Re-id: Hunting Attributes in the Wild

Ryan Layne
r.d.c.layne@qmul.ac.uk
Timothy M. Hospedales
t.hospedales@qmul.ac.uk
Shaogang Gong
s.gong@qmul.ac.uk

Computer Vision Group
Queen Mary University of London
London, E1 4NS, U.K.
http://qml.io/vision

Abstract

Person re-identification is a crucial capability underpinning many applications of public-space video surveillance. Recent studies have shown the value of learning semantic attributes as a discriminative representation for re-identification. However, existing attribute representations do not generalise across camera deployments. Thus, this strategy currently requires the prohibitive effort of annotating a vector of person attributes for each individual in a large training set – for each given deployment/dataset. In this paper we take a different approach and automatically discover a semantic attribute ontology, and learn an effective associated representation by crawling large volumes of internet data. In addition to eliminating the necessity for per-dataset annotation, by training on a much larger and more diverse array of examples this representation is more view-invariant and generalisable than attributes trained at conventional small scales. We show that these automatically discovered attributes provide a valuable representation that significantly improves re-identification performance on a variety of challenging datasets.

1 Introduction

Person re-identification addresses the task of recognising a person in diverse scenes, as viewed by non-overlapping surveillance cameras. Particularly, when monitoring distributed locations over time, we want to differentiate the target person from all other candidates and match them when they appear in another camera view - potentially from another angle, against clutter, or subject to variable levels of occlusion and lighting conditions (see Figure 2 for examples). This re-identification capability is often a critical component in a variety of safety, security, and analytics tasks, where it is necessary to maintain awareness of consistent identities across space and time. In particular, it is a fundamental capability for long-term tracking across multiple disjoint camera views [15]. Manual re-identification by trained human operators is prone to human error and may be impossible due to information overload when real-time monitoring is required. Moreover, performance levels vary between operators, and individual attentiveness varies throughout the working day [31]. Thus demand has grown for automated re-identification capabilities.

Due to the increasing importance of CCTV as a deterrent and investigative tool, as well as the intrinsic challenge of the problem, research into re-identification is now extensive [9, 14].

Much re-identification research breaks down into two main areas; developing effective representations that are discriminative for identity whilst invariant to lighting and viewpoint change [5, 9, 17] and development of learning methods trained to discriminate identities [1, 14, 28]. Feature-centric approaches [9] suffer from the problem that it is extremely challenging to obtain features that are discriminative enough to distinguish people reliably, while simultaneously being invariant to all the practical covariates such as motion blur, clutter, view angle and pose change, lighting and occlusion. In contrast, learning approaches [14] better use a given set of features, by discriminatively training models to maximise re-identification performance, for example metric learning [14] and support vector machines (SVM) [1, 28]. These lines of inquiry are nevertheless synergistic because better feature representations tend to improve a given discriminative method, while applying better discriminative methods to a given representation also tends to improve results.

A recent line of work [19, 23, 29, 34] in feature/representation learning draws inspiration from the practices of human experts. Human operators focus their attention on noting and matching distinct semantic characteristics, or *attributes*, to simplify their task. These may correspond to distinct soft-biometric, appearance or functional properties such as gender or clothing-style. Attribute-centric approaches learn a low-dimensional feature representation that corresponds to such semantic properties. They typically approach this by: asking expert operators to define an ontology of such characteristics, collecting and annotating site specific training data with a vector of attributes per person, training computer vision models to detect attributes, and then using the estimated attributes of each person as a representation for re-identification. However, this top-down human-defined attribute approach has some critical limitations: (i) It requires costly attribute annotation of site-specific training data. This is significantly more costly than person-identity information used to train discriminative matching models. (ii) Top-down definition of attributes does not guarantee that they are visually computable by computer vision techniques given visual surveillance data. (iii) Due to the limited scalability of the annotation approach, the annotated data is likely to be too small scale to learn accurate and robust detectors for each attribute of interest.

In this paper we address these issues by taking a very different data-driven [6, 26] approach to learning attributes for re-identification. We automatically construct a bottom-up attribute ontology, and learn an effective associated representation by large-scale mining of noisy but abundant content on social photo sharing sites. Specifically, rather than asking an expert to define an ontology [19, 23, 29, 34], we discover it automatically by clustering photo tags and comment data. These clusters are used to train a large bank of detectors, resulting in a number of visually detectable attributes[1]. This process is significantly more scalable than manually annotating data per surveillance site for attribute learning. Moreover, the greater volume and diversity of data used to train these automatically discovered attributes results in a more reliable and generalisable attribute representation than conventional attribute representation approaches on surveillance datasets can normally achieve. We validate our contribution by using our representation to obtain excellent results on a set of four of the most challenging re-identification datasets to date.

[1]This is in contrast to expert defined ontology, which while intuitive to experts, may correspond to properties not possible to detect reliably with current vision techniques

2 Related Work and Contributions

Re-identification A central challenge for re-identification is obtaining a feature representation that is discriminative for identity but invariant to view, lighting and other covariates. Conventional approaches to re-identification (re-id) typically attempt to address this by exploiting visual cues such as color, texture, spatial structure, and combinations thereof [2, 5, 9, 28], as well as more recent representations such as saliency [36]. Some classic representations such as ELF [11] are fully hand designed, while others such as SDALF [9], Fisher Vectors [25] and saliency [36] are learned to a a greater or lesser extent. However, a challenge with feature design/learning is that many do not perform consistently – depending on the particular re-identification dataset / camera pair, a different basic feature may be best. Most feature representations can be used directly for re-identification by nearest neighbour (NN) matching or in conjunction with a model-based matching procedure such as metric learning [4, 14] or SVM [1, 28]. These are discriminatively trained to improve matching performance on a particular dataset / camera pair, and usually improve on direct NN matching for a given feature.

Attributes for Re-identification Inspired by the success of attribute representations in other computer vision tasks, a recent line of work [19, 20, 23, 29, 34] has studied applying attributes to learn an informative representation for re-identification. The strategy has typically been to annotate binary or categorical clothing, object and soft-biometric properties on the training portion of a dataset, and then train models (such as topic models [23], SVM [19], or latent-SVMs [20]) to predict these mid-level properties based on some base low-level feature. Interestingly – *assuming attributes are reliably detectable* – only about twenty binary attributes are necessary to achieve unprecedented near perfect matching accuracy on challenging benchmarks [18]. The main bottleneck is actually accuracy/robustness of attribute detection. This is hard to achieve because surveillance video is often of poor quality. However more fundamentally, it is challenging because obtaining sufficient annotated data to train reliable attribute detectors for each camera is prohibitively costly or impossible. In this paper, we thus take a different approach to the attribute strategy, by mining attributes and attribute training data from social photo sharing sites. Automatically generating attribute detectors that both do not require manual annotation and are trained from sufficiently large scale data could be more scalable and generalisable. However, the challenge then becomes how to learn meaningful bottom-up attributes from large scale internet data, given that such mining delivers highly noisy images and annotations.

Attribute Discovery As ever growing amounts of visual data are being shared on the public web, the computer vision community has begun to exploit this resource for obtaining large scale datasets and text/visual data mining [7]. Meanwhile, the availability of cheap crowd-sourced annotation has begun to make annotation of large-scale datasets more feasible [7]. However, crowd-sourced annotation at scale still incurs expenses in terms of time and human effort, and the results are often prone to bias and noise [32]. An alternative is to develop algorithms to mine data on the internet [3, 20] with little or no human intervention. This may take the form of obtaining (noisily labeled) training data by image search using keyword query [6], or mining socially shared photos and associated tags/annotations [7].

With regards to attributes specifically, NEIL [6] has performed semi-supervised learning of attribute detectors based on large scale internet image sets, starting with a small seed

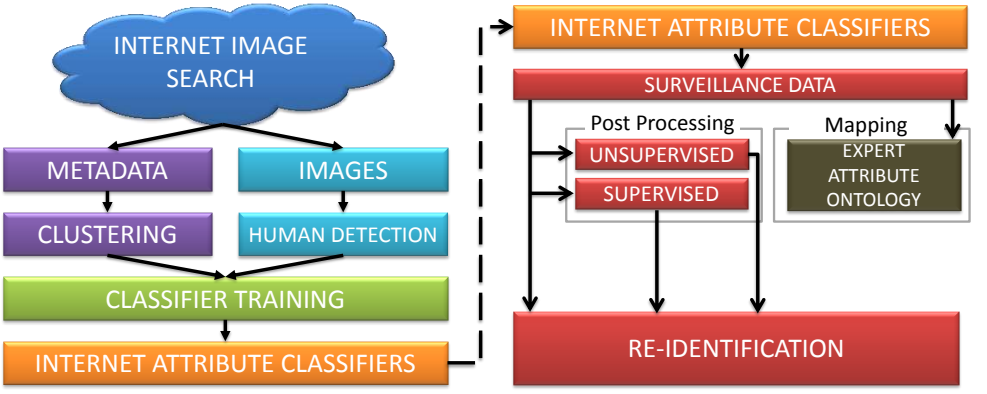


Figure 1: Schematic overview of our pipeline; Post-Processing modules such as distance-metric learning or domain-adaptation can be applied depending on the level of supervision available in order to boost "rank 1" or overall system performance as needed

.

amount of annotated data. Meanwhile in the context of retail photos, [3] has clustered product photo annotations to automatically discover an ontology of putative attributes, for which detectors are then trained.

We employ a similar strategy to Berg *et al*. in [4], but we must discover attributes from deeply noisy and unconstrained data; rather than metadata and images from a noisy, but otherwise semantically-curated website.

Contributions In this paper we show how to (i) leverage web data in order to discover and learn semantically meaningful attributes that are effective for re-identification and (ii) use this discovered attribute representation in conjunction with discriminatively trained matching techniques to obtain state-of-art performance on a wide variety of re-identification datasets.

3 Discovering and Learning Attributes for Re-id

In this section we outline how we first acquire a space of attributes from uncurated internet data (Figure 2, Sections 3.1, 3.2), then how to train detectors for each attribute (Section 3.3) and fuse them with compatible representations for re-identification (Section 3.4). We include a schematic overview of our entire pipeline in Figure 1.

In order to alleviate the burden of annotating vast amounts of attribute training data, we first aim to acquire a large volume of *uncurated* and weakly labelled data from the internet. Clearly, the kinds of photographs we might find online without a directed search stand a low probability of being immediately suitable for our purposes - Berg *et al*. [4], neatly summarise our problem as identifying "wheat from amidst a great deal of chaff". Therefore we define a "broad" search query that is likely to return photographs that contain depictions of people in everyday attire. We construct a boolean search query comprising of frequent synonyms of the word "person", such as "man", "woman", "pedestrian", etc, and combine this with multiple negative terms such as "car", "tree", "cat", and download 220,000 images with their

associated metadata. This approach differs from most work on conventional recognition [33] where images are categorically and strongly annotated - or derived from a heavily curated source such as an eCommerce website selling a catalogued array of products. In our case, there is no guarantee that a photograph and associated metadata will have any meaningful semantic link, let alone whether or not the metadata refers to what we're really interested in: tags and keywords that describe the appearance of the people, if any, in the photograph.

The metadata for each photograph comprises of a variety of noisy but potentially useful information; location information is not used in our work, but present in approximately 8% of the photographs at at least country-level, which could potentially be used to learn region-specific attributes in later work. For our purposes, we merge the photograph title and meta-tags, and employ common pre-processing measures to standardise the meta-text string somewhat; we tokenise and remove stop-words, remove numerical characters, and stem words to conflate semantically identical words to their common root. We do not apply a spelling-check so as to preserve any popular internet vernacular, names or other bespoke allegorical terms that may be relevant or insightful at a semantic level in themselves, but may not have entered official spelling dictionaries. For example, a user's specific choice of tag for a city from all available toponyms may reveal some information about the rationale behind the annotation; tagged images may also be somehow visually distinct as a result. Each photograph's meta-text is represented as a bag-of-words (BOW) histogram of bigrams with term frequency-inverse document frequency weighting (tf-idf) which ensures that salient words are more prominently represented. Lastly, we constrain each bigram's constituent tokens to being at least 3 characters long.

3.1 Discriminative text features from meta-text

As a first step to discovering latent attributes from the internet data, we apply self-tuning Spectral Clustering [27, 35] based on the BOW tf-idf metatext representations with a vocabulary of $\approx 5,000$ bigrams (see Figure 2 for examples). Unlike Marchesotti *et al.* [26], we calculate the similarity between the frequency of the unigrams and bigrams rather than using the Levenshtein distance on the second gram within each bigram. Our intuition is that in our case it is the co-occurrence of the grams that is semantically relevent, not the similarity to other bigrams. Spectral clustering has the advantage versus other common clustering methods of performing well regardless of the spatial arrangement of the underlying clusters, making it suitable for our needs. We manually specify $N_a = 200$ clusters, which will correspond to potential attributes.

3.2 Person Detection

Many retrieved images are unsuitable for learning attribute-models suitable for surveillance because they contain landscape or objects instead of persons; or because persons are present but too close-up. To filter the data to obtain suitable images, we select Dollar *et al.*'s person detector [8][2] and employ both pre-trained models supplied by the authors to extract bounding boxes of people from this extremely varied collection of photos. This person detector is a vital component in dealing with the vast amount of noise inherent in the internet-sourced images, since it affords us the ability to (i) determine if people are in an image with a measure of confidence, and (ii) be selective about how confident the detections we use for classifier

[2] http://vision.ucsd.edu/ pdollar/toolbox/doc/index.html

Figure 2: (Left) Uncurated images from our internet search. Many images are unsuitable for surveillance attribute learning as they contain no people or are too close-up. (Right) Some person detections automatically extracted from uncurated internet photographs. Each row comprises images from a discovered cluster, explicitly labelled on the left side. Noteworthy are the high variations in pose and appearance, fashion and background, as well as lighting and how the fashion varies according to location.

training – in order to trade off data volume and label noise. After conservatively thresholding person detection confidence, we are left with 69,000 person crops with corresponding meta-text features.

3.3 Classifier Training

Due to memory limitations, using traditional Support Vector Machines strategies for training attribute detectors [18] was not tractable and therefore we select Linear Discriminant Analysis (LDA). Despite being a mature approach LDA still out-performs some contemporary machine learning methods, particularly for cases where there are many classes and comparatively few positive examples per-class. This, combined with being computationally less expensive and less sensitive to class imbalance, make it useful for our purposes. Using all 69,000 crops, we train an independent LDA model for each of the $N_a = 200$ discovered attributes. Finally we build a representation for any person's image X in an internet-attribute semantic-space by stacking the positive-class posteriors from each detector into a N_a dimensional vector: $IA(X)$.

We train on internet-sourced data, which one expects to have somewhat different statistics to typical surveillance crops. For example, surveillance crops typically come from lower quality cameras with more motion blur and compression artefacts. This may negatively affect the ability of our internet data trained representation to effectively encode surveillance detections in practice. We therefore investigate applying unsupervised domain-adaptation to better align the internet training data and surveillance test data. In particular, we align the projected subspaces of the two datasets, using Gong *et al.*'s geodesic flow kernel domain adaptation (DA) method [10].

3.4 Re-identification, Calibration and Fusion

The attributes obtained thus far are trained directly from discovered text clusters. There is variability in their reliability of detection based on image data, or their usefulness for re-identification. We therefore address learning a linear weighting $\mathbf{w}$ to rescale the attributes IA such that they are weighted according to their maximum utility for re-identification. Standard

choices of optimisation criteria for re-identification include the first rank (R1) percentage, which reflects how often the first result in a ranked list is a perfect match to the probe, or expected rank (ER) or normalised area under curve (nAUC) of the cumulative match characteristic curve (CMC). We wish to enforce both a strong early-rank score, and good overall performance. To achieve this, we maximise the *product* of the CMC curve values $\hat{p}(k)$ at each rank k

$$\hat{P}_{\mathbf{w}}(k) = CMC_{\mathbf{w}}(k) = \frac{1}{n}\sum_{p=1}^{n} \mathbf{1}\left(k_p \leq k\right) \tag{1}$$

where k_p is the distribution of the ranks based on NN re-identification using $L1$ distances $D(IA_p, IA_g)$ between each attribute encoded probe $IA_p \in \mathcal{P}$ and all gallery member, $IA_g \in \mathcal{G}, g = 1,...,n$. Specifically we use greedy search to select the weight $\mathbf{w}$ that maximises the following metric when used to scale each dimension/attribute a:

$$\min_{\mathbf{w}} \prod_{k=1}^{n} \hat{P}_{\mathbf{w}}(k) \tag{2}$$

Fusion with Low-Level Features Finally, we integrate our representation with metrics based on other low-level features. Specifically, we fuse BR-SVM [1] (trained on ELF features), SDALF [9] and our weighted internet attributes after further discriminative training using KISS [16]. The resulting pseudo-metric's fusion weights *beta* can be trivially selected with standard optimisation methods:

$$\begin{aligned} D(X_p, X_g) &= d_{KISS}(IA(X_p), IA(X_g)) && (3)\\ &+ \beta_{SDALF} \cdot d_{SDALF}(X_p, X_g) && (4)\\ &+ \beta_{BRELF} \cdot d_{BRELF}(X_p, X_g). && (5) \end{aligned}$$

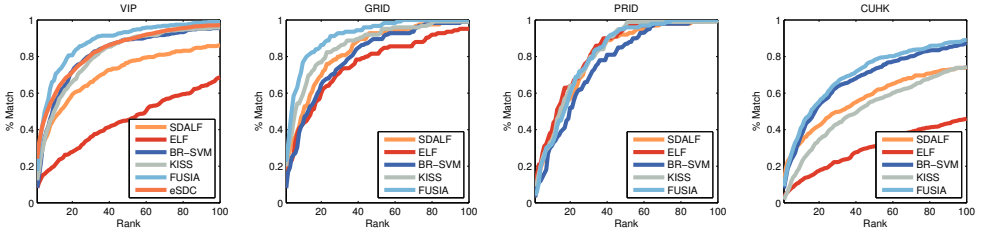
For re-identification, we perform standard NN re-identification based on the fused metric in Eq(3), which we denote FUSIA.

4 Experiments

We validate our contributions on four challenging public datasets, quantifying re-identification performance in the standard way[11] with CMC curve visualisations (CMCs), and expected-rank scores (ERs). CMC curves indicate the likelihood of a probe's true match appearing by the k^{th} rank, whilst ER represents the average rank of the true match to each probe – corresponding to the relevant metric of how far a human operator would have to search down a ranked list of matches before verifying the true target. High CMC values and ERs indicate better overall system performance.

4.1 Datasets and Settings

REID Datasets We tested the model using four publicly available re-id datasets: VIPeR [30], PRID [12], GRID [24] and CUHK [21], which provide 316, 200, 250, and 971 matched pairs respectively. These datasets cover a diverse variety of image sizes (in the region of [128x48] to [128x64].), typical view angles and camera conditions. For supervised learning experiments, we take a standard 2-fold partition approach to training and testing.



	VIPeR		GRID		PRID		CUHK	
Method	R1 ↑	ER ↓	R1 ↑	ER ↓	R1 ↑	ER ↓	R1 ↑	ER ↓
ELF [11]	0.08	84.27	0.13	26.42	0.11	18.26	0.04	160.55
BR-SVM [1] (ELF)	0.08	21.45	0.08	21.15	0.03	24.80	0.08	43.28
KISS [16] (IA)	0.12	21.51	0.20	14.73	0.09	19.27	0.02	73.16
FUSIA	0.17	13.39	0.22	9.55	0.04	19.90	0.09	38.59
SDALF [9]	0.16	44.02	0.16	17.86	0.03	20.79	0.12	72.96
eSDC [36]	0.24							
Liu *et al.* [22]	0.16							
RANKSVM [28]	0.15				0.10			
Hirzer *et al.* [13]	0.21		0.15					

Figure 3: Overall re-identification performance of our FUSIA representation versus alternatives. Top: CMC Curves; Bottom: "Rank 1" and Expected Rank summaries.

Person Detection, Representation and Domain Adaptation We discard detections with confidence $c < 0.5$, in order to minimise false positives which degrade classifier performance. Cropped person detections are normalised to 128x48 pixels prior to feature extraction. For our visual features we employ the commonly used ensemble of local features[11] (ELF), which encodes both color and texture in 6 horizontal strips [28] for final features with 2784 dimensions, and reduce dimensionality to 100 with PCA; for feature fusion, we also use symmetry-driven accumulation of local features (SDALF) as detailed in [9]. Note that SDALF provides a distance matrix directly, rather than a feature. For Domain Adaptation we have only one parameter to select, and use 10 dimensions as recommended by [10].

4.2 Visual Detectability of Internet Attributes

We first evaluate the visual detectability of the discovered internet attributes. We train the binary attribute classifiers using semantic meta-text cluster assignments as labels, and divide each cluster into training and validation partitions, containing 75% and 25% of the available data respectively. Across all folds and 200 attributes, detection accuracy in the test-folds is 70.28%, which is significant considering that text-based attribute discovery is not guaranteed to produce attributes with visual correlates, and class imbalance between positive and negative classes may negatively impact discriminative learning models. Notably these numbers for detection reliability are comparable to 66-70% obtained using an expert-designed ontology purpose designed to be visually detectable and learned with extensively manual annotation of attribute training data [18].

Dataset	ELF [11]	IA (raw)	IA	KISS[16] (IA)	BRSVM[1] (ELF)	SDALF [9]	FUSIA
VIP	91.03	71.23	44.66	21.25	21.45	44.02	12.94
GRID	33.12	26.05	23.05	17.33	21.15	17.86	10.22
PRID	31.99	19.38	17.63	21.91	76.20	20.79	19.89
CUHK	161.39	138.41	128.13	72.25	43.28	72.96	38.09

Table 1: Breaking down re-identification performance by components of our full FUSIA model. See text for details. We report Expected Rank, lower scores are better.

4.3 Attributes as a Representation for Re-Identification

Figure 3 summarises the re-identification performance of our complete system, FUSIA, on all four datasets along with a variety of state of the art alternatives. The top plot shows CMC curves with our final model FUSIA - or Fused Internet Attributes, along with KISS [16], Binary-Relation SVM [1], SDALF [9] and saliency (eSDC) [36].

In the lower table we report scores obtained using our implementations of the cited methods in the first four rows. The remaining rows report results obtained from the cited works and blank results reflect where alternatives have not published results on a given dataset or format. In all cases we summarise with Rank 1 (perfect match rate), and expected rank. Our Rank 1 is comparable to state of the art alternatives, although not always best – however, our overall performance as evidenced by the CMC curves and their expected rank scores, outperform most alternatives by an often significant margin. This margin demonstrates the discriminative strength of our semantic attribute representation. Meanwhile the consistency of this margin across this wide batch of state of the art datasets demonstrates that the quantity and variety of source data is indeed leveraged to learn a highly generalisable representation.

Table 1 breaks down our method according to the different components and contributions. Plain internet attributes (IA (raw)) fail to outperform the ELF (upon which IAs are constructed). However, the full calibrated (weighted) and domain-adapted variant (IA), boosts overall re-identification performance dramatically to near state-of-art levels on VIPeR, GRID and CUHK, and maintaining comparable performance with other representations on PRID. Finally, applying metric learning to our attributes (KISS(IA)) provides further improvement. The three columns to the left of FUSIA show the component metrics that are fused together to obtain this final result.

5 Discussion

A major advantage of our approach is that effectively unlimited numbers of person images can be obtained. Thus performance is expected to only improve with further application of computer time to crawling and learning more and better attributes. Nevertheless, a disadvantage, is that our attributes (Figure 2) are somewhat less transparent than conventional expert-defined attribute ontologies [18], that (by defining attributes such as "blue shirt" and "red shirt") map more clearly onto descriptive person search tasks. To provide some insight into the mechanism of our contribution, we illustrate the relation between these two interpretations of attributes: We use our framework to encode the VIPeR dataset in 200 dimensional IA representation, and then use existing VIPeR attribute annotation [18] to train a linear SVM mapping from a conventional attribute ontology to our representation. This corresponds to defining conventional expert-defined attributes in terms of a linear combination of internet attributes. Using "red shirt" and "blue shirt" as query terms, we demonstrate the top retrievals in our 69,000 person dataset in Figure 4. The results clearly albeit qualitatively illustrate that

Figure 4: Querying "red shirt" and "blue shirt" in our 69,000 non-labelled internet-sourced images via a transfer mapping between our attributes and expert-ontologies from [18]

our representation provides a simple distributed encoding of conventional expert-defined attributes. This has implications for applications beyond re-identification in surveillance: by connecting existing expert-defined attribute ontologies from surveillance to internet data sources, we gain the ability to query internet images for attributes without additional annotation of internet data or training new classifiers for the internet domain.

Conclusion We have shown how effective mid-level semantic attributes can be automatically discovered from internet data without supervision. These are semantic by construction due to creation via mining of textual tags and comments. Although they vary in visualness, overall they can be detected comparably reliably to expert designed attributes thanks to the effectively unlimited quantity of internet image data available to train them. We demonstrate that this internet attribute representation of person images is generalisable and discriminative for re-identification, a property that is unlocked through domain adaptation and metric learning, and furthermore is synergistic and amenable to fusion with conventional techniques. In future work we would like to investigate in more detail strategies for mapping expert defined and automatically generated ontologies to better enable image-free description-based person search, as well as image-based re-identification.

Acknowledgement Ryan Layne is supported by a EPSRC CASE studentship and by UK MOD SA/SD.

References

[1] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning Implicit Transfer for Person Re-identification. In *European Conference on Computer Vision, International Workshop on Re-identification*, 2012.

[2] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by hpe signature. In *International Conference on Pattern Recognition*, 2010.

[3] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, 2010.

[4] H. Bischof, P. M. Roth, M. Hirzer, P. Wohlhart, and M. Kostinger. Large scale metric learning from equivalence constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[5] S. Bąk, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In *European Conference on Computer Vision*, 2012.

[6] X. Chen, A. Shrivastava, and A. Gupta. NEIL : Extracting Visual Knowledge from Web Data. In *IEEE International Conference on Computer Vision*, 2013.

[7] J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, K. Stratos, K. Yamaguchi, Y. Choi,

H. Daume, A. C. Berg, and T. L. Berg. Detecting visual text. In *North American Chapter of the Association for Computational Linguistics*, 2012.

[8] P. Dollár, R. Appel, S. Belongie, P. Perona, and P. Doll. Fast Feature Pyramids for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[9] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[10] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[11] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, 2008.

[12] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA - Proceedings of the 17th Scandinavian conference on Image analysis*, SCIA'11, 2011.

[13] M. Hirzer, P. M. Roth, and H. Bischof. Person Re-identification by Efficient Impostor-Based Metric Learning. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2012.

[14] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision*, 2012.

[15] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109, 2008.

[16] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[17] R. Layne, T. M. Hospedales, and S. Gong. Towards Person Identification and Re-identification with Attributes. In *European Conference on Computer Vision, International Workshop on Re-identification*, 2012.

[18] R. Layne, T. M. Hospedales, and S. Gong. Attributes-based Re-Identification. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-Identification*. Springer London, 2013.

[19] R. Layne, T. M. Hospedales, and S. Gong. Domain Transfer for Person Re-identification. In *Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS)*, 2013.

[20] A. Li, L. Liu, and S. Yan. Clothes Attributes Assisted Person Re-identification. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Re-identification*, chapter 6. Springer London, 2014.

[21] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *Asian Conference on Computer Vision*, 2012.

[22] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person Re-identification: What Features Are Important? In *European Conference on Computer Vision, International Workshop on Re-identification*, 2012.

[23] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, and J. Bu. Attribute-restricted latent topic model for person re-identification. *Pattern Recognition*, 45(12), 2012.

[24] C. C. Loy, T. Xiang, and S. Gong. Time-Delayed Correlation Analysis for Multi-Camera Activity Understanding. *International Journal of Computer Vision*, 90(1), 2010.

[25] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In A. Fusiello, V. Murino, and R. Cucchiara, editors, *Workshop on Re-Identification, ECCV*, volume 7583 of *Lecture Notes in Computer Science*. Springer, 2012.

[26] L. Marchesotti and F. Perronnin. Learning beautiful (and ugly) attributes. In *British Machine Vision Conference*, 2013.

[27] A. Y. Ng, M. I. Jordan, and Y. Weiss. On Spectral Clustering Analysis and an algorithm. *Neural Information Processing Systems*, 2001.

[28] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person Re-Identification by Support Vector Ranking. In *British Machine Vision Conference*, 2010.

[29] R. Satta, G. Fumera, and F. Roli. A General Method for Appearance-Based People Search Based on Textual Queries. In *European Conference on Computer Vision, International Workshop on*

Re-identification, 2012.

[30] W. R. Schwartz and L. S. Davis. Learning Discriminative Appearance-Based Models Using Partial Least Squares. *Computer Graphics and Image Processing (SIBGRAPI)*, 2009.

[31] G. J. D. Smith. Behind the screens: Examining constructions of deviance and informal practices among CCTV control room operators in the UK. *Surveillance and Society*, 2, 2004.

[32] H. Su, J. Deng, and L. Fei-Fei. Crowdsourcing annotations for visual object detection. In *Workshop on Human Computation, Association for the Advancement of Artificial Intelligence*, 2012.

[33] A.-M. Tousch, S. Herbin, and J.-Y. Audibert. Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, 45(1), 2012.

[34] D. Vaquero, R. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *IEEE Workshop on Applications of Computer Vision (WACV)*, 2009.

[35] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. *Neural Information Processing Systems*, 17, 2004.

[36] R. Zhao, W. Ouyang, and X. X. Wang. Unsupervised Salience Learning for Person Re-identification. In *IEEE International Conference on Computer Vision*, 2013.