Retrieving Relative Soft Biometrics for Semantic Identification

Daniel Martinho-Corbishley, Mark S. Nixon and John N. Carter
School of Electronics and Computer Science,
University of Southampton, United Kingdom.
{dmc,msn,jnc}@ecs.soton.ac.uk

***Abstract*—Automatically describing pedestrians in surveillance footage is crucial to facilitate human accessible solutions for suspect identification. We aim to identify pedestrians based solely on human description, by automatically retrieving semantic attributes from surveillance images, alleviating exhaustive label annotation. This work unites a deep learning solution with relative soft biometric labels, to accurately retrieve more discriminative image attributes. We propose a Semantic Retrieval Convolutional Neural Network to investigate automatic retrieval of three soft biometric modalities, across a number of 'closed-world' and 'open-world' re-identification scenarios. Findings suggest that relative-continuous labels are more accurately predicted than absolute-binary and relative-binary labels, improving semantic identification in every scenario. Furthermore, we demonstrate a top rank-1 improvement of 23.2% and 26.3% over a traditional, baseline retrieval approach, in one-shot and multi-shot re-identification scenarios respectively.**

I. Introduction

Conventionally searching hours of surveillance footage for suspects is extraordinarily time consuming. Automatically describing and identifying pedestrians from eye-witness testimony is therefore a pivotal challenge. Soft biometrics are human characteristics, designed to precisely and reliably describe subjects through semantic attributes [1], [2]. This enables human accessible, semantic identification, without the need to re-identify a prerequisite image. Soft biometrics are also applicable in less constrained environments and when hard biometrics e.g. face, fingerprint or gait are unavailable. This paper investigates the automatic retrieval of absolute and relative soft biometric labels from images, evaluating their semantic identification performance.

A. Problem

Convolutional Neural Networks (CNNs) and deep learning techniques are now common place for re-identification metric learning [3]–[5] and image attribute prediction [6]–[8]. However, almost all works alluding to attribute-based re-identification assume binary or categorical ground-truth labels [6], [8]–[10]. Meanwhile, comparative soft biometrics, that describe subjects with *relative* continuous values, have been shown to outperform their categorical counterparts for subject recognition [11]–[13]. Although such novel labels are more discriminative, their interactions with automatic image retrieval and attribute-based re-identification have yet to be fully investigated.

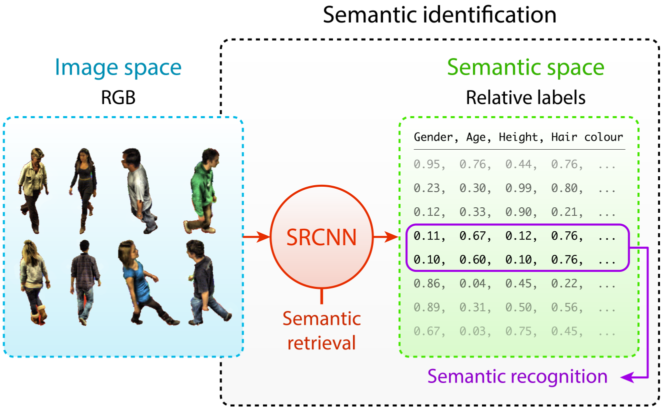


Fig. 1: Visual overview of semantic identification, illustrating semantic retrieval with relative attributes estimated by SRCNN.

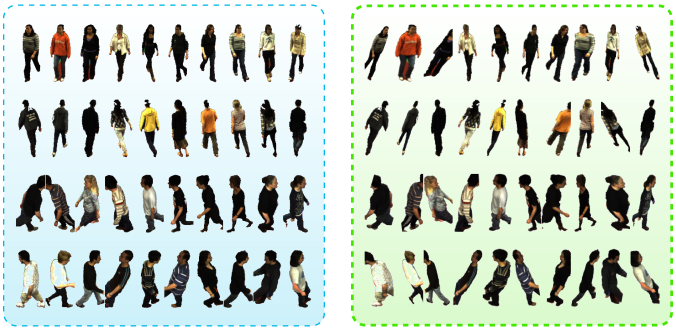
Consequently, performing semantic identification encapsulates two sets of challenges. First, the traditional reidentification challenges of large intra-class variations and inter-class ambiguities owing to disjointly captured person images. Secondly, the distinctiveness, predictability and reproducibility of ground-truth labelling methods. Both sets of issues affect semantic retrieval accuracy, the resulting semantic space and overall identification performance.

B. Proposal

We propose to jointly retrieve binary and continuous attributes from subject images, using a deep learning Semantic Retrieval Convolutional Neural Network (SRCNN). Subjects are identified in a 'semantic space', matching predicted image attributes to subject descriptions, illustrated in Figure 1.

SRCNN is used to evaluate three modalities of soft biometric label from the public Soft Biometric Retrieval (SoBiR) dataset [13]. The characteristics of each labelling technique are explored by performing semantic retrieval in one-shot and multi-shot re-identification scenarios, and the challenging 'open-world' zero-shot identification scenario. We follow the evaluation methodology of [13], directly comparing our results to a baseline solution.

Similarly to [6], our proposed SRCNN incorporates a grid of convolutional layers to jointly predict attributes, with several important distinctions. Firstly, we investigate retrieving both binary and continuous label measures, discussing alterations



(a) Pre-processed originals

(b) Augmented variations

Fig. 2: SoBiR camera views top-to-bottom; front, back, top and side.

to the learning process and scrutinising variations in behaviour during training. Secondly, we present an extended performance enhancing training strategy, including image augmentation, early stopping and attribute recognition weighting. Finally, we emphasise the facilitation of soft biometric retrieval, over the application of state-of-the-art deep learning techniques.

C. *Our Main Contributions*

(1) A deep learning SRCNN architecture and training strategy, to jointly learn and retrieve soft biometric labels. (2) The evaluation of three modalities of soft biometric label, across several challenging surveillance scenarios. (3) A demonstration of semantic identification using relative labels and SRCNN, and its improvement over a baseline approach.

II. Related Work

Two soft biometrics surveys discuss estimating attributes from whole body images and the progression from categorical to relative descriptions [1], [2]. Continuous and relative attributes are now widely accepted over traditional binary and multi-class annotations [14], and can be jointly learnt with minimal guidance. Thus far, relative attributes have been established for facial verification [15], to facilitate zero-shot learning [16] and generate fine-grained image descriptions [17]. In soft biometrics, comparative annotations are proven to outperform categorical annotations for subject recognition from body [11], [12], face [18] and clothing [19]. We investigate semantic retrieval with the SoBiR dataset, introduced by [13], and extend its methodology.

The first attribute-based re-identification work [9] predicted a set of 23 binary attributes for re-identification and reported results from a zero-shot identification scenario. Since then, advances in image attribute prediction with CNNs [20] have influenced the study of pedestrian re-identification. At present, some CNN approaches exist for deep learning similarity metrics between pairs of images [3]–[5] and predicting pedestrian attributes [6]–[8], [10], [21]. While many of these works discuss the estimation of a large number of 'fine-grained' attributes, they all perform binary classification. One study found that regression outperformed classification for demographic age estimation from faces [22]. It is therefore imperative to investigate the amalgamation of state-of-the-art pedestrian attribute prediction techniques and enhanced relative labels, to indicate the direction of future research.

Name	Annotation	Measure	Label type	Combi.	Bal.
abs-bin	Categorical	Absolute	Binary	4096	No
rel-bin	Comparative	Relative	Binary	4096	Yes
rel-con	Comparative	Relative	Continuous	∞	-

TABLE I: Semantic space characteristics of SoBiR labels [13].

Recently, several works investigate the domain transfer of semantic representations, performing semi-supervised and unsupervised semantic recognition [7], [23]–[25]. Attributes are often learnt from images captured in ideal conditions, and transferred to images captured in more unconstrained environments. These methods attempt to address the scalability issue of camera specific annotations and facilitate zero-shot recognition using binary attributes.

III. Soft Biometric Retrieval Dataset

The Soft Biometric Retrieval (SoBiR) dataset is designed to be a pragmatic, flexible and challenging framework with which to investigate automatic semantic retrieval [13]. SoBiR is a relatively small dataset of 1,600 images of 100 subjects, captured from four pairs of viewpoints. Instead of pursuing a large number of image samples, it emphasises a comprehensive set of 4,800 soft biometric ground-truth labels, derived from over 100,000 human annotations. Image resolutions and view orientations are such that pedestrian faces are unobservable in detail, necessitating a reliance on body characteristics for identification, as seen in Figure 2a.

A. *Soft biometric labels*

SoBiR comprises a compact lexicon of 12 soft trait, semantic attributes, drawn from two sources of categorical and comparative ground-truth annotations. In this study we investigate three labelling techniques, outlined in Table I.

Absolute-categorical annotations are first presented by [26], collecting a number of visually assessable, global and body features. Subjects are described in an absolute sense, using pre-defined categories e.g. 'very short', 'short', 'average', 'tall', 'very tall' for height. However, we exclude absolute-categorical labels following consistently poorer performance reported in [13]. Instead, absolute-binary (*abs-bin*) representations are derived from these multi-class labels, by combining classes into two semantic groups, e.g. 'shorter' and 'taller', 'lighter' and 'darker' etc. Groupings are formed such that the new binary labels are as equally balanced as possible.

Relative labels are objectively crowdsourced by [12], annotating pairwise comparisons between each pair of subject images. Annotations are expressed as an ordered relation from one image to another e.g. 'much more feminine', 'more feminine', 'same', 'more masculine', 'much more masculine' for gender. Relative-continuous (*rel-con*) labels are derived from these responses, by applying a similarity constrained RankSVM [17] to all pairwise comparisons. By ranking subjects on a bi-polar scale, a continuous value is attained to describe a subject's possession of each soft trait, e.g. from 'most

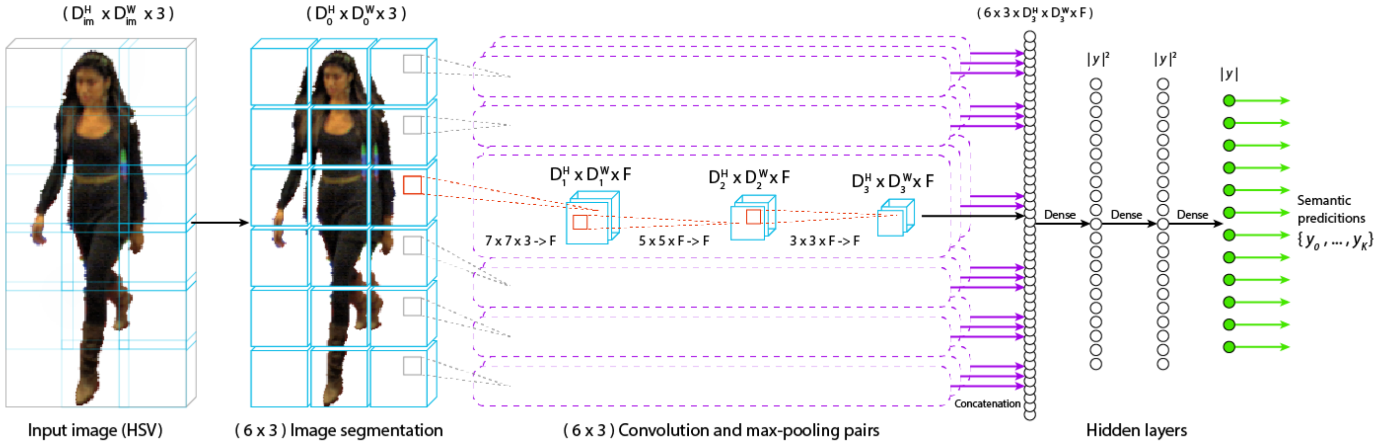

(D^H_{im} x D^W_{im} x 3)
(D^H_0 x D^W_0 x 3)
(6 x 3 x D^H_3 x D^W_3 x F)
D^H_1 x D^W_1 x F
D^H_2 x D^W_2 x F
D^H_3 x D^W_3 x F
7 x 7 x 3 -> F
5 x 5 x F -> F
3 x 3 x F -> F
Concatenation
|y|²
|y|²
|y|
Dense
Dense
Dense
Semantic predicitions $\{y_0, \dots, y_K\}$
Input image (HSV)
(6 x 3) Image segmentation
(6 x 3) Convolution and max-pooling pairs
Hidden layers

Fig. 3: Semantic Retrieval Convolutional Neural Network architecture.

feminine' to 'most masculine', from 'shortest' to 'tallest' etc. Relative-binary (*rel-bin*) representations are derived by separating subject ranks into two balanced halves, forming binary classes. These binary labels are coarser estimations than the fine-grained continuous values, but are still relative measurements.

IV. Semantic Retrieval Convolutional Neural Network

We propose a deep learning, feed-forward, SRCNN architecture to jointly learn and predict a set of semantic attributes from input images, illustrated in Figure 3. We explain the overall architecture of our SRCNN and detail the training strategy employed to alleviate overfitting.

By designing the neural network as a whole, image features and attributes are learnt in conjunction, overcoming many of the challenges associated with empirically matching feature descriptors to machine learning methods. Our solution could also perform demographic estimation, by scaling human perceived ground-truths, in order to denote real-world anthropometric measurements.

A. SRCNN Architecture

Input images of size $D^H_{im} \times D^W_{im} \times 3$ are fed into the network, represented as three channels in the HSV colour space, portraying semantic concepts of colour and shade.

Three convolutional and max-pooling layer pairs are applied sequentially, learning low-level features from image samples. Each layer is fully-connected to the last, causing learnt filters to have global spatial invariance within the image. Although highly variant, person images do exhibit some regularities in alignment around the sagittal axis. We aim to preserve this global spatial information, learning attribute-centric detectors for specific body regions.

Therefore, images are divided into a grid of 6×3 overlapping cells in place of body-part detection, similarly to [6]. Cells are of dimensions $D^H_0 \times D^W_0$, chosen to be $D^H_0 = D^W_0 = 24$. Each layer pair convolves its input with square kernels in decreasing sizes, $K_1 = 7, K_2 = 5, K_3 = 3$, and square pool sizes of $P = 2$. All layers learn $F = 16$ filters, with output maps of size $D^H_i = D^H_{i-1} - K_i + 1$ and $D^W_i = D^W_{i-1} - K_i + 1$.

The final layers of max-pooling are concatenated as a layer of size $6 \times 3 \times D^H_3 \times D^W_3 \times F$. Outputs are then fed through two dense hidden layers of size $|y|^2$. The last fully-connected layer represents the final output, of size $|y|$. In this way, attributes are jointly learnt, exploiting any relationships that occur between labels. Unlike [6], we do not predefine connections between image regions and semantic attributes, as many of our soft traits are global descriptions, learning correlations automatically through training to generalise our solution. We use a sigmoid activation for the final layer and Rectified Linear Unit (ReLU) activations for all other layers.

B. Loss Functions

As labels are learnt together, we define two separate multi-label loss functions for classification and regression formulations. Loss values are averaged over all K attributes. For binary classification we define Binary Cross-Entropy (BCE):

$$L_{BCE} = -\frac{1}{|K|} \sum_{k \in K} \left(y^t_k \cdot log(y^p_k) + (1 - y^t_k) \cdot log(1 - y^p_k) \right)$$

where y^t_k is the ground-truth target label and y^p_k is the predicted label of the k-th attribute. For continuous regression, we define a Mean Squared Error (MSE):

$$L_{MSE} = \frac{1}{|K|} \sum_{k \in K} (y^t_k - y^p_k)^2$$

To optimise all 1,259,436 parameters, the SRCNN is trained through back propagation with the ADADELTA stochastic gradient descent method [27]. The solution is implemented in Python using the Theano library and run on a GPU using CUDA and CuDNN.

V. Training Strategy

We employ several training strategies to reduce overfitting and help find a robust solution. For classification tasks, dropout regularisation [28] is applied between convolutional layers, with a dropout ratio of 0.5. It was found that including dropout for certain regression experiments excessively prolonged training time, due to the characteristics of the MSE loss function, discussed in Section VI-A. No subsequent fine-tuning is required after implementing our training strategy.

A. Data Augmentation

During taining, input images are randomly augmented, artificially increasing the training set size, to resemble variations in pose and camera angle. We employ five label-preserving data transformations in the augmentation pipeline; horizontal reflection, horizontal scaling, rotation, shearing and horizontal translation.

Half the training images are mirrored at random. Horizontal reflection is the most common data augmentation method, significantly reducing overfitting. The next pipeline stage involves rotation, shearing and horizontal scaling around the image mid-point, sampled uniformly from respective ranges $\theta \in [-\frac{\pi}{12}, \frac{\pi}{12}]$, $\varphi \in [-\frac{\pi}{12}, \frac{\pi}{12}]$ and $x_s \in [-\frac{D_{im}^W}{5}, \frac{D_{im}^W}{5}]$. Finally, horizontal translation is applied in the range $x_t \in [-\frac{3D_{im}^W}{10}, \frac{3D_{im}^W}{10}]$. Rotation and shearing echo disparities in pose, namely the position of the head and legs through the walking action and viewpoint rotation around the longitudinal axis. Horizontal scaling reproduces the affects of rotation around the frontal axis, caused by variations in camera elevation. Horizontal translation compensates for discrepancies in bounding-box alignment and is especially important as images are subdivided into non-continuous regions. Images are cropped to their original size around the mid-point and edge pixels are repeated to fill any gaps. Example augmentations of SoBiR images can be seen in Figure 2b.

B. Early Stopping

To mitigate overfitting of the training data, we define an early stopping function, based on the semantic recognition accuracy of the validation set, rather than on its loss value as is common, reasoned in Section VI-A. Training is halted if $AR(e) < AE(e-w)$ and $e > w$, where $AR(e)$ represents the average semantic recognition rank of the validation set at epoch e. A trailing window of $w = 30$ epochs is chosen, balancing premature stopping against responsiveness.

C. Attribute Recognition Weighting

Soft traits do not have equal discriminative ability, affected by label distributions, retrieval accuracy and ground-truth annotation methods. Therefore a set of attribute weightings are discovered, with which to perform semantic recognition, following the formulation of [13].

The objective function, O, finds a weight vector $\mathbf{w}$, such that semantic recognition ranks for the validation set are minimised. We wish to attain a lower loss value between probe i's predicted and ground-truth labels $L(p_i, t_i)$, than between probe i's predicted label and gallery subject j's ground-truth labels $L(p_i, t_j)$, for all probe and gallery subjects i and j:

$$O = \sum_{i \in I} \left(\sum_{j \in J} \begin{cases} 1, & \text{if } \mathbf{w}^T L(p_i, t_j) < \mathbf{w}^T L(p_i, t_i) \\ 0, & \text{otherwise} \end{cases} \right)^{\lambda}$$

where L is either a Hamming loss vector for binary classification, or a Mean Squared Error loss vector for continuous regression. By setting $0 < \lambda < 1$, precedence is given to improving already low ranks over higher ones. We empirically choose $\lambda = 0.8$ to optimise the number of low-end ranks, while minimising the average overall rank. To prevent overfitting, elements of $\mathbf{w}$ are randomly initialised in the range of $[0.5, 1.5]$.

Experiment	No. Samples		No. Subjects		No. Cameras	
	Tr.	Va.+Te.	Tr.	Va.+Te.	Tr.	Va.+Te.
SoBiR One-shot	100	100	100	100	1	1
SoBiR Multi-shot	700	100	100	100	7	1
SoBiR Zero-shot	720	80	90	10	8	8

TABLE II: Non-overlapping train (Tr.), validation (Va.) and test (Te.) set criteria.

VI. Experiments

We present three experiments, each evaluating abs-bin, rel-bin and rel-con soft biometric label modalities. Experiments follow the non-overlapping, train-validation-test split criteria in Table II.

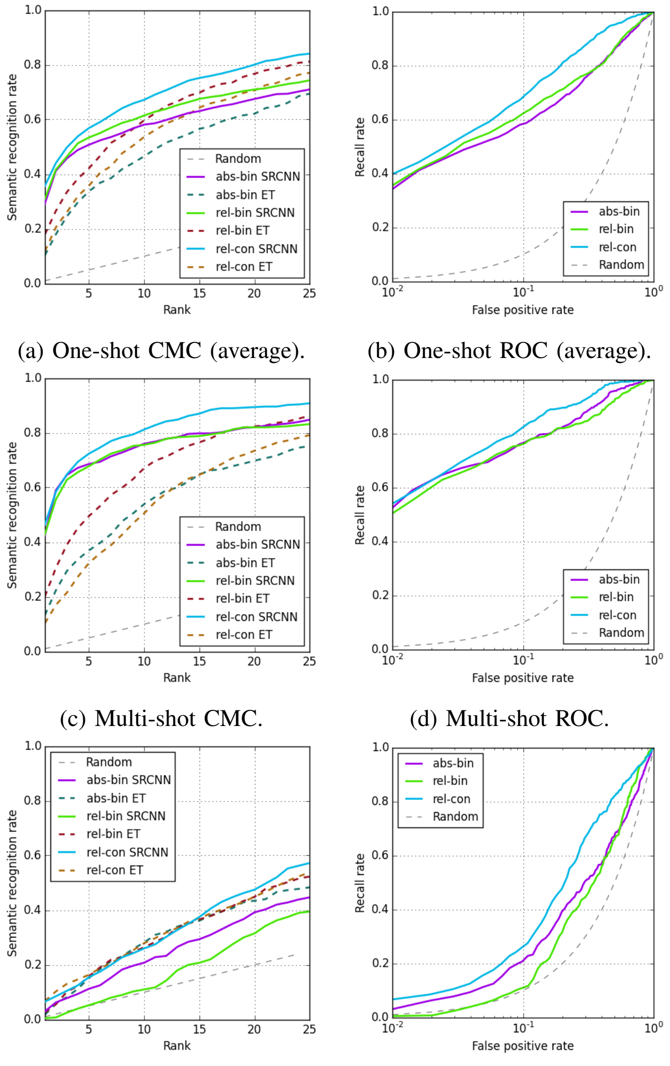
One-shot re-identification performs semantic retrieval with one camera pair at a time, randomly selecting alternative train-test viewpoints per subject. The process is repeated across all four camera pairs from the SoBiR dataset. Multi-shot re-identification samples one camera image per subject for the test set and allocates the remaining 7 camera images to the training set. Zero-shot identification is the most challenging scenario, simulating identification of a previously unseen suspect, given only an eye-witness description. In this scenario, train-test sets are split across subjects, allocating all 8 camera samples of 10 subjects to the test set and training on the remaining subject images.

Results are reported as the average of 10-fold cross validation, with equally divided validation-test splits. In all experiments, probe subjects are identified by first retrieving semantic labels from a sample image, and then performing semantic recognition against a gallery of known descriptions. Results are contrasted against a traditional semantic retrieval approach that uses hand-crafted descriptors and an Extra Trees ensemble learning method (ET), reported as a baseline on SoBiR [13].

A. One-shot Re-identification

Figure 4a and 4b summarise the semantic identification results achieved by applying SRCNN to *front*, *back*, *top* and *side* camera views from SoBiR. All three labelling modes clearly surpass the ET semantic retrieval approach, with a top rank-1 increase of 23.2%. Rel-con recognition accuracy outperforms both abs-bin and rel-bin modes, gaining an average of 6.5% and 4.8% at rank-1 respectively, Table IIIa.

To investigate why this is, we plot a side-by-side example of the validation loss and average label prediction accuracy during training time in Figure 5. In rel-con retrieval mode, the validation loss is consistently minimised, while attribute prediction accuracy steadily increases.



(a) One-shot CMC (average).

(b) One-shot ROC (average).

(c) Multi-shot CMC.

(d) Multi-shot ROC.

(e) Zero-shot CMC.

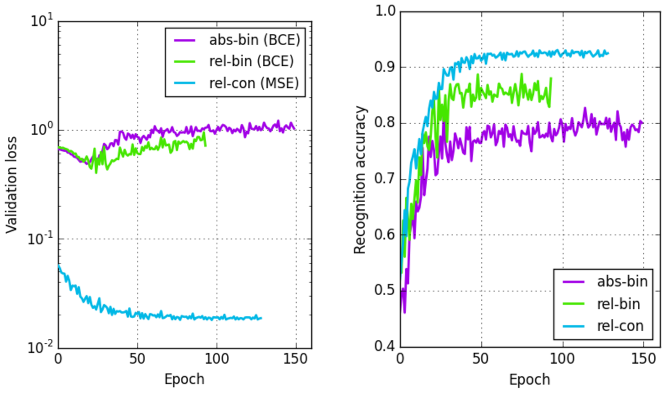
(f) Zero-shot ROC.

Fig. 4: SoBiR one-shot, multi-shot and zero-shot semantic recognition with SRCNN and ET [13].

Labelling	r=1	r=5	r=10	r=25	r=50	r=75	nAUC
(a) One-shot re-identification (average)							
abs-bin	29.2	50.8	58.1	71.0	85.8	94.5	80.7
rel-bin	30.9	53.4	61.5	74.3	86.2	95.3	82.0
rel-con	**35.7**	**56.9**	**67.2**	**84.1**	**95.1**	98.7	**88.1**
best ET	12.5	36.0	53.7	80.4	85.4	**99.3**	84.8
(b) Multi-shot re-identification							
abs-bin	43.0	68.8	75.8	85.2	95.6	98.6	89.8
rel-bin	43.2	67.6	76.0	82.8	91.6	97.6	88.1
rel-con	**46.4**	**72.2**	**81.8**	**90.2**	**98.8**	**99.6**	**92.8**
best ET	20.1	49.5	77.1	86.3	95.4	99.4	88.1
(c) Zero-shot identification							
abs-bin	3.9	11.3	20.8	44.7	67.0	83.9	61.9
rel-bin	0.5	5.3	11.0	39.85	65.8	90.0	60.9
rel-con	6.5	15.5	26.0	**57.35**	**82.0**	92.5	**71.4**
best ET	**6.7**	**16.4**	**28.2**	53.9	81.4	**94.1**	70.8

TABLE III: SoBiR semantic recognition CMC% and normalised Area Under Curve (nAUC) with SRCNN and best ET scores [13].

Meanwhile, in binary classification modes, validation loss values reverse around epoch 25, as recognition accuracy continues to increase, breaking their monotonicity. This suggests that the SRCNN is still learning important prediction decisions, enhancing recognition performance, while overfitting in terms of prediction error.



(a) Validation loss.

(b) Recognition accuracy.

Fig. 5: Example training time characteristics, contrasting validation loss and average recognition accuracy.

	Attribute weightings average±std				
	SRCNN				ET [13]
Trait	abs-bin	rel-bin	rel-con	Average	Average
Gender	**1.3±0.4**	1.1±0.1	1.2±0.5	1.2±0.4	**2.1±0.6**
Height	1.2±0.4	1.2±0.3	**1.5±0.6**	**1.3±0.5**	1.6±1.0
Age	1.2±0.2	1.2±0.2	1.2±0.1	1.3±0.4	0.5±0.8
Weight	1.2±0.2	1.2±0.4	0.9±0.6	1.1±0.5	0.5±0.5
Figure	**1.3±0.7**	1.3±0.2	0.8±0.7	1.1±0.6	0.9±0.7
Chest size	1.3±0.3	**1.3±0.1**	1.0±0.6	1.2±0.4	0.7±0.5
Arm thickness	1.0±0.5	1.1±0.2	1.1±0.3	1.1±0.4	1.1±0.7
Leg thickness	1.2±0.4	**1.4±0.1**	1.2±0.7	**1.3±0.5**	0.2±0.6
Skin colour	**1.4±0.3**	**1.5±0.5**	**1.7±0.7**	**1.5±0.5**	**1.6±0.4**
Hair colour	1.2±0.4	1.2±0.3	**1.3±0.6**	1.2±0.4	**3.1±0.7**
Hair length	1.2±0.5	1.2±0.1	**1.6±0.5**	**1.3±0.4**	**2.0±0.6**
Muscle build	**1.4±0.4**	**1.3±0.2**	1.1±0.5	1.3±0.4	0.7±0.9

TABLE IV: Average attribute recognition weightings for one-shot scenario (top four emboldened).

This influenced our choice of early stopping function, which analyses the semantic recognition rate of the validation set, rather than its loss value. In fact, evaluating Spearman's rank-order correlation, we find a strong negative coefficient (-0.95) between validation loss and prediction accuracies for rel-con training, compared to weak positive coefficients for rel-bin and abs-bin training modes (0.17 and 0.63). As a result, rel-con performs particularly well, as SRCNN's negated loss values correlate more closely to final semantic recognition rates.

We also compare learnt attribute weightings to weights found after applying the ET method [13], Table IV. With ET, attribute weightings are quite divisive, as better performing attributes are weighted distinctly higher than others. However, with our SRCNN approach, the range of weightings is narrower, implying that attributes are retrieved with similar accuracy (being jointly learnt). Interestingly, skin colour and hair length are strongest in both approaches, while leg thickness and muscle build are given more significance than gender, at odds with the weightings of ET. Continuous regression takes around 1.6× longer to train than binary classification, in this experiment.

B. Multi-shot Re-identification and Zero-shot Identification

Figure 4 reports our last set of results, again performing semantic recognition with SoBiR, but now in ‘open-world’ mutli-shot and zero-shot scenarios. In both tasks, SRCNN is trained and tested on all camera samples.

Multi-shot re-identification far outperforms the ET approach’s top rank-1 by 26.3%, Table IIIb. By providing a larger number of training samples per probe subject, recognition performance is radically increased.

In stark contrast, zero-shot identification attains relatively low recognition performance across all labels, with rel-bin fairing particularly poorly. In fact, only rel-con labelling offers any improvement over the ET approach in this scenario, gaining 1.2% nAUC, Table IIIc. Intriguingly, in both scenarios, abs-bin outperforms rel-bin labels, in contrast to the baseline solution. This indicates that imbalances in labelling are less detrimental to SRCNN, and that the abs-bin split may actually better describe the demographic distribution of SoBiR.

By excluding all probe subject images from the training set, attribute retrieval rates are drastically reduced, similarly to Layne et al.’s findings [9]. This shows that while SRCNN is able to semantically retrieve viewpoint-invariant descriptions of known subjects, there is some difficulty in learning stand-alone semantic attributes that are independent of the subjects who possess them. Improving this scenario is a crucial step in being able to perform automatic semantic pedestrian identification, given only a human description or eye-witness testimony of a suspect. Compared to single-shot, the multi-shot experiment requires on average 1.6× more training epochs and zero-shot requires 0.4× fewer epochs, indicating the extent and depth of each learning process.

VII. Conclusions

In this paper we have demonstrated the semantic retrieval of three soft biometric modalities, using a deep learning Semantic Retrieval Convolutional Neural Network (SRCNN). SRCNN jointly learns and predicts semantic attributes, enabling semantic identification from only an eye-witness testimony. Our approach achieves a top rank-1 increase of 23.2% and 26.3% over a traditional retrieval method in ‘closed-world’ and ‘open-world’ re-identification scenarios on the SoBiR dataset.

Our findings indicate that relative continuous labels not only provide more discriminative labels than binary alternatives, but also enhance semantic identification performance when automatically retrieved using SRCNN. Furthermore, we have shown that both attribute predicition and semantic identification are facilitated by uniting deep learning techniques with relative attributes.

Future work points towards improving the zero-shot identification scenario, perhaps by applying domain transfer learning techniques to relative attributes. Very large-scale comparative annotation may also be alleviated by extending a subset of labelled data to the whole, moving towards real-world applications of semantic identification.

References

[1] M. Nixon, P. Correia, and K. N. et al., “On soft biometrics,” *Pattern Recognition Letters*, 2015. 1, 2

[2] A. Dantcheva, P. Elia, and A. Ross, “What else does your biometric data reveal? A survey on soft biometrics,” *WIFS*, 2015. 1, 2

[3] D. Yi, Z. Lei, and S. Z. Li, “Deep metric learning for practical person re-identification,” *arXiv preprint arXiv:1407.4979*, 2014. 1, 2

[4] W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in *CVPR*. IEEE, 2014. 1, 2

[5] E. Ahmed, M. Jones, and T. Marks, “An improved deep learning architecture for person re-identification,” in *CVPR*. IEEE, 2015. 1, 2

[6] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Li, “Multi-label cnn based pedestrian attribute learning for soft biometrics,” in *ICB*. IEEE, 2015. 1, 2, 3

[7] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan, “Deep domain adaptation for describing people based on fine-grained clothing attributes,” in *CVPR*. IEEE, 2015. 1, 2

[8] H. A. Perlin and H. S. Lopes, “Extracting human attributes using a convolutional neural network approach,” *Pattern Recognition Letters*, 2015. 1, 2

[9] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, “Person re-identification by attributes.” in *BMVC*, 2012. 1, 2, 6

[10] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, “A richly annotated dataset for pedestrian attribute recognition,” *arXiv preprint arXiv:1603.07054*, 2016. 1, 2

[11] D. Reid and M. Nixon, “Using comparative human descriptions for soft biometrics,” in *TPAME*. IEEE, 2011. 1, 2

[12] D. Martinho-Corbishley, M. S. Nixon, and J. N. Carter, “Soft biometric recognition from comparative crowdsourced annotations,” in *ICDP*. IET, 2015. 1, 2

[13] D. Martinho-Corbishley, M. Nixon, and J. Carter, “Soft biometric retrieval to describe and identify surveillance images,” in *ISBA*. IEEE, 2016. 1, 2, 4, 5

[14] B. Qian, X. Wang, N. Cao, Y. Jiang, and I. Davidson, “Learning multiple relative attributes with humans in the loop,” *Image Processing*, 2014. 2

[15] N. Kumar, A. Berg, and P. B. et al., “Attribute and simile classifiers for face verification,” in *CVPR*. IEEE, 2009. 2

[16] C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning to detect unseen object classes by between-class attribute transfer,” in *CVPR*. IEEE, 2009. 2

[17] D. Parikh, A. Kovashka, A. Parkash, and K. Grauman, “Relative attributes for enhanced human-machine communication.” in *AAAI*, 2012. 2

[18] N. Almudhahka, M. Nixon, and J. Hare, “Human face identification via comparative soft biometrics,” in *ISBA 2016*. IEEE, 2016. 2

[19] E. Jaha and M. Nixon, “Viewpoint invariant subject retrieval via soft clothing biometrics,” in *ICB*. IEEE, 2015. 2

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*, 2012. 2

[21] C. Ng, Y. Tay, and B. Goi, “A convolutional neural network for pedestrian gender recognition,” in *ISNN*. Springer, 2013. 2

[22] X. Wang, R. Guo, and C. Kambhamettu, “Deeply-learned feature for age estimation,” in *WACV*. IEEE, 2015. 2

[23] Z. Shi, T. M. Hospedales, and T. Xiang, “Transferring a semantic representation for person re-identification and search,” in *CVPR*. IEEE, 2015. 2

[24] A. Schumann and R. Stiefelhagen, “Transferring attributes for person re-identification,” in *AVSS*. IEEE, 2015. 2

[25] R. Layne, T. M. Hospedales, and S. Gong, “Re-id: Hunting attributes in the wild,” in *BMVC*. BMVA, 2014. 2

[26] S. Samangooei and M. Nixon, “On semantic soft-biometric labels,” in *Biometric Authentication*. Springer, 2014. 2

[27] M. D. Zeiler, “Adadelta: An adaptive learning rate method,” *arXiv preprint arXiv:1212.5701*, 2012. 3

[28] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *arXiv preprint arXiv:1207.0580*, 2012. 3