN FORMU

ENSTİTÜ

Fen Bilimleri Enstitüsü ☐

Sosyal Bilimler Enstitüsü ☐

Uygulamalı Matematik Enstitüsü ☐

Enformatik Enstitüsü ☑

Deniz Bilimleri Enstitüsü ☐

YAZARIN

Soyadı : OĞUL ……
Adı : Burçin Buket ……...
Bölümü : Bilişim Sistemleri ……....

TEZİN ADI (İngilizce) :
A LEARNING-BASED METHOD FOR PERSON RE-IDENTIFICATION.
……....……..…
……....……..…
……....……..…

TEZİN TÜRÜ : Yüksek Lisans ☑ Doktora ☐

1. Tezimin tamamından kaynak gösterilmek şartıyla fotokopi alınabilir. ☐
2. Tezimin içindekiler sayfası, özet, indeks sayfalarından ve/veya bir bölümünden kaynak gösterilmek şartıyla fotokopi alınabilir. ☐
3. Tezimden bir (1) yıl süreyle fotokopi alınamaz. ☑

TEZİN KÜTÜPHANEYE TESLİM TARİHİ : ……