Person Re-identification in Appearance Impaired Scenarios

Mengran Gou
www1.coe.neu.edu/~mengran/
Xikang Zhang
zhangxk@ece.neu.edu
Angels Rates-Borras
ratesborras.a@husky.neu.edu
Sadjad Asghari-Esfeden
www1.coe.neu.edu/~sadjad/
Octavia Camps
camps@coe.neu.edu
Mario Sznaier
msznaier@coe.neu.edu

Robust Systems Lab
Electrical and Computer Engineering
Northeastern University
Boston

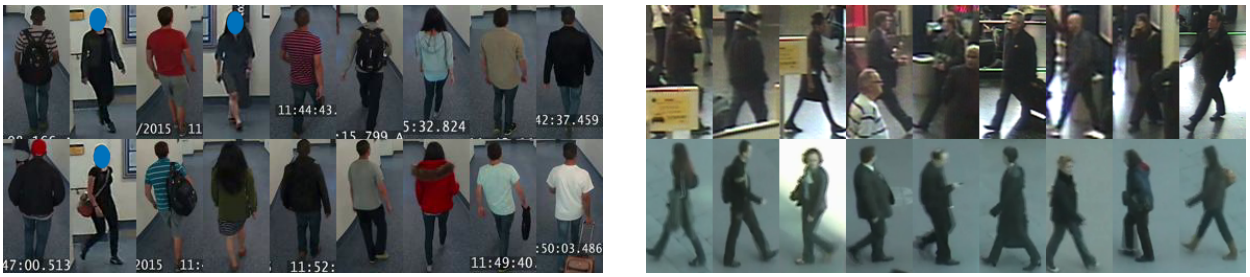
Abstract

Person re-identification is critical in surveillance applications. Current approaches rely on appearance-based features extracted from a single or multiple shots of the target and candidate matches. These approaches are at a disadvantage when trying to distinguish between candidates dressed in similar colors or when targets change their clothing. In this paper we propose a dynamics-based feature to overcome this limitation. The main idea is to capture soft biometrics from gait and motion patterns by gathering dense short trajectories (tracklets) which are Fisher vector encoded. To illustrate the merits of the proposed features we introduce three new "appearance-impaired" datasets. Our experiments demonstrate the benefits of incorporating dynamics-based information into re-identification algorithms.

1 Introduction

The problem of human re-identification (re-id) is essential to visual surveillance, especially when the cameras have little or no overlapping field of view [3, 6, 7, 13, 14, 23, 39]. This is challenging because the targets often have significant variations in appearance, caused by changes in illumination, viewpoint and pose. Furthermore, in some scenarios, targets may reappear a few days later wearing different clothes. Some approaches address these challenges by learning a mapping function such that the distances between features from the same person are relatively small, while those from different persons are relatively large [17, 20, 30, 38, 42]. Other approaches focus on designing robust and invariant descriptors to better represent the subjects across views [2, 4, 24, 25, 40], or fuse different appearance features, [11, 29, 41], to achieve state-of-the-art performance on several benchmark datasets.

Pages 48.1-48.14
DOI: https://dx.doi.org/10.5244/C.30.48



(a) (b)

Figure 1: (a) Examples of images of the same person but wearing different clothing. Each column of images shows the same person; (b) Examples of persons wearing black suits. The first row was collected from the iLIDSVID dataset and the second row was collected from the PRID dataset.

Currently, most re-id methods rely on appearance-based features such as color and texture statistics, which are extracted from either a single image or a small set of images of the target. However, these features may be misleading rather than informative when matching images of the same person wearing different clothing (Figure 1(a)) or distinguishing people in similar clothes (Figure 1(b)) .

In real surveillance applications, the vision system is able to track the individuals for a while [8, 9, 19], providing useful temporal/dynamic information. Yet, very few approaches take advantage of this capability. In this paper, we propose an approach that exploits this information to capture soft bio-metrics such as gait and motion patterns by using Fisher vector encoding of temporal pyramids of dense short trajectories.

In order to evaluate the benefits of the proposed features, we compiled three new challenging "appearance impaired" re-id datasets. Two of these are subsets of the iLIDSVID and PRID2011 datasets, and are entirely comprised of videos with people wearing black clothes. The third set, collected by us, was captured by surveillance cameras from a train station. This set contains video sequences where the same people appear wearing different clothing and accessories. Our experiments on the full standard re-id datasets as well as the appearance impaired scenarios show that combining the proposed features with existing appearance-based features improves the re-id performance in overall, and specially on appearance impaired sequences.

The main contributions of this paper are: (i) A novel dynamics-based and Fisher vector encoded feature DynFV for re-id. The proposed feature captures subtle motion patterns to aid re-id, in particular in appearance-impaired scenarios; (ii) Three new challenging "appearance impaired" datasets for re-id performance evaluation; and (iii) A comprehensive evaluation of the effect of choosing different spatio, spatio-temporal, and dynamics-based features on the performance of (unsupervised) re-id methods.

1.1 Related Work: Re-id Approaches

Appearance-based features have been widely used in person re-identification studies [6]. Xiong *et al*. [38] used color histograms and LBP features and evaluated the effect of different spatial splitting schema. Covariance matrices were used as features by [2, 25]. Zhao *et al*. [39] used SIFT features to introduce a re-id method based on salience parts. Bazzani *et al*. [4] proposed a global appearance model with three complementary visual features. On the other hand, the use of temporal information in the re-id literature is very limited. Gheis-

sari *et al*. [12] incorporated temporal information to extract salient edge features. Bak *et al*. [3] used tracking trajectories to compensate for viewpoint variance. Bedagkar-Gala *et al*. [5] investigated two different gait features and demonstrated the effectiveness of fusion gait feature with color features in long term re-id. Kawai *et al*. [15] applied spatio-temporal HOG (STHOG) features to represent motion information. Although it did not require background subtraction, STHOG features are not viewpoint invariant. Wang *et al*. [36] split the video sequences into several candidate fragments based on motion energy, in which the most discriminating fragments were selected and ranked simultaneously by a multi-ranking method. Liu *et al*. [22] achieved state-of-the-art performance with an improved temporal cycle, semantic spatial segmentation and a modified LDFV feature [24]. However, none of these methods explicitly captured the underlying dynamics contained in the temporal sequences.

Dense trajectories capturing temporal information and moving patterns [35] have been proved to be powerful features for activity recognition [31, 37]. More recently, Li *et al*. [18], while addressing the problem of cross-view activity recognition, proposed to encode dense trajectories using Hankelet descriptors. There, they showed that Hankelets carry useful viewpoint and initial condition invariant properties.

Re-identification performance can also be improved by using better ways to compare or classify the features being used. For example, [42] proposed a relative distance comparison model to maximize the likelihood of distances between true matches being smaller than distances between wrong matches. In [38], Xiong *et al*. reported a comprehensive evaluation on several metric learning algorithms and presented extensions to PCCA [27] using a regularized term, and to LFDA [30] using a kernel trick. More recently, [21] applied cross-view quadratic discriminant analysis (XQDA) to learn the metric on a more discriminative subspace.

1.2 Related Work: Fisher Vector Encoding

In [18], Hankelet descriptors were encoded with Bags-Of-Words (BOW), which has been shown to be a sub-optimal encoding method [37]. On the other hand, Fisher vector encoding methods combining discriminating descriptors with generative models [32] have achieved excellent recognition performance. For example, [37] showed that Fisher vectors are one of the best encoding methods for activity recognition while [31] showed that a two-layer Fisher vector incorporating mid-level Fisher vectors representing semantic information achieved even better performance. As shown in our experiments, using this method allows us to aggregate multiple dynamic (and spatial) features in an effective way. Next, for the sake of completeness, we briefly summarize the main concepts of Fisher vector encoding.

Let $X = \{x_1, x_2, ..., x_N\}$ be a set of feature vectors that can be modeled by a distribution $p(X|\theta)$ with parameters θ. Then, this set can be represented by the gradient vector of the log-likelihood w.r.t. the parameters θ. Following the assumptions in [32], $p(X|\theta)$ is modeled using a Gaussian mixture model (GMM) with $\theta = \{\pi_1, \mu_1, \Sigma_1, ..., \pi_K, \mu_K, \Sigma_K\}$, where K is the number of mixture models and π_k, μ_k, and Σ_k are the mixture weight, mean and covariance of Gaussian k. Assuming all covariances are diagonal matrices $\Sigma_k = \sigma_k I$, X can be encoded by equations (1):

$$\begin{aligned} F^X_{\mu,k} &= \frac{1}{N\sqrt{\pi_k}} \sum_{n=1}^{N} \gamma_n(k) \left(\frac{x_n - \mu_k}{\sigma_k} \right), \\ F^X_{\sigma,k} &= \frac{1}{N\sqrt{2\pi_k}} \sum_{n=1}^{N} \gamma_n(k) \left\{ \frac{(x_n - \mu_k)^2}{{\sigma_k}^2} - 1 \right\}, \end{aligned} \tag{1}$$

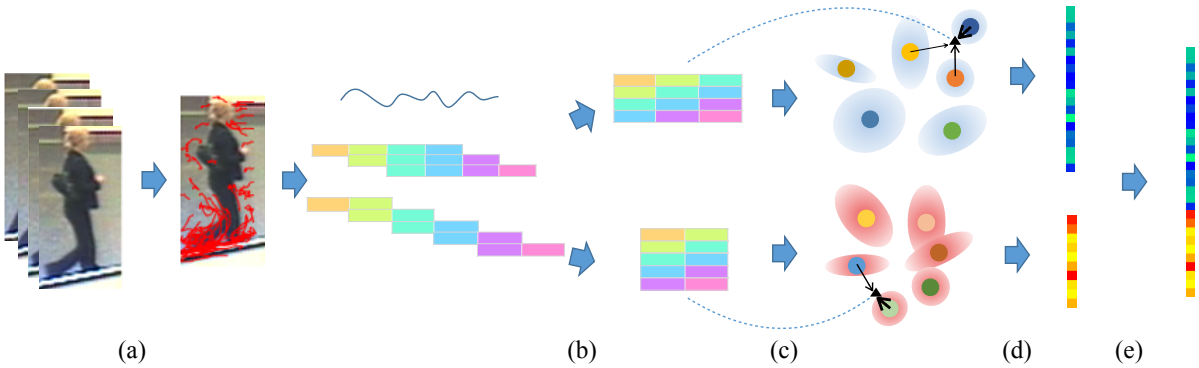


Figure 2: Pipeline of the proposed dynamics-based feature extraction. (a) Dense trajectories are extracted from video sequences and divided into small grids. (b) Temporal pyramids of the original trajectories are built using sliding windows of different sizes. (c) A GMM model is learned for each level of the pyramid. (d) The trajectories at each level of the pyramid are encoded using Fisher vectors based on the corresponding GMM. (e) The Fisher vectors at all scales are pooled to obtain the final feature vector.

where $\gamma_n(k)$ is the posterior probability of x_n given Gaussian model k. Then, the Fisher vector of set X will be the concatenation of $F^X_{\mu,k\in\{1,2,\ldots K\}}$ and $F^X_{\sigma,k\in\{1,2,\ldots K\}}$ along all K models.

2 The DynFV Feature

One of the main objectives of this paper is to address the problem of re-identification in appearance impaired scenarios such as the ones illustrated in Figures 1(a) and (b). In such cases, gait and idiosyncratic motion patterns offer a natural complementary source of information that is not affected by the lack of discriminating appearance-based features.

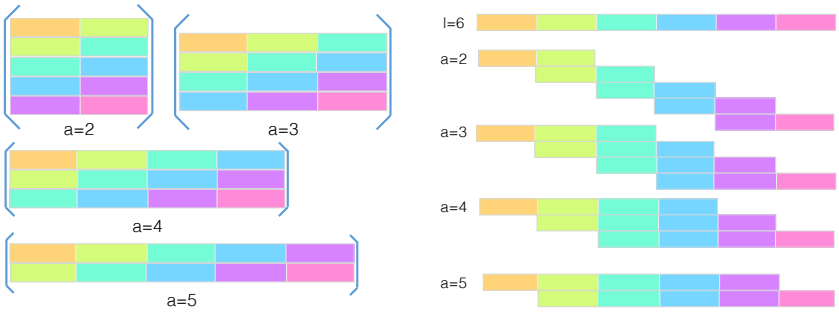
However, reliable estimation of motion-based biometrics, such as gait, is very challenging in crowded surveillance videos. In particular, it is very difficult to locate and consistently track the joints of the targets which would be required for model-based gait extraction [33]. Although researchers recently proposed model-free gait estimation [26, 34], they require explicit contour/silhouette extraction for each frame, which is difficult to obtain in low quality and crowded videos [36]. Because of this, we propose to use instead soft-biometric characteristics provided by *sets of dense, short trajectories (tracklets)*, which have been shown to carry useful invariants [18].

A potential drawback of using dense tracklets is that there are many of them and that they can exhibit large variability. Thus, it is important to have an effective way to aggregate the information they could provide. Towards this goal, we propose to use *pyramids of dense trajectories* with *Fisher vector encoding*, as illustrated in Figure 2 and described in detail below.

2.1 Temporal Pyramids of Dense Trajectories

Given a short tracklet $(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_N)$, its Hankelet [18] is defined as the Hankel matrix:

$$\mathbf{H_Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 & \ldots & \mathbf{Z}_k \\ \mathbf{Z}_2 & \mathbf{Z}_3 & \ldots & \mathbf{Z}_{k+1} \\ \vdots & \vdots & \ldots & \vdots \\ \mathbf{Z}_l & \mathbf{Z}_{l+1} & \ldots & \mathbf{Z}_N, \end{bmatrix}, \tag{2}$$



(a) (b)

Figure 3: (a) Each trajectory of length l is associated a set of Hankelets, with increasing number of columns, where the rows are obtained by splitting the trajectory into shorter tracklets of length a, using a sliding window with full overlap (stride of 1), as shown in (b).

where the constant off-diagonal structure of the matrix (depicted by elements painted with the same color in Figure 3(a)) captures the dynamics of the data and carries properties that are invariant to viewpoint changes and initial conditions [18]. The rank of the Hankelet measures the complexity of the dynamics of the data, and the more complex the dynamics, the more columns it can have before becoming column rank deficient.

Here, we propose to build *a family* of Hankelets for each trajectory (with an increasing number of columns a) to capture possibly different dynamic complexities. The rows of these Hankelets are obtained by splitting each trajectory of length l into shorter and shorter tracklets, using a sliding window with full overlap (stride of one) as illustrated in Figure 3.

Intuitively, the rows of the Hankelets, i.e. short tracklets of increasing length, constitute a set of *temporal pyramids* capturing different levels of dynamic complexity (since higher order dynamics are represented using bigger Hankelets/longer tracklets). Since different human body parts may have different dynamics, we split the bounding box of the person into G grids and process each grid separately. Then, the temporal pyramid of the dynamics-based features is built and Fisher vector encoded (see Figure 2) using Algorithm 1.

Algorithm 1 DynFV feature extraction

Require: number of grids G, length of tracklets l, set of temporal window size $\mathcal{A}$

for grid $g = 1$ to G **do**
  Extract dense tracklets Z^g of length l in current grid[1]
  Compute velocity vector V^g
  for $a_i \in \mathcal{A}$ **do**
    Extract $l - a_i$ shorter velocity vectors for each vector in V^g with length a_i and stride of 1
    Estimate a GMM model P for each velocity vector set ${V_{a_i}}^g$
    Compute Fisher vector ${F_P}^{{V_{a_i}}^g}$
    Apply power normalization, followed by L2 normalization
  end for
end for
Pool Fisher vectors by concatenating ${F_P}^{{V_{a_i}}^g}$ along all grids

3 Performance Evaluation

The second major objective of this paper is to present a thorough performance evaluation of different features used in multi-shot unsupervised re-identification systems. We evaluated the performance based on different combinations of features using five datasets. Two of these datasets are standard in the re-id literature: iLIDSVID and PRID 2011. In addition to these, we also used three new challenging datasets that we compiled to better evaluate the re-id performance in "appearance impaired" scenarios.

3.1 Datasets and Experiment Protocol

`Standard datasets:` The **iLIDSVID** dataset [36] is a random collection of 300 persons from the iLIDS Multiple-Camera Tracking Scenario [1]. For each person there are two cropped (64×128 pixels/frame) image sequences and the lengths vary from 23 to 192 frames. The **PRID 2011** dataset [13] consists of cropped (64×128 pixels/frame) sequences of 385 persons from camera A and 749 persons from camera B. To be consistent with previous work, as proposed by [36], we only use sequences of 178 persons that have more than 21 frames.

`Appearance-Impaired datasets:` To illustrate the need for dynamic-based features we collected three more challenging "appearance-impaired" datasets. Two of them consist of video sequences of people wearing black/dark clothing. They are subsets of the iLIDS-VID and PRID 2011 datasets and we named them **iLIDSVID BK** and **PRID 2011 BK**, respectively. The third dataset, named the **Train Station dataset (TSD)**, has sequences of persons with different clothing and accessories. The BK extension datasets were collected from the original datasets by manually selecting persons wearing black clothing. We collected 97 and 35 identities for iLIDSVID BK and PRID BK, respectively. **The TSD dataset** was collected with a single HD surveillance camera mounted at a public train station. Figure 1(a) shows sample frames from this set. The dataset has 81 sequences, including 9 targets with 3 sequences wearing different clothing and 54 sequences of randomly selected distractors. The length of the sequences vary from 41 to 451 frames with frames normalized to 64×128 pixels. While all the sequences were captured by the same camera, the relative viewpoint varies significantly when persons enter, re-enter and exit the scene.

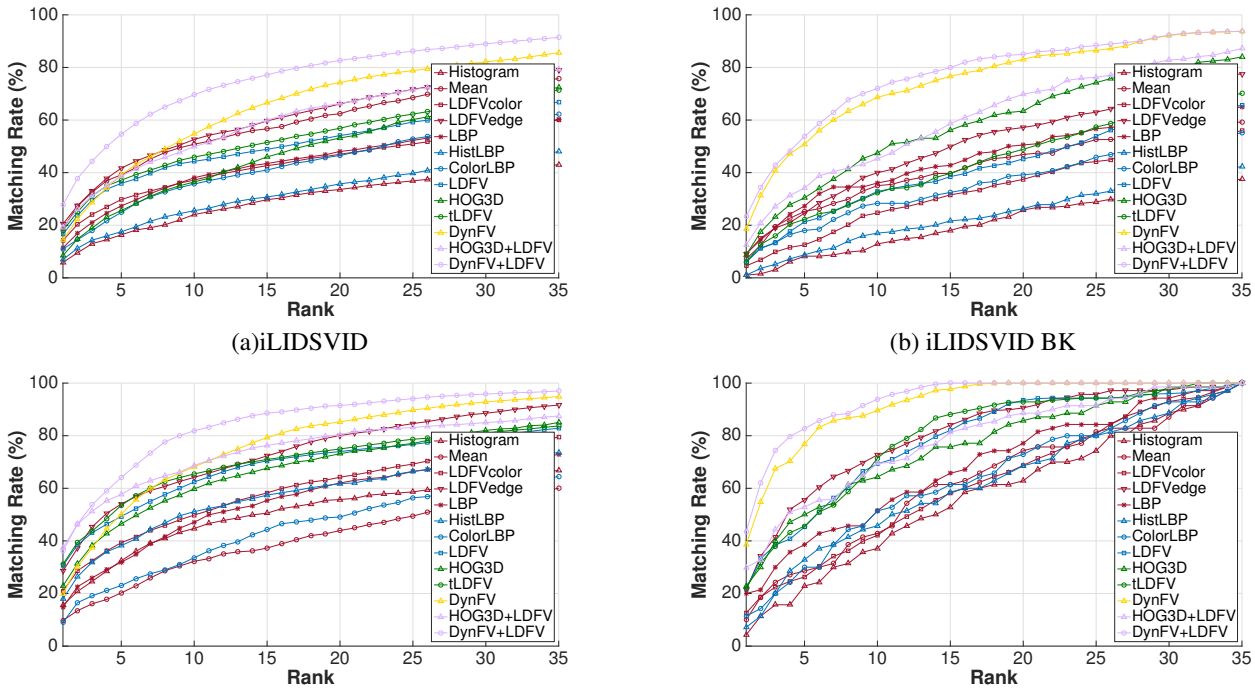
`Experimental protocol:` To evaluate the merits of the proposed dynamic based features, we *do not* apply any supervised metric learning method in the experiments. All ranking results are obtained directly by using the Euclidean distance between feature vectors. For a fair comparison, we only use the training data to learn the GMM model for the dynamic features (DynFV) and the local descriptor (LDFV) features. For the iLIDSVID and PRID datasets, we follow the protocol in [36][2]. For the BK extension datasets, because the sizes of the datasets are fairly small, we randomly pick the same size of training data in the non-BK part (i.e. 89 and 150 persons for PRID and iLIDSVID respectively) and run two experiments while a different camera is fixed as the probe set. For the TSD dataset, we only use the distractors to learn the GMM model. During the testing, we randomly pick one sequence for each target combined with all distractors to form the gallery set. This procedure is repeated 10 times. All experiments for DynFV and LDFV are repeated 10 times to remove the uncertainty in the GMM learning step. We *do not* need any ground truth in the feature extraction step.

[1] Location of each tracklet is determined by its starting point.

[2] We directly use the partition file from the project page of [36].

3.2 Features

We compare unsupervised re-id performance when using different combinations of features. We use six different types of features. Three of them are *purely spatial* features: Local Descriptors encoded by Fisher Vector (LDFV) [24], Color & LBP [14] and Hist & LBP [38]; two are *mixed local spatio-temporal* features: histogram of Gradients 3D (HOG3D) [16] and temporal LDFV (tLDFV) [22]; and one is our proposed *purely dynamics-based feature*: Dynamics-based features Fisher Vector encoded (DynFV). In all cases, before extracting the features, every frame is normalized to 64×128 pixels. In order to conduct a fair comparison, we use the same spatial split for tLDFV and LDFV and applied mean pooling along the temporal dimension for all features, except DynFV, before comparing. To get DynFV feature, we extract sets of dense trajectories (15 frames long) using the code provided by [35], in 18, 32×36 grids with 50% overlap. We set window size $\mathcal{A} = \{5, 9, 14\}$ and $K = 12$. To remove the impact of the background, a mask was applied to each frame. The mask was learned for each camera separately, as follows. First, semantic edges were obtained for each frame from the training sequences using a structured forest [10]. Then, the resulting images were averaged followed by a thresholding step to obtain a region covering semantic edges with high average scores. When we combine different features, a *simple* score-level fusion is applied. Source code is available online at http://robustsystems.coe.neu.edu and one can find all parameters in detail there.



(a)iLIDSVID (b) iLIDSVID BK

(c) PRID (d) PRID BK

Figure 4: CMC curves for iLIDSVID, PRID and the BK extension datasets

4 Experiments and Results

Next we present a series of experiments and discuss the results. In all cases we evaluate performance by comparing ranking scores. Furthermore, for the analysis of the merits of the features we also give Cumulative Match Curve (CMC) plots and report the Proportion of Uncertainty Removed (PUR) [30] score.

4.1 Feature Analysis

In these experiments we studied how different features affect re-id performance. The results are reported in Table 1 to 3, where each row shows the performance when using a different (sub)set of features and each column shows the performance of the same rank or PUR. Features are grouped as follows. Rows 1-5: a *single* component (i.e. color *or* texture) of a spatial feature; rows 6-8: *multiple* spatial features; rows 9-11: features that incorporate temporal information; and rows 11-12: combinations of spatial and temporal features. The best performance in each group is shown in bold.

Table 1: Results for iLIDSVID dataset and iLIDSVID BK dataset

	iLIDSVID					iLIDSVID BK				
Feature	1	5	10	20	PUR	1	5	10	20	PUR
Histogram[38]	5.9	16.3	24.1	33.6	6.5	1.0	8.2	12.9	25.8	7.3
Mean[14]	18.7	39.2	50.8	62.4	23.8	7.2	25.3	35.1	46.9	11.4
LDFV-color[24]	14.2	29.8	37.2	47.0	12.6	4.6	12.5	24.7	37.5	5.8
LDFV-edge[24]	**20.5**	**41.7**	**52.6**	**66.1**	**25.0**	9.0	24.6	**40.0**	**57.2**	12.8
LBP[28]	11.4	27.3	38.1	48.1	12.3	**9.3**	**27.3**	36.1	50.5	**15.1**
HistLBP[38]	6.9	17.5	25.5	35.7	7.3	1.0	8.8	17.0	26.3	5.9
ColorLBP[14]	10.9	24.9	35.7	46.5	14.2	**6.2**	18.0	28.4	39.2	8.8
LDFV	**16.5**	**35.6**	**44.2**	**54.3**	**17.8**	6.0	**21.3**	**32.8**	**45.3**	**9.9**
HOG3D[16]	8.5	25.6	36.5	53.2	15.9	8.8	30.4	47.4	63.4	20.9
tLDFV[22]	**18.0**	37.1	45.9	56.9	19.7	5.8	22.3	32.4	48.9	10.4
DynFV (Ours)	14.4	**39.2**	**54.7**	**74.3**	**26.7**	**18.6**	**50.8**	**68.7**	**83.1**	**30.6**
HOG3D+LDFV	19.2	38.6	49.9	66.3	24.1	12.6	34.2	45.3	69.8	21.4
DynFV+LDFV	**27.8**	**54.6**	**69.6**	**82.7**	**37.1**	**23.4**	**53.8**	**72.0**	**85.0**	**33.0**

Table 2: Results for PRID dataset and PRID BK dataset

	PRID					PRID BK				
Feature	1	5	10	20	PUR	1	5	10	20	PUR
Histogram[38]	15.6	32.5	44.7	55.7	12.2	4.3	22.9	37.1	62.9	7.7
Mean[14]	9.8	20.2	32.2	43.9	6.7	10.0	28.6	42.9	74.3	11.5
LDFV-color[24]	20.9	39.3	49.8	64.2	18.0	12.6	26.3	42.1	68.7	4.5
LDFV-edge[24]	**28.6**	**54.0**	**64.2**	**80.1**	**29.5**	**21.4**	**55.6**	**72.7**	**90.6**	**20.6**
LBP[28]	14.8	31.9	47.1	62.0	13.7	20.0	38.6	51.4	77.1	16.7
HistLBP[38]	18.0	38.2	51.1	61.8	16.1	7.1	32.9	45.7	68.6	7.0
ColorLBP[14]	9.0	23.0	33.6	49.1	8.5	11.4	30.0	51.4	72.9	11.4
LDFV[24]	**30.7**	**49.3**	**62.4**	**74.0**	**26.6**	**21.9**	**45.3**	**69.3**	**93.4**	**18.5**
HOG3D[16]	22.9	46.5	59.8	73.1	23.5	22.9	50.0	64.3	85.7	20.6
tLDFV[22]	**31.2**	**53.5**	65.5	75.0	28.4	22.0	45.6	71.4	92.9	20.2
DynFV (Ours)	19.9	50.2	**68.2**	**85.3**	**29.4**	**38.6**	**76.9**	**89.6**	**100.0**	**41.6**
HOG3D+LDFV	**37.4**	57.7	68.7	80.4	34.0	29.9	52.9	69.4	88.6	24.8
DynFV+LDFV	36.2	**64.1**	**81.8**	**91.5**	**41.4**	**43.6**	**82.7**	**93.9**	**100.0**	**47.4**

iLIDSVID (BK), PRID (BK) Sets: Figure 4 shows the CMC curves and Tables 1, 2 show the re-id performance scores for the iLIDSVID and iLIDSVID BK, and for the PRID and PRID BK datasets, respectively.

Spatial-based Features-Components: Color mean achieves the best performance among the three color features for the iLIDSVID, iLIDSVID BK and PRID BK datasets, but LDFV-edge outperforms color and LBP in the iLIDSVID, PRID and PRID BK datasets on rank-1 identification rate.

Spatial Features: Among all these features, LDFV gives the best performance. The reason is two-fold. First, in general, Fisher vector encoding performs better than average

pooling and histogram; and second, because the data consists of multiple frames, LDFV has many samples to get better estimates of the underlying GMM. As expected, the performances for all the spatial features, except LBP and ColorLBP in PRID BK dataset, decrease notably when the features are used on the appearance impaired datasets, especially considering that these datasets have smaller galleries. This situation also holds for the spatial features-components group.

Temporal information incorporated Features: HOG3D is a spatial *and* temporal based feature. Its performance is significantly lower in comparison with purely spatial features in the standard datasets. However, its performance does not degrade when used in the BK sets. These results suggest that the temporal component of this feature helps distinguishing different targets with similar appearance. It should be noted that we obtain better accuracy using HOG3D in the PRID dataset than the results reported in Table 4 of [36]. A reason for this is that instead of using only 4 uniformly sampled candidate fragments, we use all HOG3D features from dense sampled cells to do average pooling, which provide more stable and less noisy features. tLDFV incorporates local temporal information to the original LDFV, thus it gives slightly better results among all datasets. However, the matching accuracy significantly drops on the BK extension datasets, indicating that tLDFV highly relies on the spatial part. In the original datasets, DynFV has rank-1 performance worse than tLDFV but similar to HistLBP and HOG3D and better or similar PUR performance among all single type features, which is remarkable since DynFV does not use any type of spatial information. In the appearance impaired sets, the DynFV significantly outperforms all spatial features and is almost twice better than HOG3D and tLDFV, which illustrates the merit of using dynamic information in this type of scenario.

Feature Fusion: The last two rows show that the performance improves when using together spatial LDFV and temporal features together. As seen in both tables, joint use of DynFV and LDFV gives much better results in the original iLIDVID and PRID datasets. More precisely, using this combination provides a relative improvement of 108.4% and 55.6% at PUR performance with respect to using LDFV alone. These results show that DynFV can be used as a powerful complementary feature in video sequences-based re-id. On the other hand, including dynamic features in the impaired datasets increases the performance considerably. At the rank-1 performance, it has a relative improvement of 290.0% and 99.1%, respectively, which displays the advantage of DynFV under appearance impaired scenarios.

Table 3: Results for TSD dataset.

	TSD				
Feature	1	5	10	20	PUR
Histogram[38]	35.6	56.1	69.4	75.0	35.9
Mean[14]	13.9	34.4	41.7	64.4	17.4
LDFV-color[24]	47.1	**80.8**	**89.4**	97.2	53.0
LDFV-edge[24]	**53.6**	71.7	**89.4**	**99.4**	**54.0**
LBP[28]	19.4	56.1	61.7	72.8	27.5
HistLBP[38]	35.6	56.1	69.4	75.0	36.6
ColorLBP[14]	13.9	33.3	40.6	63.3	18.1
LDFV[24]	**44.5**	**79.2**	**95.2**	**99.4**	**54.7**
HOG3D[16]	40.0	63.3	73.9	90.6	42.4
tLDFV[22]	44.3	**78.9**	**96.9**	**99.4**	**55.5**
DynFV (Ours)	**45.7**	72.7	89.9	92.2	48.5
HOG3D+LDFV	53.2	82.8	90.2	**99.7**	59.4
DynFV+LDFV	**61.0**	**83.6**	**95.1**	98.5	**61.3**

Table 4: Results with different window sizes on PRID BK dataset

a_i	1	5	10	20	PUR
5	36.0	73.7	**90.9**	**100.0**	40.0
9	38.4	**77.6**	88.7	98.6	40.3
14	31.3	72.9	89.9	98.9	35.9
5,9,14	**38.6**	76.9	89.6	**100.0**	**41.6**

TSD Set: Table 3 shows the performance scores for the TSD dataset. Since several targets in the dataset only partially change their appearance, spatial features perform fairly well here. Though LDFV-edge performs the best among all single type features, DynFV feature still provides significant complementary information. Combining LDFV with DynFV features leads to a relative improvement of 37.1% at rank-1 accuracy.

4.2 Effect of Length of Sliding Window

This experiment evaluates the impact of the size of the sliding window used to generate the pyramid of dense trajectories for DynFV. The first three rows in Table 4 show that, instead of full length trajectories ($a_i = 14$), the shorter, fully overlapped trajectories provide significant performance gain at rank-1 accuracy. In particular, setting $a_i = 5$ has a relative increase of 15% and $a_i = 9$ has 22.7%. After combining all three pyramid levels, we achieved the best performance both on rank-1 accuracy and PUR score.

5 Conclusion and Future Work

Until now, most re-id state-of-the-art approaches relied on appearance-based features, extracted from a single or a few images. These approaches do not use the videos which are typically available in surveillance applications and are at a disadvantage in appearance impaired scenarios. In this paper, we proposed DynFV features to address these limitations and introduced three new challenging appearance impaired re-id datasets. The proposed DynFV feature exploits soft-biometrics, encapsulated in short dense trajectories associated with the targets and benefits from the powerful Fisher vector encoding method. Our extensive experiments show that DynFV features carry complementary information to previously used features and that combining them with the state-of-art LDFV features, results in a relative performance improvement at rank-1, compared against using LDFV alone, of 68% and 18% for the iLIDSVID and PRID datasets, respectively, and of 290%, 99%, and 37% for the appearance deprived scenarios in the BK and TSD datasets, respectively. In the future, we will investigate more intelligent feature fusion methods and incorporate supervised learning to improve the performance.

6 Acknowledgement

This work was supported in part by NSF grants IIS–1318145 and ECCS–1404163; AFOSR grant FA9550-15-1-0392; and the Alert DHS Center of Excellence under Award Number 2013-ST-061-ED0001.

References

[1] Uk home office: i-lids multiple camera tracking scenario definition. 2008.

[2] Slawomir Bak, Etienne Corvee, Francois Brémond, and Monique Thonnat. 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, 2010.

[3] Slawomir Bak, Sofia Zaidenberg, Bernard Boulay, and François Bremond. Improving person re-identification by viewpoint cues. In *Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on*, pages 175–180. IEEE, 2014.

[4] Loris Bazzani, Marco Cristani, and Vittorio Murino. Sdalf: Modeling human appearance with symmetry-driven accumulation of local features. In *Person Re-Identification*, pages 43–69. Springer, 2014.

[5] Apurva Bedagkar-Gala and Shishir K Shah. Gait-assisted person re-identification in wide area surveillance. In *Asian Conference on Computer Vision*, pages 633–649. Springer, 2014.

[6] Apurva Bedagkar-Gala and Shishir K Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286, 2014.

[7] Alina Bialkowski, Simon Denman, Patrick Lucey, Sridha Sridharan, and Clinton B Fookes. A database for person re-identification in multi-camera surveillance networks. In *Proceedings of the 2012 International Conference on Digital Image Computing Techniques and Applications (DICTA 12)*, pages 1–8. IEEE, 2012.

[8] O. Camps, M. Gou, T. Hebble, S. Karanam, O. Lehmann, Y. Li, R. Radke, Z. Wu, and F. Xiong. From the lab to the real world: Re-identification in an airport camera network. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99): 1–1, 2016. ISSN 1051-8215. doi: 10.1109/TCSVT.2016.2556538.

[9] Caglayan Dicle, Octavia I Camps, and Mario Sznaier. The way they move: Tracking multiple targets with similar appearance. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2304–2311. IEEE, 2013.

[10] Piotr Dollár and C. Lawrence Zitnick. Structured forests for fast edge detection. In *ICCV*, 2013.

[11] M. Eisenbach, A. Kolarow, A. Vorndran, J. Niebling, and H.-M. Gross. Evaluation of multi feature fusion at score-level for appearance-based person re-identification. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–8, July 2015. doi: 10.1109/IJCNN.2015.7280360.

[12] Niloofar Gheissari, Thomas B Sebastian, and Richard Hartley. Person reidentification using spatiotemporal appearance. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1528–1535. IEEE, 2006.

[13] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. Scandinavian Conference on Image Analysis (SCIA)*, 2011. The original publication is available at www.springerlink.com.

[14] Martin Hirzer, Peter M Roth, Martin Köstinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In *Computer Vision–ECCV 2012*, pages 780–793. Springer, 2012.

[15] Ryo Kawai, Yasushi Makihara, Chunsheng Hua, Haruyuki Iwama, and Yasushi Yagi. Person re-identification using view-dependent score-level fusion of gait and color features. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2694–2697. IEEE, 2012.

[16] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.

[17] Martin Kostinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE, 2012.

[18] Binlong Li, Octavia I Camps, and Mario Sznaier. Cross-view activity recognition using hankelets. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1362–1369. IEEE, 2012.

[19] Yang Li, Ziyan Wu, Srikrishna Karanam, and Richard J Radke. Real-world re-identification in an airport camera network. In *Proceedings of the International Conference on Distributed Smart Cameras*, page 35. ACM, 2014.

[20] Yang Li, Ziyan Wu, S. Karanam, and R.J. Radke. Multi-shot human re-identification using adaptive fisher discriminant analysis. In *BMVC 2015-26th British Machine Vision Conference*. British Machine Vision Association, 2015.

[21] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan 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*, pages 2197–2206, 2015.

[22] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3810–3818, 2015.

[23] Xiaokai Liu, Hongyu Wang, Yi Wu, Jimei Yang, and Ming-Hsuan Yang. An ensemble color model for human re-identification. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 868–875. IEEE, 2015.

[24] Bingpeng Ma, Yu Su, and Frédéric Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 413–422. Springer, 2012.

[25] Bingpeng Ma, Yu Su, and Frédéric Jurie. Bicov: a novel image representation for person re-identification and face verification. In *British Machive Vision Conference*, pages 11–pages, 2012.

[26] Ju Man and Bir Bhanu. Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322, 2006.

[27] Alexis Mignon and Frédéric Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2666–2672. IEEE, 2012.

[28] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[29] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton 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.

[30] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3318–3325. IEEE, 2013.

[31] Xiaojiang Peng, Changqing Zou, Yu Qiao, and Qiang Peng. Action recognition with stacked fisher vectors. In *Computer Vision–ECCV 2014*, pages 581–595. Springer, 2014.

[32] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *Computer Vision–ECCV 2010*, pages 143–156. Springer, 2010.

[33] David K Wagg and Mark S Nixon. On automated model-based extraction and analysis of gait. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 11–16. IEEE, 2004.

[34] Chen Wang, Junping Zhang, Liang Wang, Jian Pu, and Xiaoru Yuan. Human identification using temporal information preserving gait template. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2164–2176, 2012.

[35] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[36] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *Computer Vision–ECCV 2014*, pages 688–703. Springer, 2014.

[37] Xingxing Wang, LiMin Wang, and Yu Qiao. A comparative study of encoding, pooling and normalization methods for action recognition. In *Computer Vision–ACCV 2012*, pages 572–585. Springer, 2013.

[38] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *Computer Vision–ECCV 2014*, pages 1–16. Springer, 2014.

[39] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person re-identification by salience matching. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2528–2535. IEEE, 2013.

[40] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 144–151. IEEE, 2014.

[41] Liang Zheng, Shengjin Wang, Lu Tian, Fei He, Ziqiong Liu, and Qi Tian. Query-adaptive late fusion for image search and person re-identification. In *Computer Vision and Pattern Recognition*, volume 1, 2015.

[42] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(3): 653–668, 2013.