Appearance Features in Encoding Color Space for Visual Surveillance

Lingxiang Wu[1], Min Xu[1*], Guibo Zhu[2], Jinqiao Wang[3], Tianrong Rao[1]

[1] *Global Big Data Technologies Centre, University of Technology Sydney, Australia*
[2] *Research Center for Brain-inspired Intelligence, Institute of Automation*
Chinese Academy of Sciences, 100190, Beijing, China
[3] *National Laboratory of Pattern Recognition, Institute of Automation*
Chinese Academy of Sciences, 100190, Beijing, China

Abstract

Person re-identification and visual tracking are two important tasks in video surveillance. Many works have been done on appearance modeling for these two tasks. However, existing feature descriptors are mainly constructed on three-channel color spaces, such like RGB, HSV and XYZ. These color spaces somehow enable meaningful representation for color, yet may lack distinctiveness for real-world tasks. In this paper, we propose a multi-channel Encoding Color Space (ECS), and consider the color distinction with the design of image feature descriptor. In order to overcome the illumination variation and shape deformation, we design features on the basis of the Encoding Color Space and Histogram of Oriented Gradient (HOG), which enables rich color-gradient characteristics. Additionally, we extract Second Order Histogram (SOH) on the descriptor constructed to capture abstract information with layout constrains. Exhaustive experiments are performed on datasets VIPeR, CAVIAR, CUHK01 and Visual Tracking Benchmark. Experimental results on these datasets show that our feature descriptors could achieve promising performance.

Keywords: Person re-identification, tracking, encoding color space, HOG.

2010 MSC: 00-01, 99-00

[*] Corresponding author: Min Xu

1. Introduction

As the applications of video surveillance, person re-identification and visual tracking have recently attracted increasing research efforts in Multimedia research community [1, 2, 3]. Person re-identification refers to matching pedestrians across disjoint camera views in non-overlapping camera networks. Automatic judging whether or not a person has appeared across the camera views can save human efforts on exhaustively manual searching in large scale datasets. Visual tracking, which is to estimate locations of an object in a sequence of images, benefits a wide range of applications, such as robotic vision, action analysis and human-computer interaction. Both person re-identification and visual tracking are confronted with some serious challenges in real-world scenarios. Illumination variation and shape deformation may make the same pedestrians or objects look different. Partial occlusion, low resolution and background clutter could further complicate the problems. Some challenging examples from two datasets are shown in Fig. 1.

Facing up the above-mentioned appearance challenges, appearance modeling is a crucial step in both tasks. Constructing a robust and descriptive feature representation plays a critical role for improving the performance in visual tracking [4, 5, 6, 7, 8] and person re-identification [9, 10, 11, 12, 13]. Many local or global appearance modeling approaches have been proposed for the above-mentioned two tasks. For person re-identification, Yang *et al.* [14] made use of distribution over salient color names to describe pedestrians. In [15], the authors tackled person re-identification by salience matching. In [16], Farenzena *et al.* exploited symmetry property on person to handle pose variation problem, and proposed the symmetry-driven accumulation of local feature (SDALF). For object tracking, Danelljan *et al.* [17] utilized color names as feature representation, and proved that color attributes provides superior performance for object tracking. In [18], an integral of HOG [19] and color names [20] was constructed, and were used to achieve excellent performance.

Generally speaking, color and gradient features play critical roles in ap-

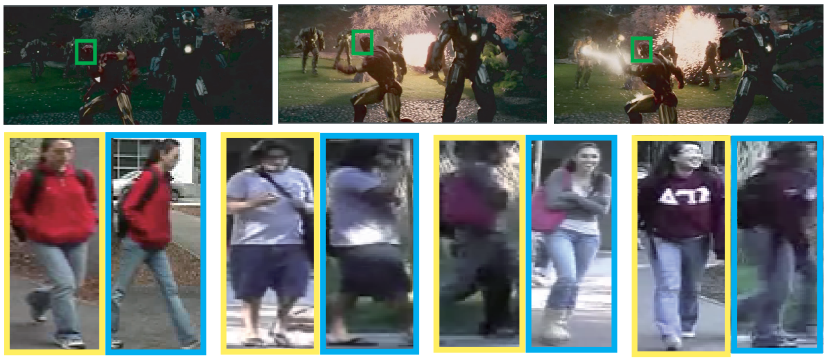


Figure 1: Challenging examples from visual tracking benchmark and person re-identification dataset. The above are three frames in a video that is under variant illumination. The below are four pairs pedestrian examples from two camera views. These images contain illumination variation, viewpoint variation and occlusion etc.

pearance modeling [21]. HOG feature is good at describing abundant gradient information in images, which is robust to illumination variation. Color descriptor, such like color histogram, is good at capturing salient regions and robust to pose variation. Instead of concatenating them directly, we'd like to investigate an appropriate color space to maximally cooperate with the HOG. Most of the existing color spaces have three channels, which enable meaningful representation for color, while lack of distinction in computer vision tasks. Color names transformed RGB color space into an 11-dimensional color attribute feature, and achieved excellent performance in object detection [22] and visual tracking [17]. The limitation is that tough training process is required to learn color names. In this paper, we propose to exploit a multi-channel color space, which could express rich color characteristics. Gradient can be affected by different color channels and/or color spaces, since HOG features get variant respondence on different color channels and/or spaces. Thus, we carefully design a color space cooperating with HOG to achieve a rich color-gradient representation in appearance modeling.

In this paper, we propose a novel Encoding Color Space (ECS). It transforms 3-channel color space into multiple channels, and present rich color characteristic simply. Upon this multi-channel color space, we extract the feature of HOG. The Encoding Color Space and HOG could work closely to present objects. In this way, the noises caused by gradient extraction are inhibited to some extent, while important and subtle gradient changes are stressed or even strengthened. Besides, the gradient changes will look sparse in each color channel, which could make the representation more descriptive. For tracking, we exploit a feature, so called Encoding Color Space based HOg (ECHO), to preserve position information. For person re-identification, upon the ECHO, a Second Order Histogram (SOH) is constructed to model the appearance of a person, which take global and local characteristics into consideration. Then, KISSME [3] metric learning method is utilized to tackle person matching in person re-identification. KCF [23] is utilized for motion estimation in tracking.

The contribution of this paper could be summarized as follows: Firstly, an Encoding Color Space is proposed to represent rich color characteristics. Secondly, we construct two rich color-gradient features, ECHO for tracking , SOH for person re-identification. For illumination variation challenges in video surveillance, we extract HOG on Encoding Color Space to construct a representation for tracking, and extract second of histogram for person re-identification. Thirdly, we execute exhaustive experiments on person re-identification and visual tracking. Our features achieve excellent performance in experiments.

1.1. Person Re-identification

Person re-identification is to match the same pedestrian in disjoint camera views. Given an image as query, when only one image in the gallery should be matched, we call this one pair images matching as single-shot recognition approach. In contrast, if the matching is between two sets of images, we call it as multi-shot recognition approach. In this paper, we focus on the single-shot method. Existing works on single-shot person re-identification can be roughly summarized into two aspects: appearance modeling and person matching.

Appearance Modeling A lot of efforts have been devoted on designing a representative feature descriptor [16, 15, 24, 2]. In [16], Farenzena *et al.*

proposed a symmetry-driven accumulation of local feature (SDALF) by exploiting symmetry and asymmetry perceptual principles to weight features extracted from different body parts. In [15], Zhao *et al.* formulated the person re-identification problem as a salience matching problem. They exploited pairwise salience distribution between pedestrian images, and match person by dense correspondence patch matching. In [24], Zhao *et al.* proposed to learn mid-level filter for person re-identification, and then matching scores of filter responses were integrated with patch matching through RankSVM training. Recently, deep learning based features are also proposed. Ahmed *et al.* [25] formulated the person re-identification problem as binary classification, and proposed a two-branch deep convolutional architecture to solve the problem. In addition to feature design, some works have been done on feature importance learning, since certain features play more important role than others under different circumstances. In [26], Liu *et al.* proposed an unsupervised approach for learning a bottom-up feature importance, so features extracted from different individuals can be weighted adaptively driven by their unique and inherent appearance attributes. Some works focus on learning feature transforms to achieve invariance in complex cross-view variations. In [27], a domain guided dropout algorithm was proposed to learn deep features for multiple domains.

Person Matching Some researchers formulated the person re-identification task as a ranking task. Given a query image, the matched image is expected to have the top position in a ranking list. In [28], Prosser *et al.* formulated it as a ranking task, and developed an Ensemble RankSVM to learn a subspace where the potential true match is given highest ranking. Loy *et al.* [29] developed a manifold ranking model to propagate the query information along the unlabeled data. Some researchers focused on learning a metric [3, 30, 31] for a Mahalanobis-like distance, which makes the distance small for intra-class but large for inter-class. Kostinger *et al.* [3] learnt a distance metric from equivalence constraints, based on a statistical inference perspective. In [32], Li and Wang proposed to learn a matric for cross-view transform, and project pedestrians into a common space for alignment. Xiong *et al.* [31] exploited multiple kernels, linear, χ^2 and RBF-χ^2, on a number of work metric learning methods.

1.2. Tracking

Besides person re-identification, object tracking is also an important task in visual surveillance to test our proposed ECHO. Existing methods for object tracking can be coarsely divided into two categories:generative methods and discriminative methods.

Generative Methods: They learn to represent the target appearance with only object information and search for the most similar image region as the predicted state. FragTrack [33] utilized histogram of local patches to represent the target. Since it took the potential fixed spatial structural information of the target, the method handled the partial occlusion problem well. However, the variation of scene and the modeling of target appearance without surrounding information made it failed in complex environment. IVT [34] learned a low-dimensional subspace representation incrementally with online update for target appearance. The lack of spatial information may lead to model drift. L1 tracker [35] proposed to use feature selection with $\ell 1$ norm minimization to represent the object which improved the robustness ability. VTD [36] combined multiple motion and observation models to account for appearance variation in the conventional particle filter tracking framework [37]. MTT [38] adopted group normalization of multi-task learning to mine the self-similarities between particles to improve the tracking performance. ALSA [39] exploited both partial information and spatial information of the target with a structural local sparse appearance model. LSHT [40] proposed a locality sensitive histogram to exploit the spatial weight for every pixel for visual tracking. Generally, generative trackers are robust to the object occlusion but sensitive to distracters in the context region of the tracked target.

Discriminative Methods: They construct visual object tracking as a binary classification or structure learning problem, which locates the object by seeking the optimal decision plane to separate the object from its background. AOB [41] used AdaBoost to select useful features for object tracking. Because the adopted feature had low representation ability, its performance was easily affected by background clutter which resulted to drift. TLD [42] integrated the modules of short-term tracking, long-term learning and onlne detector for the long-term tracking task, which used the P-N learning to guarantee the online detector's estimated error. PROST [43] combined the template correlation, mean shift optical flow and random forests in a cascade method to reduce the risk caused by the drift problem. Struck [44] proposed the structure output learning for visual tracking which evaded the label prediction problem in the traditional tracking-by-detection discriminative methods and achieved the state-of-the-art performance in the same period. SPOT [45] introduced spatial constraints between the objects in a pictorial-structures framework [46] and trained an online structured SVM for alleviating the object occlusion problem.

Correlation filter-based tracking framework has achieved great success in recent years. Bolme *et al.* [47] represented the target appearance with an adaptive correlation filter by minimizing the output sum of squared error (MOSSE). CSK [48] proposed the circulant structure based on the convolution theorem and overcame the time-consuming problem of tracking-by-detection with the exhaustive method by frequency transformation. Kernelized correlation filters (KCF) [23] was an extended version of CSK by reformulating correlation tracking using the kernelized Ridge regression with multi-channel features. Danelljan *et al.* [49] introduced color attributes to improve tracking performance in colorful sequences. Later, they proposed the DSST tracker [50] with accurate scale estimation by one separate filter for handling the scale variation problem. Zhang *et al.* [51] utilized the spatial-temporal context in the Bayesian framework to interpret correlation tracking. Zhu *et al.* [52] proposed online CUR filter for re-detection to alleviate the long-term tracking problems. Different from the above methods, which mainly focused on the model design, we realize the importance of the feature and proposed a new feature for visual tracking.

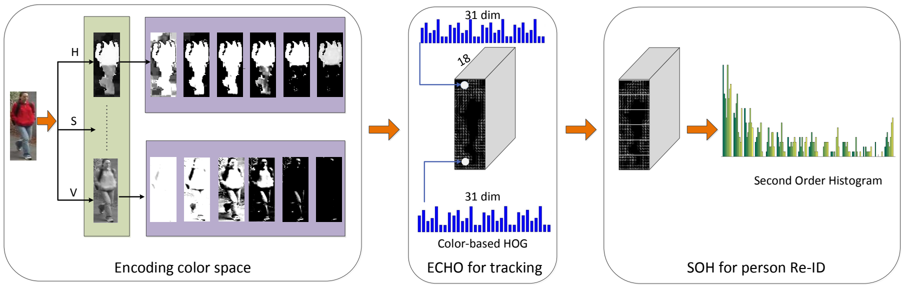


Figure 2: Overview of the proposed appearance modeling. Firstly, we propose an Encoding Color Space. Then, on this color space, we construct two features. They are ECHO for tracking, and SOH for person re-identification.

2. Our Methods

An overview of our proposed appearance model is shown in Fig. 2. Existing color spaces enable a meaningful color representation of color, such like RGB, HSV and XYZ. However, most high-level computer vision tasks may not request such representation but a mathematically descriptive one. In order to overcome the illumination variation challenge, we propose a novel multiple-channel color space, Encoding Color Space (ECS). Based on the ECS and HOG, we construct two rich color-gradient image features, ECHO for tracking, and SOH for person re-identification. In the following, we will introduce our methods in details.

2.1. Encoding Color Space

Human have the ability of color constancy, and are able to perceive stable object color in spite of serious variation in illumination. However, it is a challenge for computer system, especially in real-world circumstance. For outdoor pedestrian tracking task, surveillance cameras are usually located far from target objects, which results in low resolution and inappropriate viewpoints in captured images. Thus, color plays a crucial role in appearance modeling. When representing colors, illumination variation is a challenge we should not ignore. The illumination change across disjoint camera views or in a sequence of images could seriously affects the performance in person re-identification or tracking.

Changes in illumination could be regarded as a transformation under certain light source. It can be modeled by a diagonal-offset model [53] as

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix} + \begin{pmatrix} o_1 \\ o_2 \\ o_3 \end{pmatrix} \tag{1}$$

where u corresponds to the image taken under an unknown light source, while c corresponds to the same image transformed. Based on the diagonal-offset model, common changes in image value for illumination can be categorized into 5 types, i.e. light intensity changes, light intensity shifts, light intensity changes and shifts, light color changes, as well as light color changes and shifts.

In trichromatic theory, three channels are required to generate the full range of human visible color. Thus, three camera sensitivities need to be defined to specify the incoming spectral power distribution. The cameras are expected to "perceive" spectra as human eye does, so a number of three-channel color spaces have been proposed. Although, existing color spaces enable a meaningful representation of color, either physically or perceptually, such like RGB, HSV and XYZ, the high-level computer vision tasks do not require such a representation. In computer vision tasks, a color space incorporation with other descriptor achieving excellent performance mathematically is what we pursue. Moreover, existing color spaces were proved to be invariant to photometric changes with the assumption of white illumination. However, in real-world images, the light source might have variant spectral power distribution. The actual variety in outdoor light sources might be much larger than expected. Therefore, it is urgent to propose a novel color space, and consider the color distinction into design of image feature descriptor.

Recently, color names [20, 54] achieved excellent performance on object detection, object tracking and action recognition [22, 17, 55]. They are linguistic color labels assigned to pixel, region or objects in images. In [20], a mapping projecting RGB observations to 11 linguistic color labels were learned through real-world images. The 11 color names of the English language include: black, blue, brown, grey, green, orange, pink, purple, red, white, and yellow. The 11-dimensional color names inspired us that color space with multiple channels may have rich representative ability. While the color names were learned from real-world images from Google Image. It not only requires training time, but also might be affected by wrong labels. We'd like to design a simple but efficient multiple channel color space. We assume color and structure information are the most important information used for video surveillance tasks. Therefore, we design an Encoding Color Space to incorporate with Histogram of Gradient (HOG).

To construct the Encoding Color Space, firstly, we normalize all pixel values in the original color space into the range of [0, 1]. Take RGB color space as an example, invariant values can be obtained through normalizing RGB value by their intensity ($I = R + G + B$). According to [56], pixel values of a homogeneously colored surface might generate streaks in RGB color space, and these streaks are mainly caused by intensity changes. Our first step normalization process prevents the intensity variation. Secondly, a fixed discretization scale, N, is to be set. N is also the number of sub-channels one original channel would transform to. Thirdly, the normalized pixel value is expressed in discretization method. For $n < N$ $(n \in R)$, we denote $U(n)$ as the discretization representation of number n as:

$$U(n) = \left\{ i_1, i_2 \cdots i_{\lfloor n \rfloor}, n - \lfloor n \rfloor, 0 \cdots 0 \right\} \tag{2}$$

where $i_t = 1$ $(1 \leq t \leq \lfloor n \rfloor)$. For the normalized pixel value x, the discretization encoding can be calculated as follows:

$$\phi(x) = U(Nx) \tag{3}$$

For instance, when normalized pixel value $x = 0.4$ and $N = 6$, then $\phi(0.4) = \{1, 1, 0.4, 0, 0, 0\}$. At last, we apply the discretization encoding from pixel level to channel level, which means we transform each channel (e.g. green channel, and blue channel) into N sub-channels based on Eq. (3). In this way, a three-channel original color space is encoded into a 3 × N-channel color space.

It should be noted that we keep the decimal in the discretization representation instead of totally binary discretization to preserve more detailed information. With the decimal, the discretization representation can be recovered to the original one conversely. Without training, our encoding color space enables richer color representation in a simple but efficient way. To be specific, the similar pixels (difference within range $255/N$) may contain the same value 0 or 1 in most sub-channels, but keep different on the decimals. While the variant pixels may apart one on some sub-channels. This property benefits the gradient descriptors. It can inhibit noises caused by gradient extraction, but stress even magnifies those significant gradient changes.

2.2. Color Based HOG

Many gray-scale descriptors have been extended to their color-based counterparts, because color has high discriminative power. As one of the most successful descriptors, SIFT [57] encodes the distribution of Gaussian gradient within an image region. It performs well in representing intensity pattern, and robust to small deformation and localization errors. In [58], it has been proved that the transformed color SIFT is scale-invariant, shift-invariant and invariant to light color changes. In [59], a HOG feature pyramid is defined by calculating HOG features on each level of standard image pyramid. It captures the local shape properties, and achieves invariant to deformations in small areas.

Color and gradient features have shown better performance for appearance modeling, where color feature is robust to viewpoint variance and gradient feature provides a rich representation robust to illumination change. In order to integrate color features and gradient features, we propose a feature descriptor, Encoding Color space HOg (ECHO). ECHO is presented by calculating HOG in each channels of the Encoding Color Space, and concatenating HOG features for all the channels in the third dimension as a feature descriptor. HOG is computed on a dense grid of uniformly spaced cells with overlapping local contrast normalization. A 32-bins histogram on each cell is built based on the standard 9 orientations. By removing the last all-zero dimension, we only use the 31-bins

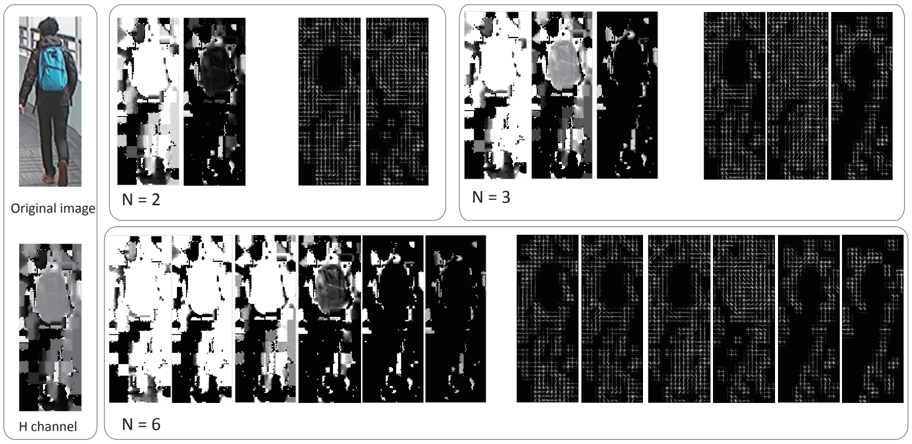


Figure 3: Encoding Color Space and HOG visualization with variant discretization level N. We present examples on H channel from original HSV color space.

information for each cell. HOG has different presentative ability on different channels. In Fig. 3, we present some examples in Encoding Color Space and HOG visualization with variant discretization levels. It can be seen that HOG get variant respondence on different channels. With higher discretiation level, the HOG and sub-channel may describe more details in certain range.

HOG for all channels can be concatenated to be either a three-dimensional matrix or to be a long vector. For visual tracking, since the position information is very useful, we utilize the position invariant matrix. The HSV color space is transformed into a HSV encoding color space firstly. Then, HOG features are extracted for each channel in the HSV encoding color space. N is set as 6 experimentally. The three-dimensional matrix forms the ECHO feature, which is used in our experiments. Considering the speed requirement in tracking, we only use the HSV color space. We refer ECHO as the first order histogram. Since the position information is not necessary for person re-identification, ECHO is processed continuatively as follows.

2.3. Second Order Histogram

Most of existing research utilized first order gradient information such as SIFT, SURF, LBP and HOG to present the geometric properties of an object, while, in [60], Morgan *et al.* mentioned that the first order gradient might be insufficient to accurately capture the perceived visual features as human beings. Considering the high order visual features may contain more abstract information, and histogram can be used to represent an estimation of the probability distribution of numerical data, we construct a Second Order Histogram (SOH) upon the first order histogram.

The first order histogram, the HOG map, is a three-dimensional matrix concatenated by HOG on all channels, whose length l of the third dimension is 31 multiples the number of Encoding Color Space channels. We compute the second order histogram with respect to each third dimension with a layout constrain. We divide the HOG map into 6 horizontal stripes to roughly capture person's head, upper and lower torso, upper and lower legs. Afterwards, on each stripe, l histograms are extracted. Then, histograms are also extracted for the whole image to its capture global distribution property. Our second order histogram for person re-identification can be represented as:

$$\mathrm{F} = [\mathrm{H}_1,\mathrm{H}_2,\cdots,\mathrm{H}_6,\mathrm{H}_G]^T \tag{4}$$

where

$$\mathrm{H}_i = [\mathrm{h}_{i1},\mathrm{h}_{i2},\cdots,\mathrm{h}_{il}] \tag{5}$$

h denotes the 16-bins histogram.

Here, we combine three encoding color spaces (HSV, YCbCr and Lab) to make them complement with each other. Then, a PCA process is executed to reduce feature dimensions. We apply PCA for two reasons. Firstly, our SOH somehow contains redundant features as it is exploited on a combination of three color spaces. Consequently, the SOH is in a very high dimension (3*3*6*31*7*16=187488 dimension per image). We use PCA to reduce dimensions from 187488 to 77 experimentally. Secondly, we reduce the feature dimension for efficient and practical computation in metric learning method KISSME which will be described in Sec. 4.1.

Our Second Order Histogram is able to capture high-order color-gradient property, and reduce feature dimensions in a certain extent. In addition, it reflects the statistical characteristics within a spatial layout. For instance, without the second order information, the first order histogram cannot accurately measure whether those cells with analogue oriented statistics are positioned together or separately.

We construct two rich color-gradient features for video surveillance. One is ECHO for tracking, and the other is SOH for person re-identification.

3. Visual Tracking

3.1. KCF Tracker

For visual tracking, we base our appearance model on a KCF tracker [23]. KCF is an extension of CSK tracker [48] by reformulating correlation tracking using a kernelized Ridge Regression, which can achieve real-time performance through tracking-by-detection.

In KCF, the goal of training is to find a function to minimize the squared error over samples x_i and their regression targets y_i.

$$\min_{\mathrm{w}} \sum_i \left(f\left(\mathrm{x}_i\right) - y_i\right)^2 + \lambda ||\mathrm{w}||^2 \tag{6}$$

where λ is a regularization parameter that controls overfitting. The cost function Eq. (6) is minimized by $\mathrm{w} = \sum\limits_i \alpha_i \varphi\left(\mathrm{x}_i\right)$, where $\varphi\left(\mathrm{x}\right)$ is the mapping induced by kernel k, and the coefficients α are

$$\hat{\alpha} = \frac{\hat{\mathrm{y}}}{\hat{k}^{\mathrm{xx}'} + \lambda}. \tag{7}$$

Here, the hat denotes the DFT (Discrete Fourier Transform) operator, $\hat{\mathrm{x}} = \mathcal{F}\left(\mathrm{x}\right)$. Each elements of y is the regression target y_i. Gaussian kernel is exploited here as:

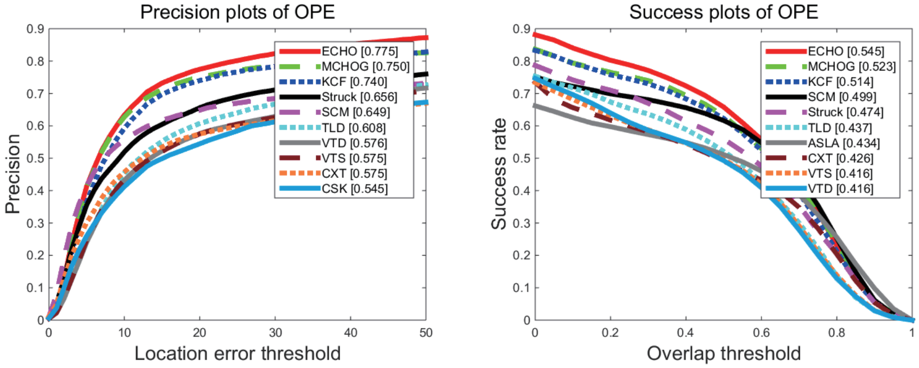


Figure 4: Precision and success plots of one-pass evaluation for 50 videos in [61]. The mean precision scores at threshold 20 and AUC are reported in the legends. The discretization level N is set as 6 experimentally. In both cases, our ECHO achieves favorable performance compared with the state-of-the-art methods.

$$k^{\mathrm{xx}'} = \exp\left(-\frac{1}{\sigma^2}\left(||\mathrm{x}||^2 + ||\mathrm{x}'||^2 - 2\mathcal{F}^{-1}\left(\sum\nolimits_C \hat{\mathrm{x}}^*_C \odot \hat{\mathrm{x}}'_C\right)\right)\right). \tag{8}$$

where C denotes the length of x vector. The detection score is formulated as:

$$f\left(z\right) = \mathcal{F}^{-1}\left\{\hat{k}^{\mathrm{xz}} \odot \hat{\alpha}\right\}. \tag{9}$$

More details can be found in [23].

3.2. Evaluation

We evaluate our method on CVPR2013 Visual Tracker Benchmark [61], and follow their evaluation protocol. Benchmark is a popular dataset with 50 challenging videos. We report performance with precision plots and success plots. The precision plot shows the percentage of frames whose estimated location is within a given threshold distance of ground truth. A higher precision at low thresholds indicates the tracker is more accurate. The mean precision scores at threshold 20 are reported in legend. The success plot presents the ratios of successful frames, whose overlap is larger than the given threshold, at the thresholds varied from 0 to 1. The tracking algorithms are ranked by the area under curve (AUC) of each success plot. One-pass evaluation (OPE) is reported.

In Fig. 4, the precision plots and success plots are shown with some state-of-the-art trackers. Our ECHO can achieve a promising result. We compare the performance of different kinds of trackers, including MCHOG [18], KCF [23], Struck [62], SCM [63], TLD [64], VTD [65], CXT [66], CSK [48], VTS [67] and our ECHO. The KCF used HOG features on gray-scale and Gaussian kernel. Our ECHO approach outperforms the baseline KCF tracker in a large extent. The ECHO enhances the mean precision rate from 0.740 to 0.775. MCHOG is also based on KCF, but exploited HOG features upon 11-dimensional color naming space. Here, we set the parameters exactly the same for KCF, MCHOG and our ECHO. Our ECHO boosts the precision rate of MCHOG from 0.750 to 0.775, and AUC of success plot from 0.523 to 0.545. It should be noted that color naming space seriously relies on practical training, and is inflexible on the number of dimension. As shown in [18], the tracker achieves 74.2% when using HOG in traditional HSV color space. While our ECHO achieves 77.5% using HSV Encoding Color Space.

In Fig. 5, we show success plots on sequences with annotated attributes in [61]. ECHO is ranked top. It can be seen that our ECHO is robust to illumination variation and shape deformation as we expected.

In order to find the optimal discretization level N, we perform experiments with respect to variant discretization levels from 2 to 8. The mean precision and the computational complex (mean frames per second (M-FPS)) are listed in Table 1. We can see that the optimal discretization level for tracking is 6. The lager the N is, the more expensive the computing cost would be.

Table 1: Comparison of ECHO with variant discretization level N.

N	2	3	4	5	6	7	8
Mean precision	0.699	0.696	0.731	0.721	**0.775**	0.705	0.723
Mean FPS	79.88	73.28	69.72	66.04	65.35	65.06	64.96

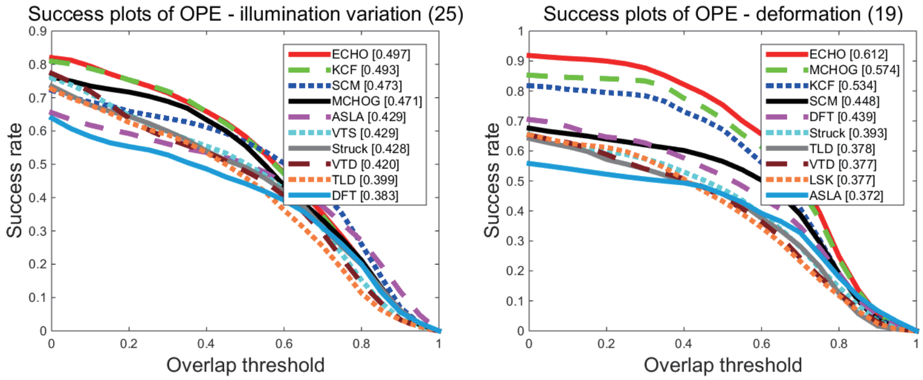


Figure 5: Success plots for videos with attributes: illumination variation and deformation. Compared with existing methods, our ECHO is the most resilient in these two cases as expected.

4. Person Re-identification

4.1. KISSME Metric Learning

In person re-identification, the matched pedestrians are expected to be ranked top at the rank list. We utilize KISSME [3] for metric learning. It learns a metric M for a Mahalanobis-like distance by maximizing the inter-class distances while minimizing the intra-class distances. KISSME views the statistical decision whether a pair is dissimilar or not by the log likelihood ratio test of the two Gaussian distribution. KISSME can be summarized as to learn metric M for distance:

$$d_M^2(\mathrm{x}_i, \mathrm{x}_j) = (\mathrm{x}_i, \mathrm{x}_j)^T M (\mathrm{x}_i, \mathrm{x}_j) \tag{10}$$

where x_i and x_j are SOH features of two samples. The M is defined as:

$$M = \sum{}_{S}^{-1} - \sum{}_{D}^{-1} \tag{11}$$

where

$$\sum{}_S = \frac{1}{|S|} \sum_{\mathrm{x}_i, \mathrm{x}_j \in S} (\mathrm{x}_i - \mathrm{x}_j)(\mathrm{x}_i - \mathrm{x}_j)^T \tag{12}$$

$$\sum{}_D = \frac{1}{|D|} \sum_{\mathrm{x}_i, \mathrm{x}_j \in D} (\mathrm{x}_i - \mathrm{x}_j)(\mathrm{x}_i - \mathrm{x}_j)^T \tag{13}$$

are the covariance metrics for similarity pairs S and dissimilarity pairs D respectively. More details about KISSME can be referred to [3].

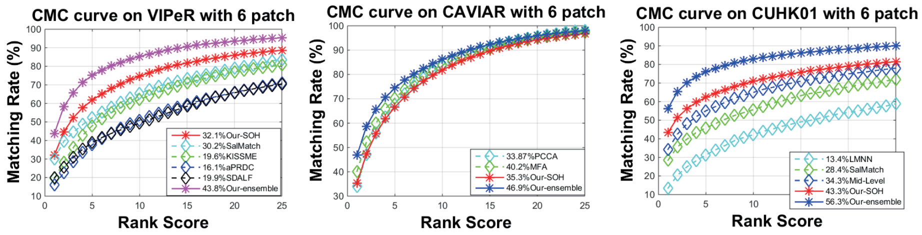
4.2. Evaluation

To evaluate the performance of our appearance model for person re-identification, we perform experiments on three popular datasets: VIPeR [68], CAVIAR [69] and CUHK01 [32]. The Cumulative Matching Characteristics (CMC) curves and Rank-1 accuracy are reported in this paper. The CMC curve represents the chance of the true matching appearing at the top 1, 2,..., N of the ranked list. The first point on the CMC curve is Rank-1 accuracy. We follow the protocol in [31], and divide samples in each dataset into two parts, 50% for training and 50% for testing. All experiments adopt the single-shot evaluating protocol. The average matching accuracy over 10 times processing repeatedly is reported.

Table 2: Rank-1 Accuracy (%) on VIPeR.

Methods	Rank-1 accuracy (%)
aPRDC[26]	16.1
KISSME[3]	19.6
SDALF[16]	19.9
SalMatch[15]	30.2
LOMO[70]	34.1
SCNCD[14]	37.8
MidLevel+LADF[24]	43.4
Ensemble[31]	36.1
LTR[71]	**45.9**
Semantic[72]	41.6
Our-SOH	32.1
Our-ensemble	43.8

Experiments on VIPeR: VIPeR is one of the most challenging datasets for person re-identification. It contains 632 pairs pedestrians captured by two outdoor cameras, each of which captures one image per person. Pedestrians



(a) VIPeR (b) CAVIAR (c) CUHK01

Figure 6: Person re-identification CMC curves on three datasets: VIPeR, CAVIAR and CUHK01.

images in Fig. 1 are from VIPeR dataset. Most images face up challenging scenarios including illumination variation, viewpoint variation, occlusion etc.

The comparison between our method and the state-of-the-art methods is presented in Fig. 6 (a) and Table 2. Popular methods without CMC curves are also listed with Rank-1 accuracy in Table 2. For the baseline method KISSME, Koestinger *et al.* used HSV, Lab color histogram and texture feature extracted by LBP. With the same metric learning method, our SOH enhances from 19.6% to 32.1% over the baseline. Compared with other two well-known person re-id feature descriptors, i.e. SalMatch [15] and SDALF[16], our descriptor also outperforms them by 1.9% and 12.3% respectively. Additionally, we fuse our SOH feature with CNN feature, BoW [73] and LOMO [70] to get a competent ensemble feature. Our ensemble feature achieves 43.8% at the Rank-1 accuracy compared to the ensemble methods introduced in [24, 31, 71, 72].

Table 3: Rank-1 Accuracy (%) on CAVIAR.

Methods	Rank-1 accuracy (%)
MFA [31]	40.2
PCCA [74]	33.9
Our-SOH	35.3
Our-ensemble	**46.9**

In order to prove the advantages of Encoding Color Space, we compare some

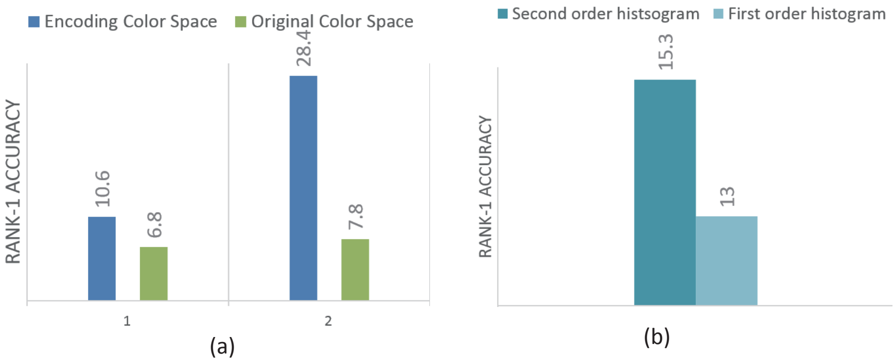


Figure 7: (a) Rank-1 accuracy comparison between Encoding Color Space and original color space on two feature descriptors. 1 is RGB color histogram. 2 is SOH on HSV. In both cases, our Encoding Color Space performs favorably batter than the original color space. (b) Rank-1 accuracy comparison between the first order histogram and the second order histogram.

features on Encoding Color Space with those on original color space in Fig. 7 (a). It can be seen that the Rank-1 accuracy of the original RGB color histogram is only 6.8%, while the color histogram on RGB Encoding Color Space achieves 10.6%. The SOH on HSV Encoding Color Space outperforms the one applying HSV directly from 7.8% to 28.4%. In Fig. 7 (b), a comparison between the first order histogram and the second order histogram is presented. The second order histogram improves 2.3% of the Rank-1 accuracy.

Experiments on CAVIAR: CAVIAR is a dataset collected in a shopping mall by two surveillance cameras. It contains 1,220 images of 72 pedestrians. Each pedestrian has 10 to 20 images. The major challenges for this dataset arises from pose variance and resolution variance. Comparison between our methods and the state-of-the-art methods [31, 74] are presented in Fig. 6 (b) and Table 3. Our features achieve promising results on CAVIAR dataset.

Experiments on CUHK01: CUHK01 dataset is collected by two cameras in a campus environment. A camera captures the front view or back view, while the other captures the side view. CUHK01 contains 971 pedestrians. For each pedestrian, two images are captured by each camera. Images in this dataset are

Table 4: Rank-1 Accuracy (%) on CUHK01.

Methods	Rank-1 accuracy (%)
SalMatch [15]	28.5
LTR [71]	53.4
LMNN [75]	13.4
Chen *et al.* [76]	50.4
Ahmed *et al.* [77]	47.5
Mid-Level [24]	34.3
Li *et al.* [78]	**59.5**
Our-SOH	43.3
Our-ensemble	56.3

of high resolution. Some pedestrians are occluded by bags.

The experiment results are shown in Fig. 6 (c) and Table 4. Our SOH achieves a certain high Rank-1 accuracy as 43.3%. Our ensemble feature achieves 56.3%.

5. Conclusion

In this paper, we propose the Encoding Color Space, which is a multiple channel color space with rich color characteristics. Upon the Encoding Color Space, two distinct appearance models are constructed, i.e. ECHO for visual tracking, and SOH for person re-identification. Our appearance models are efficient in leveraging rich color-gradient property to overcome illumination variation and shape deformation. Exhaustive experiments are performed for both visual tracking and person re-identification. The experiment results show that the proposed features achieve excellent performance in video surveillance as expected.

References

[1] B. Ma, J. Shen, Y. Liu, H. Hu, L. Shao, X. Li, Visual tracking using strong classifier and structural local sparse descriptors, IEEE Transactions on Multimedia 17 (10) (2015) 1818–1828.

[2] S. Sunderrajan, B. Manjunath, Context-aware hypergraph modeling for re-identification and summarization, IEEE Transactions on Multimedia 18 (1) (2016) 51–63.

[3] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2288–2295.

[4] P. Feng, C. Xu, Z. Zhao, F. Liu, C. Yuan, T. Wang, K. Duan, Sparse representation combined with context information for visual tracking, Neurocomputing.

[5] R. Shi, G. Wu, W. Kang, Z. Wang, D. D. Feng, Visual tracking utilizing robust complementary learner and adaptive refiner, Neurocomputing.

[6] J. Wang, W. Liu, W. Xing, S. Zhang, Two-level superpixel and feedback based visual object tracking, Neurocomputing.

[7] B. Zhuang, L. Wang, H. Lu, Visual tracking via shallow and deep collaborative model, Neurocomputing 218 (2016) 61–71.

[8] L. Ma, J. Lu, J. Feng, J. Zhou, Multiple feature fusion via weighted entropy for visual tracking, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 3128–3136.

[9] L. An, X. Chen, S. Yang, Multi-graph feature level fusion for person re-identification, Neurocomputing.

[10] Z. Liu, Z. Zhang, Q. Wu, Y. Wang, Enhancing person re-identification by integrating gait biometric, Neurocomputing 168 (C) (2015) 1144–1156.

[11] X. Wang, C. Zhao, D. Miao, Z. Wei, R. Zhang, T. Ye, Fusion of multiple channel features for person re-identification, Neurocomputing 213 (2016) 125–136.

[12] W. Fang, H.-M. Hu, Z. Hu, S. Liao, B. Li, Perceptual hash-based feature description for person re-identification , Neurocomputing.

[13] L. Ren, J. Lu, J. Feng, J. Zhou, Multi-modal uniform deep learning for rgb-d person re-identification, Pattern Recognition 72 (2017) 446–457.

[14] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, S. Z. Li, Salient color names for person re-identification, in: European Conference on Computer Vision, Springer, 2014, pp. 536–551.

[15] R. Zhao, W. Ouyang, X. Wang, Person re-identification by salience matching, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2528–2535.

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

[17] M. Danelljan, F. Shahbaz Khan, M. Felsberg, J. Van de Weijer, Adaptive color attributes for real-time visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1090–1097.

[18] G. Zhu, J. Wang, Y. Wu, X. Zhang, H. Lu, Mc-hog correlation tracking with saliency proposal, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI Press, 2016, pp. 3690–3696.

[19] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 1, IEEE, 2005, pp. 886–893.

[20] J. Van De Weijer, C. Schmid, J. Verbeek, D. Larlus, Learning color names for real-world applications, IEEE Transactions on Image Processing 18 (7) (2009) 1512–1523.

[21] S. Gong, M. Cristani, S. Yan, C. C. Loy, Person re-identification, Vol. 1, Springer, 2014.

[22] F. S. Khan, R. M. Anwer, J. Van De Weijer, A. D. Bagdanov, M. Vanrell, A. M. Lopez, Color attributes for object detection, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 3306–3313.

[23] J. F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, TPAMI.

[24] R. Zhao, W. Ouyang, X. Wang, Learning mid-level filters for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 144–151.

[25] E. Ahmed, M. Jones, T. K. Marks, An improved deep learning architecture for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3908–3916.

[26] C. Liu, S. Gong, C. C. Loy, X. Lin, Person re-identification: What features are important?, in: European Conference on Computer Vision, Springer, 2012, pp. 391–401.

[27] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1249–1258.

[28] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, Q. Mary, Person re-identification by support vector ranking., in: BMVC, Vol. 2, 2010, p. 6.

[29] C. C. Loy, C. Liu, S. Gong, Person re-identification by manifold ranking, in: Image Processing (ICIP), 2013 20th IEEE International Conference on, IEEE, 2013, pp. 3567–3571.

[30] Z. Wang, R. Hu, C. Liang, Y. Yu, J. Jiang, M. Ye, J. Chen, Q. Leng, Zero-shot person re-identification via cross-view consistency, IEEE Transactions on Multimedia 18 (2) (2016) 260–272.

[31] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: European conference on computer vision, Springer, 2014, pp. 1–16.

[32] W. Li, X. Wang, Locally aligned feature transforms across views, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3594–3601.

[33] A. Adam, E. Rivlin, I. Shimshoni, Robust fragments-based tracking using the integral histogram, in: CVPR, Vol. 1, IEEE, 2006, pp. 798–805.

[34] D. Ross, J. Lim, R. Lin, M. H. Yang, Incremental learning for robust visual tracking, IJCV 77 (1-3) (2008) 125–141.

[35] X. Mei, H. Ling, Y. Wu, E. Blasch, L. Bai, Minimum error bounded efficient ℓ1 tracker with occlusion detection, in: CVPR, IEEE, 2011, pp. 1257–1264.

[36] J. Kwon, K. M. Lee, Visual tracking decomposition, in: CVPR, IEEE, 2010, pp. 1269–1276.

[37] M. Isard, A. Blake, Condensation conditional density propagation for visual tracking, International journal of computer vision 29 (1) (1998) 5–28.

[38] T. Zhang, B. Ghanem, S. Liu, N. Ahuja, Robust visual tracking via multi-task sparse learning, in: CVPR, IEEE, 2012, pp. 2042–2049.

[39] X. Jia, H. Lu, M. Yang, Visual tracking via adaptive structural local sparse appearance model, in: CVPR, IEEE, 2012, pp. 1822–1829.

[40] S. He, Q. Yang, R. W. Lau, J. Wang, M.-H. Yang, Visual tracking via locality sensitive histograms, in: CVPR, IEEE, 2013, pp. 2427–2434.

[41] H. Grabner, M. Grabner, H. Bischof, Real-time tracking via on-line boosting, in: BMVC, 2006, pp. 47–56.

[42] Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection, IEEE-TPAMI 34 (7) (2012) 1409–1422.

[43] J. Santner, C. Leistner, A. Saffari, T. Pock, H. Bischof, Prost: Parallel robust online simple tracking, in: CVPR, IEEE, 2010, pp. 723–730.

[44] S. Hare, A. Saffari, P. Torr, Struck: Structured output tracking with kernels, in: ICCV, IEEE, 2011, pp. 263–270.

[45] L. Zhang, L. van der Maaten, Preserving structure in model-free tracking, TPAMI 36 (4) (2014) 756–769.

[46] M. Fischler, R. Elschlager, The representation and matching of pictorial structures, IEEE Transactions on Computers 22 (1) (1973) 67–92.

[47] D. Bolme, J. Beveridge, B. Draper, Y. Lui, Visual object tracking using adaptive correlation filters, in: CVPR, IEEE, 2010, pp. 2544–2550.

[48] J. Henriques, R. Caseiro, P. Martins, J. Batista, Exploiting the circulant structure of tracking-by-detection with kernels, in: ECCV, Springer, 2012, pp. 702–715.

[49] M. Danelljan, F. Shahbaz Khan, M. Felsberg, J. Van de Weijer, Adaptive color attributes for real-time visual tracking, in: CVPR, IEEE, 2014.

[50] M. Danelljan, G. Häger, F. Khan, M. Felsberg, Accurate scale estimation for robust visual tracking, in: BMVC, 2014.

[51] K. Zhang, L. Zhang, Q. Liu, D. Zhang, M. Yang, Fast visual tracking via dense spatio-temporal context learning, in: ECCV, Springer, 2014, pp. 127–141.

[52] G. Zhu, J. Wang, Y. Wu, H. Lu, Collaborative correlation tracking, in: BMVC, 2015.

[53] J. von Kries, Influence of adaptation on the effects produced by luminous stimuli, Sources of color vision (1970) 109–119.

[54] A. Lindner, S. Süsstrunk, Semantic-improved color imaging applications: It is all about context, IEEE Transactions on Multimedia 17 (5) (2015) 700–710.

[55] F. S. Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, A. M. Lopez, M. Felsberg, Coloring action recognition in still images, International journal of computer vision 105 (3) (2013) 205–221.

[56] T. Gevers, A. Gijsenij, J. Van de Weijer, J.-M. Geusebroek, Color in computer vision: fundamentals and applications, Vol. 23, John Wiley & Sons, 2012.

[57] D. G. Lowe, Object recognition from local scale-invariant features, in: Computer vision, 1999. The proceedings of the seventh IEEE international conference on, Vol. 2, Ieee, 1999, pp. 1150–1157.

[58] K. Van De Sande, T. Gevers, C. Snoek, Evaluating color descriptors for object and scene recognition, IEEE transactions on pattern analysis and machine intelligence 32 (9) (2010) 1582–1596.

[59] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8.

[60] M. J. Morgan, Features and the primal sketch, Vision research 51 (7) (2011) 738–753.

[61] Y. Wu, J. Lim, M.-H. Yang, Online object tracking: A benchmark, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 2411–2418.

[62] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S. L. Hicks, P. H. Torr, Struck: Structured output tracking with kernels, IEEE transactions on pattern analysis and machine intelligence 38 (10) (2016) 2096–2109.

[63] W. Zhong, H. Lu, M.-H. Yang, Robust object tracking via sparsity-based collaborative model, in: Computer vision and pattern recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 1838–1845.

[64] Z. Kalal, J. Matas, K. Mikolajczyk, Pn learning: Bootstrapping binary classifiers by structural constraints, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 49–56.

[65] J. Kwon, K. M. Lee, Visual tracking decomposition, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 1269–1276.

[66] T. B. Dinh, N. Vo, G. Medioni, Context tracker: Exploring supporters and distracters in unconstrained environments, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1177–1184.

[67] J. Kwon, K. M. Lee, Tracking by sampling trackers, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 1195–1202.

[68] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, in: Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), Vol. 3, 2007.

[69] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification., in: Bmvc, Vol. 2, 2011, p. 6.

[70] S. Liao, Y. Hu, X. Zhu, S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition, 2015, pp. 2197–2206.

[71] S. Paisitkriangkrai, C. Shen, A. van den Hengel, Learning to rank in person re-identification with metric ensembles, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1846–1855.

[72] Z. Shi, T. M. Hospedales, T. Xiang, Transferring a semantic representation for person re-identification and search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4184–4193.

[73] L. Zheng, L. Shen, L. Tian, S. Wang, J. Bu, Q. Tian, Person re-identification meets image search, arXiv preprint arXiv:1502.02171.

[74] A. Mignon, F. Jurie, Pcca: A new approach for distance learning from sparse pairwise constraints, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2666–2672.

[75] W. Li, R. Zhao, X. Wang, Human reidentification with transferred metric learning, in: Asian Conference on Computer Vision, Springer, 2012, pp. 31–44.

[76] S. Z. Chen, C. C. Guo, J. H. Lai, Deep ranking for person re-identification via joint representation learning, IEEE Transactions on Image Processing 25 (5) (2015) 2353.

[77] E. Ahmed, M. Jones, T. K. Marks, An improved deep learning architecture for person re-identification, in: Computer Vision and Pattern Recognition, 2015, pp. 3908–3916.

[78] S. Li, M. Shao, Y. Fu, Cross-view projective dictionary learning for person re-identification, in: IJCAI - International Joint Conference on Artificial Intelligence, 2015.