# An Human Perceptive Model for Person Re-identification

Angelo Cardellicchio[1], Tiziana D'Orazio[1], Tiziano Politi[2] and Vito Renò[1]
[1]*National Research Council, Institute of Intelligent Systems for Automation, Bari, Italia*
[2]*Politecnico di Bari, Bari, Italia*

Keywords: Color Analysis, Feature Extraction, Histograms.

Abstract: Person re-identification has increasingly become an interesting task in the computer vision field, especially after the well known terroristic attacks on the World Trade Center in 2001. Even if video surveillance systems exist since the early 1950s, the third generation of such systems is a relatively modern topic and refers to systems formed by multiple fixed or mobile cameras - geographically referenced or not - whose information have to be handled and processed by an intelligent system. In the last decade, researchers are focusing their attention on the person re-identification task because computers (and so video surveillance systems) can handle a huge amount of data reducing the time complexity of the algorithms. Moreover, some well known image processing techniques - i.e. background subtraction - can be embedded directly on cameras, giving modularity and flexibility to the whole system. The aim of this work is to present an appearance-based method for person re-identification that models the chromatic relationship between both different frames and different areas of the *same* frame. This approach has been tested against two public benchmark datasets (*ViPER* and *ETHZ*) and the experiments demonstrate that the person re-identification processing by means of intra frame relationships is robust and shows great results in terms of recognition percentage.

## 1 INTRODUCTION

Since the September 11 attacks, the increased need of security has led to the development of new techniques to *identify* and *prevent* security threats in crowded environments. One of the most studied problems is *person re-identification* (*PRID*), i.e. the task of identify the same individual given a set of frames taken from two or more non-overlapping cameras covering the same environment.

Each computer vision system that performs video surveillance tasks must deal with challenging and critical aspects due to:

- *noisy signals provided by the cameras*;
- *illumination and viewpoint variations*;
- *background clutter with occlusion phenomenas*;
- *low image quality*.

Ideally, given an initial dataset divided in a *probe set* and a *gallery set*, the *PRID* pipeline begins with a feature-extraction step, where a *robust* set of features is extracted from each frame and combined in a *signature vector* that should characterize the observation avoiding ambiguities (i.e. a person with a blue shirt and white trousers and another person with a red blouse and pink skirt should have different signatures). Finally, each signature of the probe set has to be compared with the others of the gallery set in order to find the best match.

There are several approaches used to do this comparison, but they can be divided in two macro categories:

**Unsupervised Approaches** use a fixed metric to compare the signatures extracted from different frames.

**Supervised Approaches** learn a metric matrix $M$ using an optimization criterion which maximizes the distance between different signatures in order to have isolated clusters in the vector space of the features.

In literature, the most used types of features are *texture-based* and *color-based* ones. Some approaches like (Farenzena et al., 2010) combine both types of features to extract the signature, while others ((Yang et al., 2014),(Matsukawa et al., 2014)) use only color-based ones, which have been proven to be more effective when the image resolution is low.

Finally, the effectiveness of the *PRID* pipeline can be evaluated in the *single-shot* or in the *multiple-shot*


Cardellicchio A., D'Orazio T., Politi T. and Renò V..
An Human Perceptive Model for Person Re-identification.
DOI: 10.5220/0005341906380643
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 638-643
ISBN: 978-989-758-089-5

case. In the first case, only two frames per each subject are given: one in the gallery set and one in the probe set; in the other case, multiple frames per each individual are given, so each frame of the probe set can be tested against several frames of the gallery set. Generally speaking, the single-shot case is more challenging than the multi-shot one because the algorithms must be trained with only one sample per person.

The focus of this work is to build an *human perceptive model* as a base for a video surveillance system. As a consequence, we need to focus on the way humans perform PRID, i.e. using *perceptive* information instead than mere numerical one.

Intuitively, humans use both *short-term* and *long-term* biometric features to perform PRID. While the first kind of features refere to traits which can change in short periods of time, like clothes, features of the second kind usually don't change in the whole lifetime, like fingerprints or retina. Given the low quality of images acquired by CCTV cameras and the requirement for a non-pervading PRID system, long-term features cannot be currently used in a proper way; instead, short-term features may be used for PRID, as they will be likely retained between different views of the same individual taken in near temporal spots.

Two different requirements have driven the development of the proposed methodology: *improve qualitative PRID performances* and *lower PRID algorithm computational cost*.

As for the first requirement, actual PRID methods focus primarily on *quantitative* performance, i.e. the maximization of the number of correctly matched frames, which can be statistically characterized by the first rank of a CMC curve. While this approach is theoretically correct (i.e. if a method gets a good value of the first rank of the CMC, *it works*), a change of perspective may be necessary. In fact, given a PRID dataset, it can be shown that some frame sets are *ambiguous*, i.e. the correct match is hard to be identified even from an human operator. An example of an ambiguous frame set is given in figure 1(a), while in figure 1(b) an example of non-ambiguous frame set is shown: ideally, a good PRID algorithm should never fail in the second situation.

(a) Ambiguous frameset

(b) Non-ambiguous frameset

Figure 1: Ambiguity of different frame subsets.

The second requirement implies that we should look for a (relatively) unexpensive set of feature upon which build our frame signature, as PRID systems need to run in real time in order to be effective.

Following these consideration, we model the way humans perform PRID characterizing each couple of frame by means of both chromatic *inter-relationships*, which refer to similarity between a couple of different frames, and *intra-relationships*, which refer to similarity between different areas of the same frame.

A similar approach has been proposed in (Kviatkovsky et al., 2013), where the authors divide each frame in two parts (upper and lower) and characterize color point clouds from each part by means of *shape context descriptors* (Belongie et al., 2002), whose shapes are supposed to be retained between different views of the same individual. Unfortunately, this approach doesn't analyze possible dependecies between different parts of the same image.

In order to fill this gap and compute the intra-relationships, we firstly divide each frame in $n$ different horizontal stripes, like (Yang et al., 2014), (Truong Cong et al., 2010); then, we extract several color-based features using *color histograms* (Swain and Ballard, 1992), which have been proven to be robust to pose and illumination changes; lastly, we compare obtained signature in order to find the best match. The division in $n$ different horizontal stripes allows to retain the spatial information which is not enfolded in color histograms.

The rest of the work is organized as follows: in the second section an introduction on color models is given, then in the third section the proposed methodology is exposed; the fourth section contains experiments and results carried out on two public datasets (*ViPER* and *ETHZ*). In the fifth and last section we discuss the conclusions and gives an overview on future works and perspectives.

# 2 METHODOLOGY

## 2.1 Modeling Chromatic Content

### 2.1.1 Diagonal-offset Model

We first introduce the *diagonal-offset model* (Finlayson et al., 2005), which can be used to map colors under illumination conditions $i$ to the corresponding colors under illumination conditions $f$ through a linear model:

$$\begin{pmatrix} R^f \\ G^f \\ B^f \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} * \begin{pmatrix} R^i \\ G^i \\ B^i \end{pmatrix} + \begin{pmatrix} o^r \\ o^g \\ o^b \end{pmatrix}$$

where $a$, $b$, and $c$ are the first degree terms and $o^r$, $o^g$ and $o^b$ the offsets. Various types of illumination changes can be defined exploiting this model: specifically, in (Van de Sande et al., 2010) the authors introduce:

**Light Intensity Changes** i.e. changes with a constant value of light-mapping coefficents $a = b = c = k$ and a null offset vector;

**Light Intensity Shifts** i.e. changes with a constant value of light-mapping coefficents $a = b = c = 1$ and a constant value of offset vector coefficents $o^r = o^g = o^b = h$;

**Light Intensity Changes and Shifts** i.e. changes with a constant value of light-mapping coefficents $a = b = c = k$ and a constant value of offset vector coefficents $o^g = o^g = o^b = h$;

**Light Color Changes** i.e. changes with a non-constant value of light-mapping coefficents $a \neq b \neq c$ and a null offset vector;

**Light Color Changes and Shifts** i.e. changes with a non-constant value of light-mapping coefficents $a \neq b \neq c$ and a non-constant value of light-mapping coefficents $o^r \neq o^g \neq o^b$.

The diagonal-offset model is useful to characterize chromatic variations due to different image acquisition settings and evaluate the robustness of each color model described in the next paragraph.

#### 2.1.2 Color Models

In our work, we used some of the most significant color models known in literature.

**rgb** is obtained normalizing the RGB color model. Its main advantage is it is invariant to light intensity changes.

**HSV** color model describes colors in terms of hue, saturation and value of illuminant intensity. We discard intensity information, considering only $H$ and $S$ to improve the robustness to light changes.

**Log-Chromaticity Color Space** is defined in (Gevers and Smeulders, 1999) as:

$$\chi_1 = ln\left(\frac{R}{G}\right), \chi_2 = ln\left(\frac{B}{G}\right)$$

In (Kviatkovsky et al., 2013) it is proven this model is invariant to illumination intensity and color changes and shifts.

We now give an exhaustive overview of our methodology.

## 2.2 Modeling Chromatic Relationships

Among all chromatic features, *color histograms* are the most widely used in PRID as they are invariant to variations in pose and view angle ((Farenzena et al., 2010), (Yang et al., 2014), (Matsukawa et al., 2014), (Truong Cong et al., 2010)). Unfortunately, they bring two disadvantages:

1. loss of *shape* information, i.e. information about texture or body parts shape;
2. loss of *spatial* information, i.e. information about the spatial disposition of the colors in the image.

As for shape information, video surveillance cameras usually capture low quality frames, thus making this kind of information not discriminative in PRID applications. However, spatial information *is* discriminative, because the spatial relationship between different areas of an individual is supposed to be retained in short periods of time.

Our approach overcomes the lack of spatial information introducing the concept of *Differential Spatiogram* (*DS*), a mathematical structure which enfolds both intra and inter relationships between frames. DSs are calculated combining *Inter distances Vectors* (*IrV*) and *Intra distances Matrices*(*IaM*), whose extraction and merging processes into the DS are detailed in the following paragraphs.

#### 2.2.1 Inter Distances Vector

Inter frame relationships are computed extracting a distance vector for each pair of frames, one from the probe set and one from the gallery set. Given two frames $a$ and $b$, their *Inter distance Vector* is defined as:

$$IrV'_{ab} = \begin{pmatrix} D_{1_{ab}} & \dots & D_{n_{ab}} \end{pmatrix}$$

In the above formula, $D_{i_{ab}}$ represents the distance between the $i-th$ strip of the frame $a$ and the corresponding one taken from the frame $b$.

In the single-shot case the cardinality of both the probe set $I_p$ and of the gallery set $I_g$ is $m$; as a consequence, $m^2$ IrVs have to be computed, one for each pair $(i, j)$ where $i \in I_p$ and $j \in I_g$. Therefore, we can estimate the computational complexity related to the calculation of IrVs for the whole dataset as $O(m^2 \cdot n)$ operations, where $n$ is the number of the considered horizontal stripes. $D$ can be any kind of distance metric; in the experimental section, we will show the results obtained using the Bhattacharyya one.

It is interesting to note that the computational cost can be approximated to $O(m^2)$ in the very common case where $n \ll m$.

#### 2.2.2 Intra Distances Matrix

Intra frame distances characterize chromatic relationships between different areas of the same image. Assuming a strong hypothesys about the aspect of an individual, i.e. it can not vary between two different views, it is likely that intra-distances will be retained and therefore are useful for the PRID pipeline. Intra-distances can be modeled by means of an *Intra Distances Matrix* defined as:

$$IaM_f = \begin{pmatrix} 0 & D_{12} & \dots & D_{1n} \\ \vdots & \ddots & \ddots & D_{2n} \\ \vdots & & \ddots & D_{n-1n} \\ 0 & \dots & \dots & 0 \end{pmatrix}$$

In the above formula, $n$ is the number of segment in which each frame is divided. The computation of a IaM for the whole dataset $O(m^2 \cdot n^2)$ operations. As before, $D$ can be any kind of distance metric, and the computational cost can be lowered to $O(m^2)$ if $n \ll m$.

#### 2.2.3 Differential Spatiogram

Given a couple of images $a \in I_p$ and $b \in I_g$, we can combine IrV and IaM to rely chromatic variations between different areas of both $a$ and $b$ with initial chromatic content of the two frames.

The DS is defined as:

$$DS_{ab} = \begin{pmatrix} IrV_{ab}(1) & D_{12} & \dots & D_{1n} \\ 0 & IrV_{ab}(2) & \ddots & D_{2n} \\ \vdots & \ddots & \ddots & D_{n-1n} \\ 0 & \dots & 0 & IrV_{ab}(n) \end{pmatrix}$$

In the above formula, $D_{ij}$ is defined as:

$$D_{ij} = \sqrt{[IaM_a(i,j) - IaM_b(i,j)]^2 \cdot [IrV_{ab}(i) \cdot IrV_{ab}(j)]}$$

where $i \leq j$.

This formulation allows to characterize intra-chromatic relationships while taking in account inter-chromatic relationships, i.e. even if IaM differences between two areas $a$ and $b$ of two images $i$ and $j$ may be comparable, it is not guaranteed that chromatic components of the $k$-th strip of $i$ and $j$ are similar.

DS have an upper triangular matricial structure, which gives the possibility to exploit its algebrical properties to extract a set of metrics from it. We propose three indexes:

**Trace:**

$$Tr(DS_{ab}) = \sum_{i=1}^{n} DS_{ab}(i)$$

**Determinant:**

$$Det(DS_{ab}) = \prod_{i=1}^{n} DS_{ab}(i)$$

**Sum:**

$$Sum(DS_{ab}) = \sum_{i=1}^{n} \sum_{j=1}^{n} DS_{ab}(i,j)$$

An interesting property of both $Tr$ and $Det$ is that they are linearly dependent on the eigenvalues of the DS, which correspond to IrV values (because DS has a upper triangular structure). In addition, we expect that the $Tr$ metric should behave better than the $Det$ one when the two $i$-th strips of the images $a$ and $b$ are similar, i.e. when:

$$IrV_{ab}(i) < \varepsilon, \varepsilon \to 0$$

In fact, given the above condition:

$$Det(DS_{ab}) = \prod_{i=1}^{n} DS_{ab}(i) \to 0$$

while on the contrary:

$$Tr(DS_{ab}) = \sum_{i=1}^{n} DS_{ab}(i) \neq 0$$

Ideally, $Tr$ will better characterize chromatic relationships when only one strip of the frame $a$ is very similar to the corresponding on taken from the frame $b$, while others aren't. However, in real cases it is extremely unlikely to find this situation.

Both $Tr$ and $Det$ metrics don't enfold intra frame information, but *Sum* does. As a consequence, we expect it to have slightly better PRID performances if compared to the others. We will show the results obtained with the last metric in the next section.

## 3 EXPERIMENTAL RESULTS

### 3.1 Experimental Settings

For our experiments, we test our approach against two public PRID datasets, *VIPeR* and *ETHZ*.

**VIPeR** (Gray and Tao, 2008) is one of the most used dataset for PRID. It contains 632 image pairs, observed from two different camera views, each one associated with one person. VIPeR dataset is particulary challenging, because of severe illumination changes and viewpoint variations.

**ETHZ** (Ess et al., 2007) contains three scenes captured from moving cameras. The first sequence contains 83 pedestrians, for a total of 4857 images; the second scene contains 35 pedestrian, for a total of 1936 images; the third and last scene contains 28 pedestrians, for a total of 1762 images. This dataset is not hard as VIPeR in terms of pose variations, but it offers some challenging aspects like illumination changes and occlusions.

As for settings, for VIPeR we follow the widely-used setup ((Farenzena et al., 2010), (Yang et al., 2014), (Kviatkovsky et al., 2013)) which consider half of the overall image pairs, i.e. 316 image paris. As our method is unsupervised, we don't need a training set and a validation set. The test is splitted in two phases: in the first one, images from the first camera are treated as probe set, while frames from the second as gallery set; in the second phase, probe and gallery set are switched. Finally, the average of CMC curves from first and second phase is taken as one trial, and we consider 100 trials of evaluation in order to achieve a statistical significance. As in the other methods, the average of the first 50 ranks are reported.

The setup for ETHZ is different, and it recalls the one used in (Zheng et al., 2009). Firstly, we randomly select one image per each subject in order to build the gallery set; the other images will form the probe set. We then estimate the matching between probe and gallery set for each image of the probe set. As for VIPeR, we will repeat the whole process 100 times to achieve a statistical significance, and we will show the first 7 ranks (as the cardinality of the gallery set is significantly smaller than the cardinality of VIPeR dataset).

## 3.2 Quantitative Results using Different Color Models and Metrics

In this section we show PRID quantitative performances using the various color spaces depicted in section 2 and a *fusion histogram* obtained by the concatenation of HS and log-chromaticity histograms.

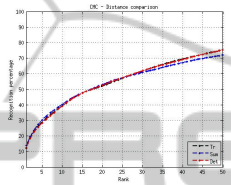
We will solve the following problem: given a pair of frames $a \in I_p$ and $b \in I_g$, and a metric $M$ extracted from $DS_{ab}$, for every frame taken from the probe set we search for the frame of the gallery set that minimizes the $M$ metric:

$$\underset{a,b}{\operatorname{argmin}} \left(M(DS_{ab})\right)$$

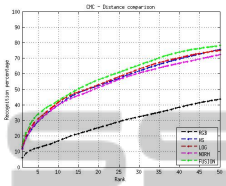
Then we use this information to calculate the rank of every matched frame, building the CMC curve and quantifying the performances of the PRID pipeline exploiting the DS.

We first evaluate the various metrics depicted in section 2.2.3. We expect the performances of *Sum* will be slight better than the performances of *Tr* and *Det*, as shown in figure 2(a). We point out that it is due to the fact that intra-relationships *are relevant*: this can change according to the dataset, and a method to dinamically choose which metric is the most relevant has to be developed.

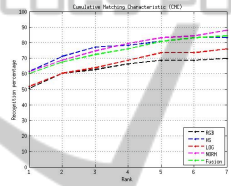
Figures 2(b-e) compare quantitative PRID performances of *Sum* metric using various color models on VIPeR and on each sequence of ETHZ, respectively.



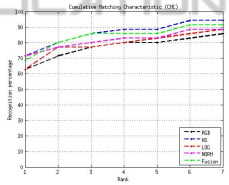
(a) Sum, Tr and Det - VIPeR



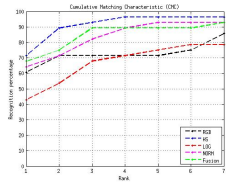
(b) Sum - VIPeR



(c) Sum - ETHZ seq.1



(d) Sum - ETHZ seq.2



(e) Sum - ETHZ seq.3

Figure 2: Comparison of different metrics on ETHZ and VIPeR datasets.

It is clear that HS, log-chromaticity and fusion color spaces behave better than RGB and normalized RGB, as expected. An interesting consideration is while fusion get best performances on VIPeR, HS behaves better on ETHZ: as a consequence, PRID performances are influenced by the color space in use, and the research of an adaptive method to identify the color space which guarantee best performances may be an interesting topic.

# 4 CONCLUSIONS AND FUTURE WORKS

In this paper an approach to model the way humans perform PRID embedding both intra and inter frame chromatic relationships has been presented. Our approach also lowers the computational cost related to the feature extraction and the signature matching.

Differential Spatiograms have been proved to be easily exploitable to characterize relationships between frames extracted from CCTV videos, as it has been show that a relatively limited number of operations is needed to calculate the feature matrix; moreover, the algebraic properties of the Differential Spatiogram can be use to add new knowledge to the whole system.

As this is a snapshot of a work in progress, it may be interesting to focus on future works.

In particular, we plan to focus on following points:

- *Exploit of DS to Elaborate New Metrics*: we plan to elaborate better metrics which exploit algebraic properties of the DS in order to enhance PRID performances;
- *Elaborate Score Matrices*, combining data from multiple features (like MSCR (Forssén, 2007) or SCR (Bak et al., 2010)) and dinamically evaluate the best one to use at runtime, according to the properties of the dataset;
- *Elaborate a Classification Module*, which can categorize different frames and narrow the cardinality of the frame sets where PRID is performed; this will improve performances both in terms of PRID and computational cost;
- *Elaborate a Supervised Approach*, as this kind of methods has been proved to be more effective than unsupervised one;
- *Elaborate a Qualitative-based Metric* to support CMCs in the evaluation of qualitative PRID performances, as only first ranks of the CMC are relevant to *effective* PRID.

# REFERENCES


Bak, S., Corvee, E., Brémond, F., and Thonnat, M. (2010). Person re-identification using spatial covariance regions of human body parts. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 435–440. IEEE.

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

Ess, A., Leibe, B., and Van Gool, L. (2007). Depth and appearance for mobile scene analysis. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE.

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

Finlayson, G., Hordley, S., and Xu, R. (2005). Convex programming colour constancy with a diagonal-offset model. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 3, pages III–948–51.

Forssén, P.-E. (2007). Maximally stable colour regions for recognition and matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

Gevers, T. and Smeulders, A. W. (1999). Color-based object recognition. *Pattern recognition*, 32(3):453–464.

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

Kviatkovsky, I., Adam, A., and Rivlin, E. (2013). Color invariants for person reidentification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(7):1622–1634.

Matsukawa, T., Okabe, T., and Sato, Y. (2014). Person re-identification via discriminative accumulation of local features. In *Pattern Recognition (ICPR), 2014 IEEE Conference on*.

Swain, M. and Ballard, D. (1992). Indexing via color histograms. In Sood, A. and Wechsler, H., editors, *Active Perception and Robot Vision*, volume 83 of *NATO ASI Series*, pages 261–273. Springer Berlin Heidelberg.

Truong Cong, D.-N., Khoudour, L., Achard, C., Meurie, C., and Lezoray, O. (2010). People re-identification by spectral classification of silhouettes. *Signal Processing*, 90(8):2362–2374.

Van de Sande, K. E. A., Gevers, T., and Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1582–1596.

Yang, Y., Shengcai, L., Zhen, L., Dong, Y., and Li, S. Z. (2014). Color models and weighted covariance estimation for person re-identification. In *Pattern Recognition (ICPR), 2014 IEEE Conference on*.

Zheng, W.-S., Gong, S., and Xiang, T. (2009). Associating groups of people. In *Proceedings of the British Machine Vision Conference*, pages 23.1–23.11. BMVA Press. doi:10.5244/C.23.23.