Attributes Co-occurrence Pattern Mining for Video-based Person Re-identification

Xiu Zhang, Federico Pala, Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside, CA 92521, USA
xiu.zhang@email.ucr.edu, fedpala@ucr.edu, bhanu@cris.ucr.edu

Abstract

Person re-identification has received considerable attention in the image processing, computer vision and pattern recognition communities because of its huge potential for video-based surveillance applications and the challenges it presents due to illumination, pose and viewpoint changes among non-overlapping cameras. Being different from the widely used low-level descriptors, visual attributes (e.g., hair and shirt color) offer a human understandable way to recognize people. In this paper, a new way to take advantage of them is proposed. First, convolutional neural networks are adopted to detect the attributes. Second, the dependencies among attributes are obtained by mining association rules, and they are used to refine the attributes classification results. Third, metric learning technique is used to transfer the attribute learning task to person re-identification. Finally, the approach is integrated into an appearance-based method for video-based person re-identification. Experimental results on two benchmark datasets indicate that attributes can provide improvements both in accuracy and generalization capabilities.

1. Introduction

Person re-identification (re-id) aims at matching pedestrians across non-overlapping cameras. It has many applications in visual surveillance such as human tracking, identification and action recognition. Despite significant advances in the area, inaccuracies caused by scene illumination changes, varying human poses and low image resolution are still challenging and make person re-identification difficult in real world surveillance scenarios.

Additionally, video-surveillance recordings require the adoption of soft biometrics since face recognition techniques may not be feasible. Indeed, a majority of the literature considers the person appearance as the main cue for person re-id. To address this task, there are two commonly used methods: (a) extracting robust hand-crafted features from the given images or videos, (b) developing a method for learning a robust similarity function between individuals. However, commonly used hand-crafted features (e.g., color, texture, anthropometric measures and gait [7, 18, 20]) are insufficiently robust to reliably model different conditions such as changes in viewpoint and occlusions.

Figure 1. An example where people's appearance (e.g., described by the global color) may be misleading. Attributes can give useful discriminative information: given the woman image (a) as a query, her brown and long hair can be used to distinguish her in (b) from the man in (c) who has black and short hair.

Today the convolutional neural networks (ConvNets) are the most effective techniques to automatically learn the best features for vision tasks [22, 25]. Thus, we propose the use of ConvNets for attributes detection, and use them for identifying pedestrians. Attributes offer a human understandable way for recognizing individuals. For example, when people are asked to describe a person, they often use statements such as: "A man with a white T-shirt and blue jeans," or "A girl wearing sunglasses and a black shirt". In Figure 1 we show some examples where attributes can give some insight for person re-identification.

Unfortunately, the use of attributes is not always reliable because of various complications: (a) surveillance cameras acquire low resolution recordings that make attribute detection extremely challenging; (b) only limited labeled data are available to classify attributes, considering that some attributes rarely exist among people (e.g., blue hair, print dresses in different designs); (c) the same attribute can have different discriminative capabilities depending on the dataset. For example, black hair is discriminative for distin-

978-1-5386-2939-0/17/$31.00

guishing people in Europe, however, it is useless in Asia.

In this paper, we propose a three-stage framework to overcome these difficulties. In the first stage, we design three ConvNets for attributes detection, one for each body part: head, torso, and legs. In the second stage, the co-occurrence of attributes is used to refine the detection results. The intuition behind this is that people tend to follow some rules in dressing themselves. In the third stage, we take the prediction scores of attributes as features, and use metric learning to learn a similarity function among image sequences. Finally, we integrate our framework into an appearance-based model [17] for the final prediction.

2. Related Work and Contribution

2.1 Appearance Based Person Re-identification. A large part of person re-identification methods subdivide the body into different parts to deal with its nonrigid nature and represent each part as an ensemble of various local and global features [7] describing their appearance using color, shape and textural hand-crafted descriptors [4, 7, 19].

Deep learning algorithms have been applied to re-identification thanks to their ability to automatically learn relevant features from the raw pixels composing people images/videos. Unfortunately, a problem that arises in the task of learning a good pedestrian representation is the scarcity of examples. Whereas plenty of images can be collected acquiring several subjects, re-identification consists of recognizing one subject given just the few frames that may be available from a single acquisition. To overcome this issue, two strategies are currently being used: 1) grouping the images in pairs [2, 15] or triplets [5], learning a representation that encapsulates the similarity between frames of the many matching and mismatching pairs that can be generated; 2) starting from the similar task of attribute detection - where more labels may be available - and then build a signature of the subject on the basis of the retrieved appearances.

In this work we used the method of [17] as a baseline for our framework to validate the effectiveness of incorporating the attributes information. It generates an invariant feature representation based on siamese networks, with a recurrent convolutional neural network (RCNN) to model the sequence of frames.

2.2 Attributes Based Person Re-identification. Since attributes are more consistent across viewpoints, they are appropriate for improving person re-identification. The work in [14] proposes a pedestrian descriptor that extracts attributes from handcrafted features, and concatenates them with other low-level characteristics. The approach in [21] predicts attributes by using a ConvNet architecture using a triplet objective function. Attributes detection minimizes the distance between attributes of the same person and maximizes the distance between attributes of different person. In this paper, we first separate the whole body into parts, and then formulate the attribute detection task as a multi-class classification problem by training separate neural networks for each body part.

2.3 Exploiting Co-occurrence Patterns. Several techniques for exploiting co-occurrence information for recognition have been proposed in the literature. Collaborative filtering (CF) collects the relevance feed-back on the co-occurrence of items and it is widely used in recommendation systems [23]. One of the challenges in CF comes from the data sparsity of large target datasets. Feng, et al. [8] construct a hierarchical concept co-occurrence representation with network community detection algorithms and use it to assist image classification. It is based on the assumption that the relationship between two concepts are symmetric, which does not applies in our case. For example, we may assume that a woman wearing a skirt is also wearing high heals, but cannot assume the reverse to be true as well. Based on the above considerations, we use association rules [1] to overcome these limitations. They are based on the concepts of attribute support and confidence.

2.4 Learning Metrics for Person Re-identification. Metric learning [13, 3, 16] has been widely used in person re-identification and aims at learning a metric to minimize the intra-class distance for the same person and maximize the inter-class distance among different persons. KISSME [13] is one of the most popular metric learning methods. It assumes that the difference between image pairs follows a Gaussian distribution with a zero mean, and constructs a metric based on the log-likelihood ratio test. Also, it adopts principal component analysis (PCA) to learn a more compact mapping. XQDA [16] is an extension of KISSME, which first learns a discriminant subspace by using a method similar to linear discriminant analysis (LDA). Then, it applies KISSME to learn a distance function in the generated subspace.

2.5 Contribution. As compared to the state-of-the-art, our contribution can be summarized in the following two points:

(1) We present a novel framework that takes into account attributes and their co-occurrence.

(2) We perform experiments that highlight the generalization capabilities of the framework. We train on a large independent attribute dataset and then test on two different re-id benchmarks. Unlike [27], our approach performs consistently on both testing datasets.

3. Technical Approach

In this section, we present our approach for person re-identification. It is composed of the elements shown in Figure 2. The (a)(b)(c) part represents the three stages of our algorithm, explained in section 3.1, 3.2 and 3.3. First, we train three ConvNets to detect the attributes, then the co-occurrence patterns are mined and used to refine the classification results. Finally, we employ the XQDA metric learn-

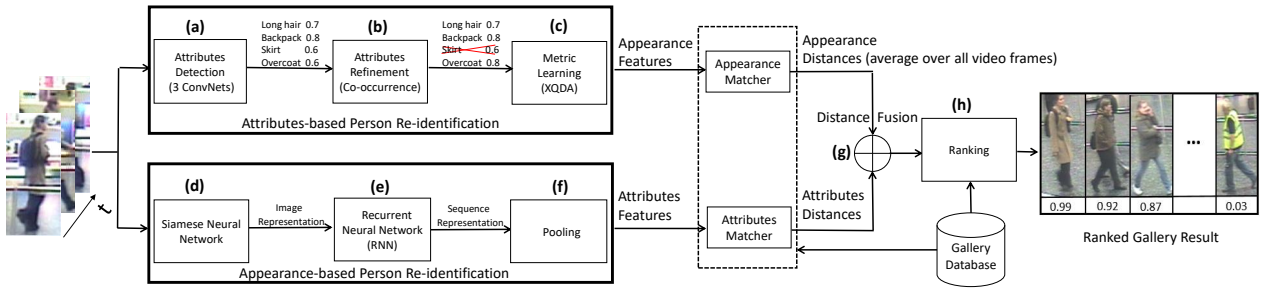

Figure 2. The architecture of our method. (a)(b)(c) represent our three-stage attributes-based person re-id module. In (a) the pedestrian video frames go through three ConvNets to extract his/her attributes. In (b) co-occurrence information among attributes is used to refine the detection results. (c) takes the attribute detection scores and concatenate them into a single feature vector. After that, it uses the XQDA metric learning technique [16] to produce the similarity matrix. (d)(e)(f) is a state-of-the-art video-based person re-identification model [17], where the sequences pairs first go through a siamese neural network (d), then the information flow is captured by RNN (e), and the corresponding features are obtained by max pooling (f). (g) integrates the outputs from (c) and (f) by addition. (h) ranks the results from (g) with respect to the gallery for the final prediction.

ing technique [16] in order to transfer the refined attributes detection scores into a representation targeted to person re-identification. The lower section of Figure 2(d) trains a feature extraction network using a Siamese network architecture. (e) is a RCNN that handles the video stream and (f) pools its outputs [17]. An integration of the outputs from the two pipelines is finally used for the final ranking.

3.1 Attributes Detection Using ConvNets. Taking advantage of the potential of deep neural networks in learning good representations of pedestrians for re-id [2, 15], we formulate the attribute detection task as a multi-label annotation problem by designing three different networks. A separate network is trained for the head, torso and leg parts, to classify the attributes located in the corresponding body parts. The three networks use a similar architecture of three convolutional layers and two fully connected layers. Each output node of the last fully connected layer represents the prediction of a specific attribute. Depending on the size of the body part and the number of relevant attributes, we adjust the filter size and the number of neurons in the last fully-connected layer. We use the output of the network as features for attribute detection. We concatenate the features from the three different networks to form a 54 dimensional appearance feature vector to distinguish persons. The ConvNet architecture for the head part is shown in Figure 3. We used Rectified Linear Units (ReLU) as non-linearities and Max-pooling for downsampling.

Binary Cross-Entropy is used as the loss function for training and it is defined as:

$$L(X,Y) = -\frac{1}{n}\sum_{i=1}^{n} y^{(i)} \log a(x^{(i)}) + \\ +(1-y^{(i)}) \log(1-a(x^{(i)})) \tag{1}$$

where $X = \{x^{(1)}, \cdots, x^{(n)}\}$ is the set of examples taken from the training set, $Y = \{y^{(1)}, \cdots, y^{(n)}\}$ is the corresponding set of the labels and $a(x^{(i)})$ is the output of the network given the input $x^{(i)}$. n is the total number of training samples.

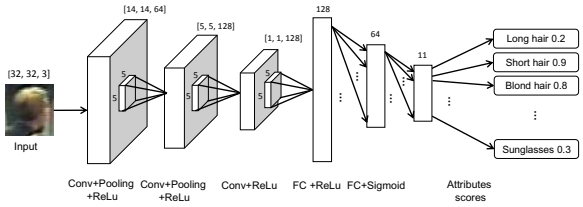

Figure 3. ConvNet architecture for attributes detection (head part). The input is a 32 by 32 pixels RGB image. The size of the convolutional kernels is 5 by 5. For each convolutional layer, the triplets shown represent the size of the outputs and the number of feature maps. The #s above the FC layers refer to the # of nodes.

3.2 Co-occurrence Pattern Mining for Attributes Detection Refinement. The results of person re-id by using the attributes detection scores, as the signature of a person, turn out to be very noisy. This is caused by the low resolution of images and the limited amount of labeled data. As attributes are local features, the higher the resolution is, the better results we get. On the other hand, the labeled data we get from the existing datasets is very limited and cannot adapt well to unknown test settings. What makes things even worse is that the data have a highly unbalanced distribution, which means that the proportion of the positive samples differs a lot among different attributes. In this paper, we employ the co-occurrence information to overcome these issues and refine the attributes detection results.

We first introduce two basic concepts: support and confidence. The support value $sup(a_i)$ is defined as the proportion of images in the database which contains the attributes a_i. The confidence value $conf(a_i \rightarrow a_j)$, with respect to a set of images I, is the proportion of the images that contain both a_i and a_j, N represents the total number of images I.

$$sup(a_i) = \frac{n(a_i)}{N} \tag{2}$$

$$conf(a_i \rightarrow a_j) = \frac{sup(a_i \cup a_j)}{sup(a_i)} \tag{3}$$

Association rules [1] aim at mining the relationships $a_i \rightarrow a_j$ under the restriction of the given lower bounds for both the support and the confidence. The advantages are the following: (1) they fit the non-symmetric nature well; (2) items can be overlapped, which complies with the pairwise relationships among different attributes; (3) they take into consideration of both the reliability (support) and the accuracy (confidence). (4) They can not only predict the existence-to-existence relationship but also the existence-to-nonexistence one. We adopted FP-growth algorithm [11] for calculating the co-occurrence information matrix because it converges very fast by virtue of its tree-like structure. Weka platform [10] has been used to implement our mining technique for its simplicity to explore the data and run learning algorithms.

The specific procedure is the following. First, we concatenate the attributes score features from the three networks into a single vector, where each element represents the prediction of the corresponding attribute. Then we refine each output vector with the following rules: if $a_i \rightarrow a_j$ describes a positive relation, which means $conf(a_i \rightarrow a_j) > 0$, then:

$$P(a_j) = max(P(a_j), P(a_i) \times conf(a_i \rightarrow a_j)) \tag{4}$$

if $a_i \rightarrow a_j$ represents a negative relation, where $conf(a_i \rightarrow a_j) < 0$, then:

$$P(a_j) = min(P(a_j), P(a_i) \times conf(a_i \rightarrow a_j)) \tag{5}$$

3.3 Transfer Learning for Person Re-identification. So far, we have focused on the attributes detection problem. To apply the attributes prediction results to the person re-identification task, the attribute scores are considered as features for the target person. Then, a metric learning method is adopted to learn a distance metric by minimizing the intra-class distance while maximizing the inter-class distance for the individuals in the gallery. Among the many different metric learning methods, we have chosen one of the best performing ones, the Cross-view Quadratic Discriminant Analysis (XQDA) [16]. We point out that for attribute detection the ground-truth is composed of the attributes for each pedestrian whereas XQDA uses the subjects' identity as labels.

3.4 Integration with an Appearance-based Re-Id System. We adopt the framework of [17] to train and get the features for our specific testing task as shown in Figure 2(d)(e)(f). In the training phase, given a pair of images, a Siamese network is used to map each video frame into a feature vector. The temporal information taken from the video sequences is processed by the RNN layer and captures motion information such as gait, body and clothing movement. Then, an integration, as in Figure 2(g), is performed by simply adding the distance metric values achieved from the two separate parts, which are calculated from appearance features and attribute features. These distances are used to rank the probes with respect to the template gallery as shown in Figure 2(h).

4. Experiments

We validate our approach on two of the most important benchmark datasets for video-based person re-identification. We start by describing the datasets and the experimental setup. Then, we present and analyze the experimental results. For the attributes detection network, we use a NVIDIA Digits DevBox, which comes with Four TITAN X GPUs with 7 TFlops of single precision, 336.5 GB/s of memory bandwidth, and 12 GB of memory/board. For the co-occurrence pattern mining, we use Weka 3 package.

4.1 Datasets. The first dataset is iLIDS-VID [24], which consists of 2 acquisitions of 300 pedestrians at an airport arrival hall. The length of videos varies from 23-192 frames with an average of 73 frames. This dataset is very challenging due to the changing illumination conditions and viewpoints, complex backgrounds and occlusions.

The PRID 2011 dataset [12] includes 200 pairs of image sequences taken from two adjacent camera views. The length of the image sequences varies from 5 to 675, with an average of 100 frames. Compared with the iLIDS-VID dataset, this is less challenging because of the relatively simple backgrounds and the rare presence of occlusions.

PEdesTrian Attribute (PETA) [6] is the dataset we used to learn attributes. It is a large-scale surveillance dataset of 19000 attribute labeled images taken from 8707 persons. Each image is annotated with 61 binary and 4 multi-class attributes, such as hair style, clothing color and accessories. The images in this dataset are from 10 different datasets including 477 images from iLIDS-VID and 1134 images from PRID. The images from iLIDS-VID and PRID 2011 datasets are handled in performing experiments as described in Section 4.2.

4.2 Experimental Setup. For each dataset, we randomly extract two equal subsets, one for training and one for testing. During the testing stage, for each query sequence, we compute the distance against each identity in the gallery set and return the top n identities. To measure the performance, the Cumulative Match Characteristic (CMC) plot is used, which represents the percentage of the test sequences that are correctly matched within the specified rank. The experiments are repeated 10 times and the average CMC plot is reported.

For evaluating the attributes detection, the corresponding 477 and 1134 images of related person identities in iLIDS-VID and PRID 2011 are removed from the PETA dataset separately. Then we randomly select 16000 images from the remaining PETA dataset for training and leave the remaining 2523 and 1866 images from the two datasets for validation. We test the performance on two datasets of 150 and 100 videos from iLIDS-VID and PRID 2011, respectively. For all the experiments, we use a starting learning rate of 0.01, momentum of 0.9, learning rate decay of 10^{-3}, weight decay of 10^{-4}, and total # of epochs is 500.

For mining the co-occurrences of attributes, the number of pedestrians for training is 8587 and 7773 for iLIDS-VID and PRID 2011, respectively, after removing the related labeled identities. We heuristically set the support parameter of FP-growth to 0.1, and the confidence value to 0.5.

Methods	r=1	r=5	r=10	r=20
Attributes	3.7	9.7	14	24.2
Ours without refinement	59.7 (0.034)	85.3	93.6	98
Ours	**60.3** (0.035)	85.3	93.6	98
RCNN [17] CVPR 2016	58.3 (0.035)	84.6	92	96.7
TDL [26] CVPR 2016	56.3	**87.6**	**95.6**	**98.3**
TAPR [9] ICIP 2016	55	87.5	93.8	97.2
SI^2DL [27] IJCAI 2016	48.7	81.1	89.2	97.3

Table 1. Comparison of the recognition rates at different ranks (%) on iLIDS-VID dataset. The first row shows the result of attributes only. The second row is the result by integrating attributes from three body parts with the RCNN [17]. The refinement using co-occurrence information is shown in the third row. The reference RCNN part result is shown in the forth line. When available, we show the standard deviation values along with the average of the rank one recognition rate.

Methods	r=1	r=5	r=10	r=20
Attributes	4.3	15.3	31	43
Ours without refinement	72.7 (0.041)	93	96.3	98.3
Ours	73.2 (0.038)	93	96.3	98.3
RCNN [17] CVPR 2016	70.6 (0.051)	92.3	95.3	97.3
TDL[26] CVPR 2016	56.3	87.6	95.6	98.3
TAPR[9] ICIP 2016	55	87.5	93.8	97.2
SI^2DL [27] IJCAI 2016	**76.7**	**95.6**	**96.7**	**98.9**

Table 2. Comparison of recognition rates at different ranks (%) on PRID 2011 dataset. It follows the same structure as Table 1.

4.3 Experimental Results and Discussion. We compare our results with the following state-of-the-art methods: Recurrent Convolutional Neural Networks (RCNN) [17], Top-Push (TDL) [26], Temporally Aligned Pooling Representation (TAPR) [9] and Simultaneously learning Intra-Video and Inter-video Distance Learning (SI^2DL) [27]. In Table 2 and 3 we show the recognition rates for the two datasets at rank 1, 5, 10, 20 respectively. From Table 1, 2 and Figure 4 we can see that for both the datasets there is an improvement in the identification rate for all the methods at all considered

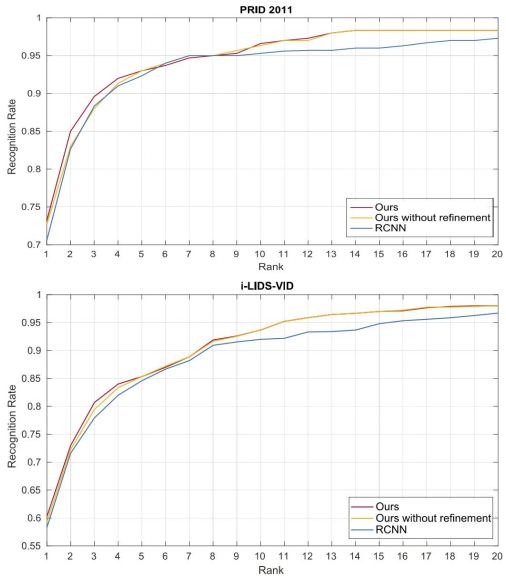

Figure 4. CMC curves for iLIDS-VID and PRID 2011 datasets, comparing the reference RCNN model [17], our model without co-occurrence refinement, and our model in different colors.

Figure 5. Comparison results of our models Vs. RCNN model.

ranks.

We achieve rank 1 identification rates of 60.3% and 73.2%, which results in improvements of 2% and 2.6% with respect to [17] for iLIDS-VID and PRID 2011 datasets respectively. This means that about three additional persons have been recognized correctly in the first attempt for both datasets. Also, our algorithm gets about 0.5% increase over without the refinement for both datasets. For iLIDS-VID dataset, our algorithm achieves the best rank 1 performance. For the PRID 2011 dataset, the results are approaching to the best result obtained by [27]. However, if we examine the results of [27] on the more chanllenging iLIDS-VID dataset, it obtains the lowest recognition rates compared to all the listed results. Instead, our method performs consistently well on both datasets. It is fair to say that attributes information and co-occurrence patterns are complementary to RCNN [17].

Figure 5 shows some examples of rank one recognitions. In Figure 5(a), we show an example where our approach works better than RCNN [17]. In Figure 5(b) instead, we show another example where the RCNN outperforms our approach. The people in Figure 5(a) wear similar clothes, but they have different hair color and style. The probe woman in Figure 5(a) has long blond hair while the wrong

man has short brown hair. Often, the cause of errors is the wrong detection of attributes. In Figure 5(b) we can see the woman in the probe wear a yellow shirt, however, in the test images, there is no yellow shirt detected. That's why our model fails. The standard deviation of rank 1 with respect to the RCNN model [17] is shown in Table 2 and 3. We can observe that the standard deviation of our algorithm over 10 runs is quite similar to that of RCNN [17].

5. Conclusions

We designed a new framework for person re-identification. The proposed approach demonstrated that attributes are important cues for person re-identification and that co-occurrence information can help to improve their descriptive capabilities. When this is combined with an appearance based approach it leads to improved performance. We trained three convolutional neural networks to first classify the attributes, and then used the corresponding co-occurrence information to refine the classification. Subsequently, we adopted the XQDA [16] metric learning method to transfer the problem from attributes detection to person re-identification. Since we used an independent dataset for training attributes, our method achieved a better and more consistent results with respect to previous work.

6. Acknowledgment

This work was supported in part by NSF grants 1330110 and 1552454 and ONR grant N00014-12-1-1026 . The contents of the information do not reflect the position or policy of US Government.

References

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD RECORD*, 1993.

[2] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[3] L. An, M. Kafai, S. Yang, and B. Bhanu. Reference-based person re-identification. In *AVSS*, 2013.

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

[5] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *CVPR*, 2016.

[6] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *ACM MM*, 2014.

[7] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing*, 2(2):127–151, 2011.

[8] L. Feng and B. Bhanu. Semantic concept co-occurrence patterns for image annotation and retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):785–799, 2016.

[9] C. Gao, J. Wang, L. Liu, J.-G. Yu, and N. Sang. Temporally aligned pooling representation for video-based person re-identification. In *ICIP*, 2016.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[11] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *ACM SIGMOD RECORD*, 2000.

[12] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011.

[13] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[14] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *BMVC*, 2012.

[15] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[16] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[17] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 2016.

[18] F. Pala, R. Satta, G. Fumera, and F. Roli. Multimodal person reidentification using rgb-d cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(4):788–799, 2016.

[19] R. Satta, F. Pala, G. Fumera, and F. Roli. Real-time appearance-based person re-identification over multiple Kinect cameras. In *VISAPP (2)*, pages 407–410, 2013.

[20] R. Satta, F. Pala, G. Fumera, and F. Roli. People search with textual queries about clothing appearance attributes. In *Person Re-Identification*, pages 371–389. Springer, 2014.

[21] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, 2016.

[22] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, 2014.

[23] S. Uchihashi and T. Kanade. Content-free image retrieval by combinations of keywords and user feedbacks. In *CIVR*, 2005.

[24] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, 2014.

[25] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng. An enhanced deep feature representation for person re-identification. In *WACV*, 2016.

[26] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *CVPR*, 2016.

[27] 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 *IJCAI*, 2016.