Recent Trends in Person Re-Identification: An Overview

Perni Dedeepya[a], and R.Vani[b]
[a]
Research scholar, ECE Department, SRM IST Ramapuram
Chennai, India. dedeepyaperni@gmail.com
[b]Professor, ECE Department, SRM IST Ramapuram, Chennai, India
vanir@srmist.edu.in

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: Person re-identification has significant applications in real time scenarios like forensics and video surveillance. Most of the person re-identification models emphasize on image to image matching or video to video matching as they are easy because of the homogenous matching features. But, in real time applications there will be situations where image is to be matched to a video. This sort of re-identification process is called image to video person re-identification (IVPR), which has significant applications in tracking the location of a lost human and criminal tracking. This method of re-identification has many difficulties due to the heterogeneous matching to be done due to the variation in features of image and video. This article focuses on existing person re-identification methods, their difficulties and further area of research in this field

Keywords: Person re-identification, Image to video person re-identification, feature, video, image.

1. Introduction

Person Re-Identification focus on identifying a person from images/videos taken from different non-overlapped cameras, given a video/image of person taken from another camera. Because of its many applications in video surveillance, public safety-critical applications and forensics, person re-identification is gaining more and more attention. Applications of person re-identification include public security and safety in video surveillance. Although a large number of person re-identification models are existing, there are still a plenty of barriers to be resolved for real time applications, because of large variations in different non overlapped cameras, which occur due to differences in the illumination, poses, viewpoints, occlusions and cluttered background.

Based on the existing techniques of person re-identification, it can be divided broadly into two categories: Image based person re-identification and Video based person re-identification. For person re-identification if matching is done between image and image, it comes under the former category. Most of the existing re-identification methods fall under this category [1-10]. The later focuses on matching between two videos [11-16]. For both the categories the two things to be matched are of same type, so have similar features, hence referred to as homogeneous matching.

But, in many real life scenarios, the matching is to be done between image of a person and video that contains him. This is a special category of re-identification process known as Image to Video Person Re-identification (IVPR). As the video and image differ in its features it falls under the category of heterogeneous matching.

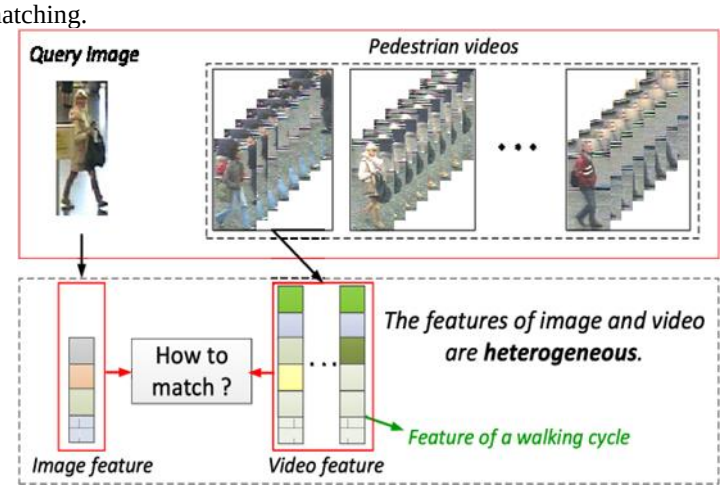

Fig. 1.Problem in matching an image to a video

Applications of IVPR involves, tracking a criminal from videos of surveillance using image of the criminal in the police records, identifying a Alzheimer patient who lost his way back to his house using his image previously taken.

Difficulties of IVPR are because of the heterogeneous matching to be done between an image and video. They include 1. Difference in feature i.e. image has only visual appearance feature while a video has both visual

appearance and spatial temporal feature. 2. Features extracted from each frame or walk cycle form a set while those extracted from an image form a point. Matching this point to set is difficult. 3. It is strenuous to effectively extract features from a video which has large noise in terms of unwanted things to match it with an image. 4. In each video there exists large variations between the frames resulting in an increased difficulty to match an image with video.

2. Existing Work/ Related Work

Existing works that focus on person re-identification are discussed in detail in this section. As already discussed person re-identification is classified into two categories along with the third complicated category. Related works in each category are as follows:

A. Image based person re-identification

These can be further classified into 1. Methods based on feature representation , in which the focus is mainly on the distinct feature to match the images. [17] Presents person re-identification based on appearance being considered as the feature for representation known as Symmetry-Driven Accumulation of Local Features (SDALF). For this two horizontal axes of symmetry that divide the body into three main regions head, torso and legs is found, and then a vertical axis of symmetry is estimated on the latter two. The complementary aspects regarding appearance of human body are detected on each part and are weighted by distance w.r.t vertical axis so as to minimise the effect of pose variations. Using these features for matching gives the similarity measure between the candidates in the image. This method can be applied to different modalities like single shot image or multiple shot frame and is observed to give highest performance on three promising public datasets VIPeR, iLIDS and ETHZ. In [18] patterns and colours on virtual bones are identified and spatially localized to exploit an articulated 3D body model. Despite of different postures and viewpoints this model creates a unique signature of each person being monitored.

A specific metric learning method is used to match the images. An improvement is observed by using this model on Microsoft Kinect dataset in comparison to a 2D body like SDALF. [19] Proposes person re-identification across non overlapping cameras based on clothing attributes. As there will be difficulty in identifying the faces in images taken with non-overlapping cameras clothing appearance can be used as clue for assisting in re-identification process. The large gap between low-level and high-level descriptors is bridged using middle level descriptors. A latent support vector machine describes the relation between part features which are low-level, clothing attributes that are middle level and labels of person pair which are high level descriptors. This method showed a significant improvement over sate of art methods when applied over VIPeR data set.

2. Methods based on distance learning, which focus mainly on calculating the metric based on optimal distance. [20] Introduces simple and effective distance learning metric using equivalence constraints for large data to reduce computational complexity. This is done based on likelihood ratio test using statistical inference perspective. This method shows outperforming results which are applied on VIPeR and Toycars datasets compared to existing sate of art methods and is also very faster in training images, as it involves just calculation of two covariance matrices which are very small in size. This is also referred to as Keep It Simple and Straight forward (KISS) metric. The above mentioned KISS metric when given to small size training set results in poor performance due to an instable estimation of inverse to the co-variance matrix. This is overcome in [21] where Regularized smoothing Keep It Simple and Straight forward (RS-KISS) metric learning is proposed. The overestimated Eigen values in former method during the calculation of covariance metric are reduced in an effective way in the latter method but at a cost of increase in time for computation. When applied on VIPeR it outperformed all the existing methods. An incremental learning to RS-KISS is also introduced in which the computational cost is reduced significantly.

B. Video based person re-identification

Most of the video based person re-identification methods depend on single frame features as in image and ignore the spatial information from image sequences available in video based scenarios. [12] Introduces a new model to select video fragments which are most discriminative from a noisy sequence of images. Most reliable space-time and appearance features are computed from the selected video fragments. A discriminative video ranking model is used to relax the assumptions made by gait recognition methods significantly. This is done by using HOG3D feature and optic flow energy profile. Using this method for PRID2011, iLIDS-VID and HDA+ image sequence datasets is considered advantageous over holistic image sequence matching, extant gait recognition, and state-of-the-art single-/multi-shot re-identification methods. Along with spatial alignment that is commonly used by treating independently the appearance of different body parts, to address the issues caused by difference in appearance of a person, due to difference in poses, illumination, viewpoints and occlusions [13] considers a temporal alignment problem to address the issues related to difference in appearance of a body part, during different phases of action. This approach exploits the periodicity exhibited by a walking person, taking the video sequence as input to generate a body action model based on spatio-temporal features. A series of body action units for different body parts are obtained based on certain action primitives. These body units are then used to train the vocabularies and fisher vectors are obtained. Finally, a fixed length feature vector is obtained by

concatenating the fisher vectors extracted from all the body action units. This final vector represents the appearance of the walking person.

Applying this method on iLIDS-VID and PRID2011 datasets, this method outperforms the previous methods which are based only on spatial alignment. Although video-based person re-identification is a homogenous matching, there is a difficulty that, there exist large variations between different pedestrian videos called inter video variations and also within each video called intra video variations. [14] Introduced a new method, simultaneous intra-video and inter-video distance learning (SI2DL) to address the above difficulty. This approach employs set based distance learning techniques which deals with intra and inter video variations, by learning a pair of intra and inter video metrics. Each video is made more compact by using intra video distance metric, while inter video distance metric can be used to reduce the metric between truly matching videos, compared to that between wrongly matching videos. A new video relationship model called video triplet is also designed, to intensify the performance of inter video metric.

This is designed using two truly matched videos and an impostor video. Using SI2DL method on iLIDS-VID and PRID2011 achieves the state-of-the-art performance. In video-based scenario some differences in inter class can be much more difficult to deal with, because there is a possibility for different people to have similar appearance, movements and actions which are very difficult for alignment. To address the above-mentioned problem [16] proposes a top push distance learning model (TDL). In this method feature representation is exploited by constituting HOG3D, average pooling of colour histogram and LBP features. Based on these intra class variations are minimized and top push distance constraint is realized to optimize a discriminative distance learning model. This is done by introducing top push into distance metric learning for person re-identification to increase the matching accuracy. Using this method on PRID2011 and iLIDS-VID datasets results in significant improvement on matching accuracy and outperforms existing methods related to distance/rank learning. [15] Introduces new recurrent neural network architecture for video based person re-identification. In this method the feature extractor is obtained by training jointly the convolution neural network, its recurrent final layer and temporal pooling layer using Siamese network architecture.

This is done by extracting features of person using a convolution neural network with recurrent final layer from each frame. These allow the information to be collected from all the time steps. The features so obtained are combined to give a complete appearance features for the overall video sequence using temporal pooling. In order to calculate motion information and appearance features this approach makes use of color and optical flow information. The existing methods of video-based re-identification are outperformed when the proposed method is applied on PRID2011 and iLIDS-VID.

C. Image to video based person re-identification

Image and video differ in various features; hence it is a challenging task to match an image with a video. [22] Proposes a joint feature projection matrix and heterogeneous dictionary pair learning (PHDL) approach for IVPR. This approach transforms the heterogeneous features of image and video into coding coefficients with the same dimensions for matching. To make this matching easier, intra-video variations are reduced using a feature projection matrix. To exploit the information contained in the video effectively a multi-view PHDL (MPHDL) is also proposed in which, different dictionaries and projection matrices are learned for video features of different kinds. Simulation results show that MPHDL can enhance the performance of PHDL when applied to datasets like iLIDS-VID, PRID2011, MARS and HDA+. Experimental results show a large matching rate than the challenging methods. [23] Addresses the IVPR problem in an end to end way by formulating a new framework.

A novel deep neural network is built to learn the representation of features and distance metric based on point to set using both image and video as input. To prevent manual selection of related frames from noisy frames a kNN-triplet unit i used. This kNN-triplet module makes the network focus successively on useful and important frames, while discarding the useless ones in the video. For evaluation of the above method new datasets modified from like iLIDS-VID, PRID2011 and MARS as like iLIDS-VID-P2S, PRID2011- P2S and MARS- P2S respectively are considered. It is observed that this model surpasses a number of methods designed alternatively for point to set problems based on face recognition and shows significant improvement over image-based methods applied on same problem.

IVPR methods are observed to have relatively lower performance as they only consider the measurement of similarity between appearance feature of an image and spatio-temporal feature of a video. But the fact is that information contained in the background of an image or video is useless for person re-identification. So, it is important to extract the discriminative features in image/video that are useful, without any unrelated information. [24] Introduces a feature extraction mechanism based on body part attention which focuses mainly on body parts such as torso, elbow, wrist, knee, ankle etc. This mechanism can extract body part attention-based appearance features and spatio-temporal features in image and video respectively by neglecting the useless area of image/video. This mechanism can be employed into CNN and LSTM networks to jointly learn attention coefficient and similarity measurements. This approach when used on iLIDS-VID and PRID2011, it is observed

to outperform all the state-of-the-art techniques and is also applicable to large data sets like MARS, CUHK03 and MARKET1501.

3. Result and Cooperative Analysis

This section includes comparison of some of the above techniques when applied to various datasets..

A. Image based person re-identification techniques

SDALF approach minimises the effects of the pose variations by accumulating the local features in the images and is simple and effective [17]. KISSME method is a simple and effective technique based on likelihood ratio with an optimization procedure which is iterative. This method can be applied to large datasets and involves fewer computations [20].

TABLE I. COMPARISON OF MATCHING RATES (%) FOR SDALF AND KISSME TECHNIQUES ON VIPER DATASET [20].

Rank	1	10	25	50
SDALF	20	50	70	85
KISSME	20	62	81	92

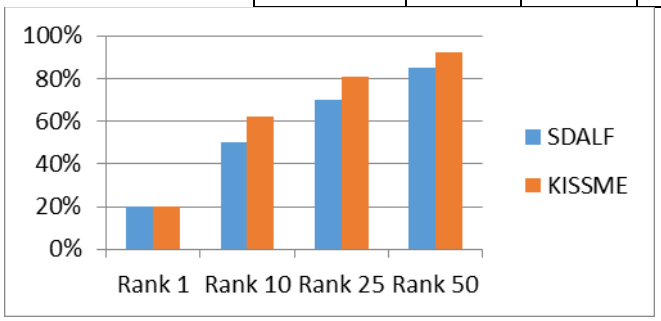

Fig. 2. Comparison of matching rates (%) for different ranks of SDALF and KISSME.

TABLE I and Fig. 2 gives the comparison of above two methods when applied on VIPeR dataset in the range of first 50 ranks. A competitive result is obtained across all the ranks and is observed that KISSME outperforms SDALF.

B. Video based person re-identification techniques

DVR is capable of discovering informative and reliable video fragments from incomplete and not so accurate image sequences. This technique significantly improves spatial appearance features of the video [12]. STFV3D aligns spatially and temporally the dynamic appearance of different people effectively [13]. SI2DL learns a pair of intra and inter video distance metrics. These make the video compact to represent it in a better way and can reduce metric value of correctly matched images compared to wrongly matched images [14].

TABLE II. MATCHING RATES (%) ON PRID2011DATASET [14].

Method/Rank	1	5	10	20
DVR	28.9	55.3	65.5	82.8
STFV3D	42.1	71.9	84.4	91.6
SI2DL	76.7	95.6	96.7	98.9

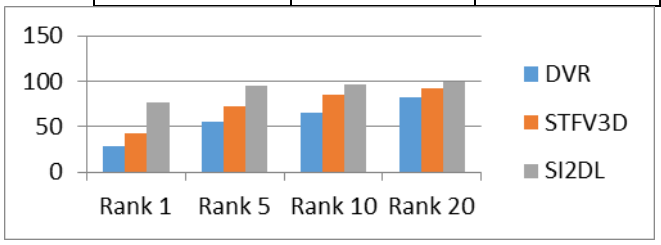

Fig. 3. Matching rates (%) on PRID2011 dataset for DVR, STFV3D and SI2DL

TABLE III. MATCHING RATES (%) ON ILIDS-VID DATASET [14].

Method/Rank	1	5	10	20
DVR	23.3	42.4	55.3	68.4
STFV3D	37.0	64.3	70.0	86.9
SI2DL	48.7	81.1	89.2	97.3

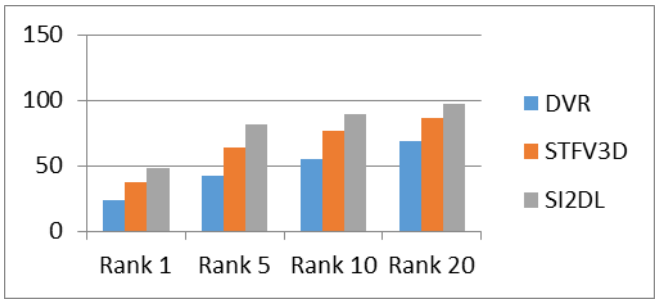

Fig. 4. Matching rates (%) on iLIDS-VID dataset for DVR, STFV3D and SI2DL.

TABLE II, TABLE III, Fig. 3 and Fig. 4 gives the comparison of the above three methods on PRID2011 and iLIDS-VID datasets. It is observed that STFV3D outperforms DVR because this method can bridge the gap between camera views due to the variation in color and view point. This is more effective in PRID2011 than iLIDS-VID because PRID2011 has significant color in consistency. iLIDS-VID dataset is challenging dye to the similarities in clothing style of people, lighting variations, variations in view point, occlusions and background cluttering. SI2DL is out performing both the above methods due to the reduction in intra video and inter video variations.

C. Image to video based person re-identification techniques

PHDL uses feature projection matrix to reduce intra video variations and heterogeneous dictionary pair to have favourable discriminability by making use of the important information in the video [22]. P2SNET builds and deep neural network with an end to end architecture, which combines feature learning and point to set distance metric. The outliers in the video are removed using kNN-triplet [23]. CBAN technique discards useless information and extracts cross media features that have important body part attention by CNN/LSTM from the image or video [24].

TABLE IV. COMPARATIVE RESULTS MEASURED BY RANK-M ACCURACY IN % OF PHDL, P2SNET AND CBAN ON MARS, ILIDS-VID AND PRID2011 DATASETS [24].

Method/dataset	**MARS**				**iLIDS-VID**				**PRID2011**			
	r-1	***r-5***	***r-10***	***r-20***	***r-1***	***r-5***	***r-10***	***r-20***	***r-1***	***r-5***	***r-10***	***r-20***
PHDL	35.7	51.5	60.9	67.3	28.1	50.3	65.8	80.3	41.9	67.2	85.4	90.0
P2SNET	55.3	72.9	78.7	83.7	40.0	68.5	78.1	90.0	73.3	90.5	94.7	97.8
CBAN	68.2	85.3	88.9	92.9	43.2	71.0	80.1	92.1	74.6	90.6	95.7	97.9

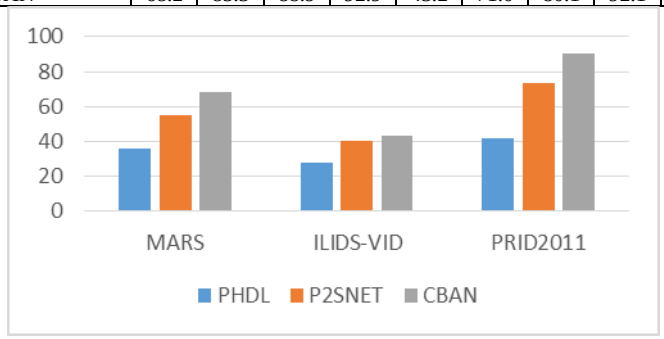

Fig. 5. Rank-1 ACCURACY IN % OF PHDL, P2SNET AND CBAN ON MARS, ILIDS-VID AND PRID2011 DATASETS.

Comparing the above three methods by applying to various datasets we obtain TABLE IV and Fig. 5. It is observed that CBAN achieves more accuracy due to the decrease in the large gap that exists in the cross media feature between image appearance feature and s[patio temporal feature of video.

4. Conclusion

Image to video person re-identification problem is disorganized because a person poses, appearance of the body, and deformation are more distinct than those of the face. Also, a pedestrian video has richer information than a face video because of the crowded public in the background. So, the models of face recognition based on point to set are not effective enough for this problem. Hence IVPR is one of the challenging issues with many applications in the real life scenario, an ongoing research topic with much more and more study to be done.

References

[1] J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in Proc. IEEE Conf. ICCV, Dec. 2013, pp. 3567–3574.

[2] C. Liu, C. C. Loy, S. Gong, and G. Wang, "POP: Person re-identification post-rank optimisation," in Proc. IEEE Conf. ICCV, Dec. 2013, pp. 441–448.

[3] Q. Qiu, J. Ni, and R. Chellappa, "Dictionary-based domain adaptation methods for the re-identification of faces," in Person Re-Identification. London, U.K.: Springer-Verlag, 2014, pp. 269–285.

[4] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in Proc. IEEE Conf. CVPR, Jun. 2015, pp. 3908–3916.

[5] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in Proc. IEEE Conf. CVPR, Jun. 2015, pp. 1565–1573.

[6] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in Proc. IEEE Conf. CVPR, Jun. 2014, pp. 3550–3557.

[7] X.-Y. Jing et al., "Super-resolution person re-identification with semicoupled low-rank discriminant dictionary learning," in Proc. IEEE Conf. CVPR, Jun. 2015, pp. 695–704.

[8] S. Li, M. Shao, and Y. Fu, "Cross-view projective dictionary learning for person re-identification," in Proc. IJCAI, 2015, pp. 2155–2161.

[9] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in Proc. IEEE Conf. ICCV Dec. 2015, pp. 1116–1124.

[10] Raja, S. Kanaga Suba, and T. Jebarajan. "Reliable and secured data transmission in wireless body area networks (WBAN)." European Journal of Scientific Research 82, no. 2 (2012): 173-184..

[11] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in Proc. ECCV, 2014, pp. 688–703.

[12] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by discriminative selection in video ranking," IEEE Trans. Pattern Anal Mach. Intell., vol. 38, no. 12, pp. 2501–2514, Dec. 2016.

[13] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for viceo-based pedestrian re-identification," in Proc.IEEE Conf. ICCV, Dec. 2015, pp. 3810–3818.

[14] X. Zhu, X.-Y. Jing, F. Wu, and H. Feng, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," in Proc. IJCAI, 2016, pp. 3552–3559.

[15] N. McLaughlin, J. M. del Rincón, and P. C. Miller, "Recurrent convolutional network for video-based person re-identification," in Proc. IEEE Conf. CVPR, Jun. 2016, pp.1325–1334.

[16] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in Proc. IEEE Conf. CVPR, Jun. 2016, pp. 1345–1353.

[17] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani,"Person re-identification by symmetry-driven accumulation of local features," in Proc. IEEE Conf. CVPR, Jun. 2010, pp. 2360–2367.

[18] D. Baltieri, R. Vezzani, and R. Cucchiara, "Learning articulated body models for people re-identification," in Proc. 21st ACM Int. Conf. Multimedia, 2013, pp. 557–560.

[19] S. Sunderrajan and B. S. Manjunath, "Context-aware hypergraph modelling for re-identification and summarization," IEEE Trans. Multimedia, vol. 18, no. 1, pp. 51–63, Jan. 2016.

[20] M. Köstinger, M. Hirzer, P.Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in Proc. IEEE Conf. CVPR, Jun. 2012, pp. 2288–2295.

[21] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, "Person re-identification by regularized smoothing KISS metric learning," IEEE Trans. Circuits Syst. Video Technol., vol. 23, no.10, pp. 1675–1685, Oct. 2013

[22] X. Zhu, X.-Y. Jing, X. You, W. Zuo, S. Shan, and W.-S. Zheng, "Image to video person re-identifcation by learning heterogeneous dictionary pair with feature projection matrix," IEEE Trans. Inf. Forensics Security, vol. 13, no. 3, pp. 717-732, Mar. 2018.

[23] Guangcong Wang, Jianhuang Lai, "P2SNet: Can an image match a video for person re-identification in an end to end way?", IEEE transaction on circuits and systems for video technology, vol.28, no.10, October 2018, pp. 2777- 2787.

[24] Benzhi Yu, Ning Xu and Jian Zhou," Cross-media body part attention network for image to video person re-identification", IEEE Access, vol.7, August 2019, pp. 94966- 94976.