Super-Fine Attributes with Crowd Prototyping

Daniel Martinho-Corbishley, Mark S. Nixon and John N. Carter

Abstract—Recognising human attributes from surveillance footage is widely studied for attribute-based re-identification. However, most works assume coarse, expertly-defined categories, ineffective in describing challenging images. Such brittle representations are limited in descriminitive power and hamper the efficacy of learnt estimators. We aim to discover more relevant and precise subject descriptions, improving image retrieval and closing the semantic gap. Inspired by fine-grained and relative attributes, we introduce super-fine attributes, which now describe multiple, integral concepts of a single trait as multi-dimensional perceptual coordinates. Crowd prototyping facilitates efficient crowdsourcing of super-fine labels by pre-discovering salient perceptual concepts for prototype matching. We re-annotate gender, age and ethnicity traits from PETA, a highly diverse (19K instances, 8.7K identities) amalgamation of 10 re-id datasets including VIPER, CUHK and TownCentre. Employing joint attribute regression with the ResNet-152 CNN, we demonstrate substantially improved ranked retrieval performance with super-fine attributes in direct comparison to conventional binary labels, reporting up to a 11.2% and 14.8% mAP improvement for gender and age, further surpassed by ethnicity. We also find our 3 super-fine traits to outperform 35 binary attributes by 6.5% mAP for subject retrieval in a challenging zero-shot identification scenario.

Index Terms—Attribute-based pedestrian re-identification, Soft biometrics, Crowdsourcing, Retrieval, Perception, PETA dataset

1 Introduction

Visual characteristics are intrinsic to the human identity [1], [2], [3]. They are essential for communicating eye-witness testimony in forensic investigations and when biological ground-truths are unknown. Gender, age and ethnicity are the most commonly reported characteristics in policing [4], criminal record keeping [5] and identity science [2], [6], [7], [8] and are proven to be critical in suspect identification [1], [9]. In remote surveillance operations, searching for a suspect entails automatically retrieving relevant images from large crowds of people, given only an eyewitness description. An example query could be:

> "Find the possibly mixed race, young adult male".

In this scenario, a human operator generates a suspect query, while the search system narrows the suspect image list by automatically estimating and matching subject attributes. Notice that the query may be unclear, ambiguous or nuanced, yet the human and machine are required to communicate descriptions through a shared lexicon of *traits* e.g. gender, age, ethnicity and associated *attributes* e.g. male, female, quite young, mixed race etc.

Problem. Pedestrian re-identification has seen a surge in performance due to advances in deep learning [10], yet attribute recognition from body still lags behind modalities such as face [11]. This is due to the highly unconstrained nature of surveillance footage, which deals with enormous inter- and intra-class variability, inherently low quality and obscured imagery. Despite this, almost all attribute-based re-identification approaches assume brittle, binary representations, which are ineffective at recounting subjective or uncertain properties in challenging images. Consequently, conventional subject labels are coarse and often inconsistent or irrelevant (Fig. 1 blue), resulting in lower quality ground-truths and degraded generalisation of learnt estimators.

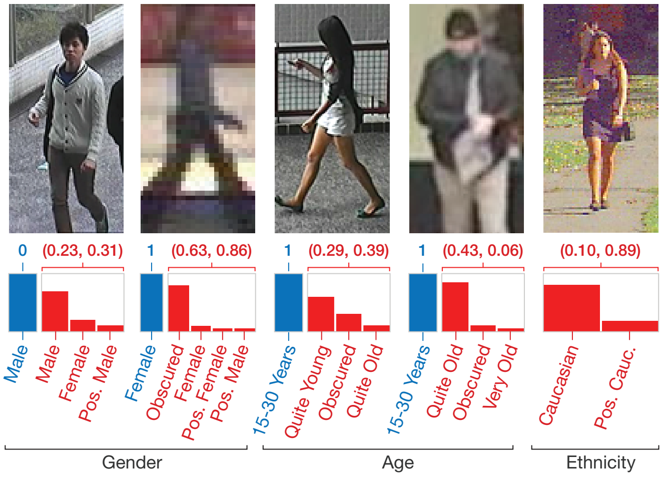


Fig. 1: **Conventional ground-truth (blue)** vs. **Super-fine (red)** labels on the PETA dataset. Conventional categories are coarse-grained and can be inconsistent and/or irrelevant. In contrast, super-fine annotations improve automatic subject retrieval with more precise and relevant descriptions, represented as multi-dimensional coordinates (red numbers).

Fine-grained [12] and relative attributes [13] have moved closer to solving true semantic image discrimination. These approaches use pre-defined sparse or ordinal representations to extend the number of traits and attribute precision. However, discrete categories can fail to capture subtle variations and one-dimensional orderings are unsuited to describing traits composed of abstract concepts e.g. *ethnicity*.

In practice, expert investigators and non-expert eyewitnesses must describe suspects with visual perception alone, as biological information is often unavailable. It is therefore imperative to discern *exactly what* users perceive in each image and to capture more illustrative subject descriptions by eliminating restrictive pre-determined vocabularies.

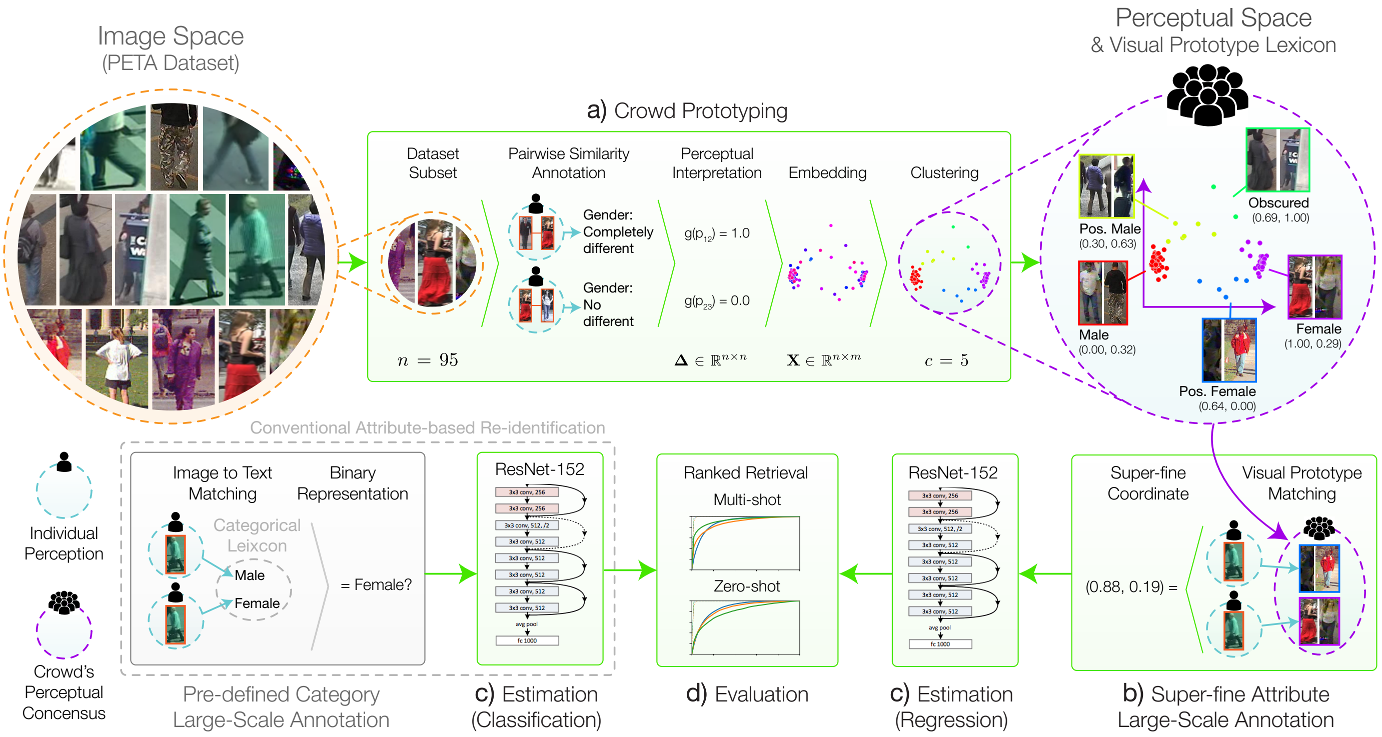


Fig. 2: Approach overview, **Contributions (green)**. **a)** Crowd prototyping previews the crowd's perception of an image subset to discover a perceptual space and salient visual prototypes (Section 4). **b)** Super-fine annotation efficiently matches unlabelled images to pre-discovered visual prototypes, generating super-fine coordinate labels (Section 5). In contrast, a conventional approach matches images to pre-defined categorical text labels. **c)** Image labels are estimated by fine-tuning the ResNet-152 CNN, classifying conventional binary attributes and regressing super-fine attributes (Section 6). **d)** Approaches are evaluated comparing ranked retrieval performance in multi-shot and zero-shot scenarios (Section 7).

Proposal. We introduce *super-fine attributes*. Similar to fine-grained and relative attributes, super-fine attributes describe both clear and subtle discriminations between images. Furthermore, super-fine attributes now simultaneously encapsulate multiple, integral concepts of a single trait as multi-dimensional coordinates. This enables vastly more powerful and intricate image descriptions. In Fig 1, rather than describing the second, highly obscured example as categorically *"Female"* or relatively *"14% Female"*, it can now be described as *"mostly Obscured but vaguely Female"* in relation to "Obscured" and "Female" image prototypes.

Importantly, these prototypes are linked to coordinates in a perceptual space that captures their relative distances. For example, "Quite Old" age prototype images are more similar and therefore closer to "Very Old" than "Very Young" prototype images. New images are then annotated by visually matching them to corresponding image prototypes, generating super-fine label coordinates from multiple prototype annotations in the perceptual space.

We propose *crowd prototyping* to discover such visual prototypes and map the crowd's consensus of each trait as a perceptual space, prior to large-scale annotation. This involves crowdsourcing similarity comparisons from a subset of images as in [14] (Fig. 2a). Similarity comparisons are then embedded to find a perceptual space (Fig 5), and clustered to find prototype images representing the most salient concepts (Fig 6). By first discovering prominent visual concepts and abstracting away from linguistic constraints, super-fine labels offer enhanced relevance and precision over expertly-defined alternatives.

The versatility of crowd prototyping is demonstrated by discovering and annotating gender, age and ethnicity traits on the PETA dataset. Our *domain agnostic* process is able to methodically discover objective visual discernments for all three traits, despite their traditionally fraught binary, ordinal and categorical measurement scales.

Lastly, we measure the effectiveness of annotating and estimating our novel super-fine attributes at scale. In all retrieval experiments, jointly regressed super-fine labels outperform automatically classified binary labels, noting that 3 super-fine attributes outperform 35 conventional attributes in a zero-shot scenario. Our radically new approach demonstrates that a popular image attribute recognition model can be employed to greater effect by *re-examining the target domain space*, applicable to any image retrieval context.

Contributions:

1) **Super-fine attributes**, a new form of image attribute which represent integral visual concepts as perceptual coordinates in multi-dimensional space.
2) **Crowd prototyping**, a domain agnostic method to discover salient visual concepts and facilitate super-fine attribute annotation of challenging images.
3) **Improved attribute-based image retrieval** and generalisation in zero-shot scenarios by jointly regressing continuous super-fine labels.
4) **Large-scale super-fine dataset**, the most comprehensive and precise visual appraisal of gender, age and ethnicity, covering 19000 surveillance images, available at: http://users.ecs.soton.ac.uk/dmc1g14.

The remainder of this work is organised as follows: Section 2 reviews topics influencing our methodology. Section 3 discusses key approach and dataset considerations. Section 4 details and evaluates crowd prototyping. Section 5 demonstrates efficient large-scale annotation, comparing super-fine and conventional labels. Section 6 discusses the estimation approach. Section 7 establishes the discriminative power of super-fine labels for ranked retrieval. Finally, Section 8 concludes our findings.

2 Related Work

We start by relating our work to soft biometrics and subsequently review the progression from categorical to ordinal and multi-dimensional attribute representations. This covers works within human attribute recognition, fine-grained attributes, relative attributes and attribute discovery. Lastly we discuss influential concepts from perceptual psychology.

2.1 Soft Biometrics

Soft biometrics studies the use of human describable characteristics for person identification. Visual soft biometric cues can be perceived at-a-distance, in partially obscured, occluded and very low quality surveillance footage and when hard biometrics e.g. fingerprint, gait or even face, are inapplicable. Two recent surveys encapsulate the wide range of traits [2] and modalities [6] employed in the field.

Identification using demography is fast becoming a primary motivation of study in biometrics. A recent biometric survey notes that biological sex and race cues are being superseded by sociological gender and ethnicity attributes, applicable in unconstrained surveillance environments where biological ground-truth is previously unobtainable [7].

2.2 Categorical Human Attribute Recognition

From face. One of the earliest approaches in human attribute recognition recognises gender from faces with a neural network [15] and suggests a special category is required for difficult to classify samples. More recently, the crowd's estimation accuracy of facial gender, age and race was surpassed by incorporating a quality-assessment method during recognition [8].

Convolutional neural networks (CNNs) are now pervasive in facial demographic recognition. Gender accuracy from challenging 'in-the-wild' datasets is up to 98% and 94% on CelebA and LFW [16]. For age, [17] achieve 84.7% (one-off) accuracy on Adience and [18] find that support-vector regression improves upon traditional classification. However, a 2012 survey in gender recognition compares face, gait and body modalities [11], reporting upwards of 99% accuracy from face but only 82.4% from body.

From body. In 2008, [9] first investigated the potential for identifying people using soft biometric semantic descriptions, finding race and sex to be most pertinent. Subsequently, [19] performed human identification using only automatically estimated soft biometric traits from face and body. In 2014, [20] introduced attributes-based re-identification (re-id), with 21 binary body descriptions on the popular VIPeR and PRID datasets, disclosing 4.5% annotator disagreement for the 'male' attribute and up to 14.5% for 'darkhair'. The paper also proposes the challenging zero-shot identification scenario, where images of the target suspect are previously unseen by the learning algorithm. Since then, advances in deep learning CNNs have permeated re-id, necessary in tackling the high degrees of inter-class variation [10] when estimating body attributes [21], [22], [23], [24], [25], [26], [27], [28].

Tackling the bottleneck of acquiring domain specific labels is often a primary focus. Two studies transfer clothing attributes from large, 'very fine-grained' fashion datasets to less constrained surveillance scenarios using an RCNN detector [28] and MRF-IBP model [29]. Rather than re-purposing potentially irrelevant attributes from disjoint domains, we are instead concerned with efficiently finding new and improved image descriptions within the surveillance domain. We do however transfer highly robust pre-trained feature representations learnt from a large corpus of data (ImageNet), by fine-tuning the ResNet-152 CNN.

Datasets & attributes. Since 2014, there has been an explosion in large-scale, 'in the wild' surveillance [24], [25], [26], [30], [31], [32] and people [33] datasets. We opt for the PETA dataset [30], amalgamating 10 prominent benchmark re-id datasets. Similarly to [27], we take an in-depth look at PETA's labels in Section 3 and explore challenging evaluation scenarios to inspect model robustness in Section 7.

Almost all conventional attribute recognition works either explicitly or implicitly aim to estimate the *biological ground-truth* of sex, age or race. However, we have mentioned several works that digress annotator disagreement [9], [20] and improve recognition for difficult to classify images with special categories [15], automatic detection [8] or manual removal [27]. Following [7], we believe that forcing the recovery of biological ground-truth from such challenging images can be misleading, especially considering annotation processes are often undocumented. Instead, we aim to first quantify the *crowd's visual perception* of the image dataset, in a more objective approach that captures different 'schools of thought' from within the crowd [34].

2.3 Fine-Grained & Relative Attributes

'Fine-grained attributes' has become a catch all term encapsulating relative attributes [13], just noticeable differences [35], large scale categorisation [31] and categorization at sub-ordinate levels [12]. Continuous and relative attributes are now widely accepted over traditional binary and multi-class annotations [36], as the field treads ever closer towards subtle and indistinguishable differences.

In 2009, [37] outperformed traditional binary classifiers by describing face regions with *pairwise similes* to a set of reference appearances e.g. "*a mouth like Barak Obama's*". In 2011, [13] introduced relative attributes as a means of ranking images by attribute strength using class-level *ordered pairwise comparisons* e.g. "*bears are furrier than giraffes*". The paper demonstrates how unseen, zero-shot object classes can be inferred by combining relative attributes with a similarity constrained RankSVM.

Contemporary soft biometrics demonstrate that comparative descriptions e.g. "*is subject A taller or shorter than B?*" are more objective, accurate and discriminative than categorical labels for unconstrained subject recognition; from the body [1], [38], face [39] and clothing [40].

Unfortunately, none of these relative attribute approaches can describe abstract traits like ethnicity, due to the restriction of single concept, one-dimensional representations. This motivates us to find a versatile, *super-fine* solution that can discover and represent descriptions given any manner of trait. We extend the work of [14] that compared a relative attributes approach to a one-dimensional embedding of similarity comparisons, finding an almost identical ordering of gender for 100 visually clear images. In contrast, when applied to the more visually ambiguous and unclear images from PETA, the technique found a much less divisive split between male and female image classes.

When dealing with subtle images, [35] highlights the psychophysics phenomena of 'just noticeable differences', concluding it is inappropriate to force a total order for any given image attribute. The approach instead infers when two images are indistinguishable using a Bayesian formulation to non-uniformly map low-level features and mid-level attributes. Alternatively, [41] accounts for differences between individuals' perceptions of an attribute by utilising an Adaptive SVM model that adjusts a generic model to user-specific notions. Tailoring the solution to individual users captures more precise intentions, improving binary and relative attribute image search. A third approach discovers shades of meaning for attributes, where a single linguistic definition may have a range of associated visual cues [34]. The paper addresses the gap between users' visual perception and their linguistic interpretation of an attribute.

Analogously to [34], we cluster a latent space to discover commonly perceived 'schools of thought' for each specific attribute. Although in contrast, we represent visual concepts through prototype images rather than pre-defined terms. We also account for common variations across all users with multiple dimensions of a single perceptual space per trait, rather than multiple adapted models as in [41].

2.4 Semantic Attribute Discovery

Attribute discovery originates in inferring descriptive classes e.g. "*fury with four legs*" from class-level names e.g. "*dog*" to alleviate expensive manual labelling [42]. These approaches find new semantic attributes by extracting co-occurrence patterns from low-level image features, enabling zero-shot recognition [43]. We instead find new attributes by quantifying the crowd's perception of a small subset of images, which are later automatically inferred from a much larger set of super-fine image labels.

In crowdsourcing, the vast majority of works expertly define categories, only collecting fixed labels from the crowd. Conversely, crowdclustering aims to discover categories and labels simultaneously through noisy crowd annotations [44]. Two such approaches employ grid-based similarity comparisons with a Baysian work preference model [44] and human-in-the-loop classifier [45]. Inspired by crowdclustering, we coin the term *crowd prototyping*, as a result of depicting crowdclusters with image prototypes.

Attribute discovery literature predominately entails finding distinct, separable categories or clusters by collecting *triplet comparisons* resulting in binary similarity measures e.g. "*is A more similar to B or C?*". This data is applicable to non-metric embedding, where only preserving the inter-point order is concerned. However, in relative attributes, pairwise comparisons often provide continuous distance measurements applicable to metric topological embeddings that seek to preserve inter-point distances.

Perceptual dissimilarity measures and interpretations are highly contentious. [46] argues against metric embedding, stating that dissimilarity magnitudes are unreliable and difficult to measure. On the other hand, [47] comprehensively investigates 5 types of similarity judgement, reporting that while triplet matching exhibits lower variance, pairwise ratings are less costly and also possess greater granularity. As super-fine attributes must embody higher descriptive power we employ *continuous pairwise comparisons* embedded with Metric Multidimensional Scaling (MDS) [48] building upon a preliminary study [14].

2.5 Perceptual Psychology

In 1927, [49] introduced the concept of a 'psychological continuum', whereby perceptual, rather than physical, discriminative measurements of object qualities are obtained from pairwise comparisons. Lately, [50] argues for a renewed focus on similarity as an explanatory concept, highlighting the task dependency of comparisons and the distinction between separable and integral (non-separable) concepts. Extensive soft biometric study has already uncovered a number of clearly *separable* trait descriptors e.g. gender, age and ethnicity. As such, rather than collect instance-level similarity measures as in [44], [45], we delve deeper into these trait subspaces, exploring their *integral* dimensions for the first time. A recent psychology study also explores the relationship between body shape and linguistic descriptions to generate realistic 3D avatars [3]. Analogously, we employ similarity spaces to supersede textual descriptions and enable linguistically unconstrained descriptions.

Due to the amorphous nature of similarity comparisons, measured conceptual spaces necessitate high-dimensionality to sufficiently store conceptually meaningful patterns. This has inspired a range of techniques to extract these non-linear patterns, such as MDS [48]. In our case, conceptual spaces are geometric structures, where points represent images and dimensions represent appearance qualities e.g. gender, age or clarity. We discover natural categories as convex geometric regions in conceptual space [51], bridging symbolic and associative cognitive representations. These regions of overlapping similarity resemble distinguishable characteristic families [52]. Such *prototypes* are depicted via a number of central images, representing the breadth of each clusters' conceptual region. New images can then be categorically matched to the most visually similar prototype.

3 Approach & Dataset Overview

In this section we outline considerations regarding our approach (Fig. 2), image dataset and performance evaluation, contrasted to conventional attribute-based identification.

Dataset. We focus on the PETA dataset, the most diverse re-identification dataset to date, amalgamating 19000 instances and 8699 unique identities across 10 prominent benchmark datasets; 3DPeS, CAVIAR4REID, CUHK, GRID, i-LIDS, MIT, PRID, SARC3D, TownCentre, VIPeR. It incorporates a very high degree of intra-class variation and is annotated with 108 binary attributes. The majority of attributes are extremely imbalanced, occurring in under 10% of the data, and do not include ethnicity, presumably due to its controversial nature. A number of works also experiment on PETA [21], [22], [23], [30], [53] for direct benchmarking.

Individuation. When conceiving an appropriate lexicon, we look towards the two principles of Leibniz's Law, raising broad ontological questions about how to individuate subjects in identity science [54]. The first principle, the 'Indiscernibility of Identicals' states, if i is identical to j, then i and j must have all the same P properties:

$$\forall i \forall j \left[i = j \rightarrow \forall P(P_i \leftrightarrow P_j) \right]$$

This holds true in the case of PETA, where i and j are subjects and image instances are labelled uniformly. However, subject-level labelling discards any intra-subject instance-level variation that may occur due to intensive changes in appearance, lighting or pose. This prohibits learning estimators that truly emulate human perception or generalise well in challenging scenarios.

We instead label image instances individually, estimating and evaluating both instance-level and subject-level image retrieval with super-fine attributes. Instance- and subject-level labelling are linked to theories of perdurance and endurance respectively. We find subject-level super-fine labels commonly attain the best performance in multi-shot retrieval, and instance-level alternatives to perform better in zero-shot retrieval. Yet all super-fine labels significantly outperform conventional subject-level binary attributes.

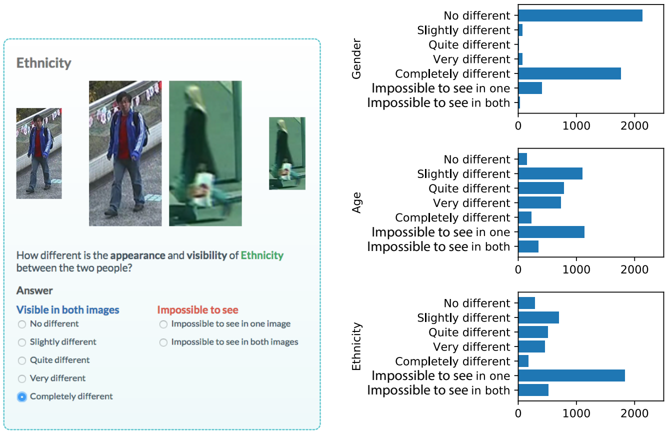
Fidelity. The second principle, the 'Identity of Indiscernibles' states, if i and j have all the same P properties, then i is identical to j:

$$\forall i \forall j \left[\forall P(P_i \leftrightarrow P_j) \rightarrow i = j \right]$$

This predicate does not hold true with PETA as only 7769 unique attribute configurations exist across all 8699 subjects. On selecting the 35 evaluation attributes, 218 subjects share a common description; male, under 30, casually dressed, carrying a backpack, with short black hair and sneakers. Worse still, when selecting only global and body traits just 128 total configurations exist. This highlights a stark problem with the current state of attribute-based re-identification; adding ever more attributes offers diminishing returns.

Consequently, we are motivated to strengthen the identity of indiscernibles principle with super-fine attributes, describing traits with greater fidelity and generating many more label configurations. However, increasing label precision can also impact relative estimation accuracy. To fairly evaluate this trade-off, we measure the gain in ranked retrieval performance given each set of estimated labels.

Representation. Traditional, one-size-fits-all lexicons often result in subjective labelling, suffering from annotator bias and anchoring effects [1]. This is especially apparent when relating textual descriptions to challenging surveillance imagery (Fig. 1). The perspicacity of our new approach lies in its ability to discern high-dimensional conceptual spaces, alleviating the constraints imposed by binary labelling and unary axes projection. By tailoring the label lexicon much more closely to image concepts perceived by the crowd, descriptions can be jointly represented in visual and conceptual domains (Fig. 2a-b). Intuitively, this leads to improved descriptive representations and more objective annotations, by matching images in visual space and accounting for confusion in conceptual space.



Ethnicity

How different is the **appearance** and **visibility** of Ethnicity between the two people?

Answer

Visible in both images
○ No different
○ Slightly different
○ Quite different
○ Very different
● Completely different

Impossible to see
○ Impossible to see in one image
○ Impossible to see in both images

(a) Example task question.

(b) Response distributions.

Fig. 3: Pairwise similarity task for crowd prototyping.

4 Crowd Prototyping

In this section we approximate the crowd's perceptual consensus of each trait, discovering a number of discrete visual prototypes and their relationships. This later facilitates efficient large-scale annotation of super-fine attributes. We detail a technique for collecting pairwise similarity comparisons, interpreting and embedding a topological perceptual consensus and clustering visual prototypes.

4.1 Pairwise Similarity Task

To capture the crowd's unconstrained perception of each trait, we propose a crowdsourcing task collecting $\binom{n}{2}$ pairwise similarity comparisons from a small subset of $n = 95$ subject images, as in [14]. In each question, respondents are asked to judge the appearance and visibility of a trait between two subject images, Fig. 3a. Images are displayed twice, once at their original resolution and a second scaled to fixed height for more direct side-by-side comparison.

Respondents rate their perception of trait dissimilarity on a 5-point Likert scale from "No different" to "Completely different". Alternatively, respondents may answer "Impossible to see in one image / both images" if there are no visible cues with which to make a distinction. This approach enables differentiation between ambiguity (open to more than interpretation) and uncertainty (having imperfect or unknown information).

We find binary measures ill-suited when describing challenging images, as in triplet and grid-based query solutions [44], [45]. We therefore elicit scaled responses with pairwise similarity comparisons, commonly used in judging perceptual distance between objects [1], [13], [55] and measuring qualitative properties without prior knowledge of their structure [14], [47]. It is widely acknowledged that counting the number of differences between objects is more objective than recalling their similarities, so we ask respondents to annotate intervals of dissimilarity. As perceptual

Pairwise Annotation	p_{ij}	u_{ij}
No different	0	0
Slightly different	0.25	0
Quite different	0.5	0
Very different	0.75	0
Completely different	1	0
Impossible to see in one image	1	1
Impossible to see in both images	0	1

TABLE 1: Pairwise annotations and associated proximity p_{ij} and uncertainty u_{ij} interval measures between subjects i and j.

judgements are asymmetric [55], we also randomly shuffle image display orders to regularise responses.

Respondents are assessed continuously throughout the annotation process, requiring a minimum of 80% test question accuracy. An initial quiz page of 10 test questions is presented, with remaining pages containing 9 genuine questions and 1 covert test question. Test questions are based on 100 initial responses, and crafted to allow an acceptable range of genuine responses without overzealous priming. This lets us assume a minimum level of annotator competence and quality, without modelling individual preferences.

4.2 Pairwise Similarity Task Results

Each task collects one response per pair for each trait, resulting in 13395 annotations from 614 respondents located across 62 countries. Fig. 3b presents the disparity in response distributions across each trait. Gender elicits a clear binary split between "No / Completely different", with few respondents answering "Impossible to see". However, age and ethnicity elicit many more "Impossible to see" responses from the same image set, showing that perceived uncertainty is non-uniform and trait specific. Respondents also shy away from extreme judgements of these two traits, indicating more indecision and the necessity of a continuous scale in allowing subtle differentiation.

Incidentally, each task consumed under $20 in crowdsourcing costs, yet provides a wealth of information. Though a single perceptual consensus is intangible, we later show with image recognition that our following efficient approximations far outperform pre-defined alternatives.

4.3 Interpretation Strategies

Table 1 represents dissimilarity annotations as pairwise proximities p_{ij} and uncertainties u_{ij} between subject images i and j, where $i, j \in 1, .., n$. A number of disparate theories interpret perceptual distances either with exponential decay [56], [57], a linear scale [48] or as binary measures [44], [45]. In light of this, we investigate three proximity mappings:

g^0, **5-Point linear**:

$$g^0(x) = x$$

g^1, **3-Point linear**:

$$g^1(x) = \begin{cases} 0.00, & \text{if } x \leq 0.25 \\ 0.50, & \text{if } x = 0.50 \\ 1.00, & \text{if } x \geq 0.75 \end{cases}$$

g^2, **Normalised exponential decay**:

$$g^2(x) = \frac{\exp(\lambda(1-x)) - e^{\lambda}}{e^{\lambda} - 1}$$

When applied to proximities p_{ij}, the two linear interpretations g^0 and g^1 represent 5-point and 3-point Likert scales respectively. Alternatively we experiment with normalised exponential decay g^2 using $\lambda \in \{-20, 4, 4, 20\}$. At $\lambda \in \{-20, 20\}$ the mapping is equivalent to a binary interpretation, taking only "Completely different" to mean "different", or only "No different" to mean "same".

In order to describe pairwise dissimilarities under one measure δ_{ij}, we assimilate both proximity and uncertainty annotations. Uncertainty measures are derived per subject u'_i as a fraction of all "Impossible to see" annotations:

$$u'_i = g\Big(\frac{\sum_{j \in N \wedge j \neq i} u_{ij}}{n-1}\Big).$$

The average pairwise uncertainty between two subjects i and j can then expressed as follows:

$$v_{ij} = |u'_i + u'_j|/2.$$

Lastly, mapped proximities $g(p_{ij})$ and absolute pairwise uncertainty difference $|u'_i - u'_j|$ are sum weighted by average uncertainties v_{ij}:

$$\delta_{ij} = (1 - v_{ij})g(p_{ij}) + v_{ij}|u'_i - u'_j|,$$

forming a symmetric, positive semidefinite distance matrix with zero diagonal $\mathbf{\Delta} = [\delta_{ij}] \in \mathbb{R}^{n \times n}$. Empirically, we find this weighted combination of uncertainty and dissimilarity produces the most stable and coherent embeddings, in comparison to interpreting uncertainty as a fixed distance, or ignoring pairwise constraints for uncertain annotations.

4.4 Perceptual Space Embedding

Given a high-dimensional distance matrix $\mathbf{\Delta}$, we employ non-linear, metric Multi-Dimensional Scaling (MDS) to find a low-dimensional conceptual space point configuration $\mathbf{X} = [\mathbf{x}_i] \in \mathbb{R}^{n \times m}$, where $m \ll n$. This is performed prior to prototype clustering, in order to minimise annotation noise and discover the most salient, global conceptual relations between subject images.

Metric MDS aims to preserve pairwise point distances in lower dimensions. In this case, a non-linear formulation is imperative, as dissimilarity annotations are limited in expressive range forming a non-convex proximity space. Highly dissimilar subject pairs will likely all elicit "Completely different" responses, yet could be describing the youngest to oldest pair, or youngest to second oldest etc. MDS is able to unwrap this space, preserving local neighbour distances relative to the global configuration.

MDS solutions are iteratively computed via the Scaling by Majorizing a Complicated Function (SMACOF) algorithm [48]. This considers minimising the normed Stress-1 function defined as:

$$\sigma_1(\mathbf{X}) = \frac{||d_{ij}(\mathbf{X}) - \delta_{ij}||_2}{||d_{ij}(\mathbf{X})||_2}.$$

where $d_{ij}(\mathbf{X})$ is the Euclidean distance between points x_i and x_j in $\mathbf{X}$. Each step increments k and computes the Guttman transform, updating $\mathbf{X}$ as follows:

$$\mathbf{X}^k \leftarrow n^{-1}\mathbf{B}(\mathbf{X}^{k-1})\mathbf{X}^{k-1},$$

where $\mathbf{B}(\mathbf{X})$ has elements:

$$b_{ij} = \begin{cases} -\delta_{ij}/d_{ij}(\mathbf{X}), & \text{for } j \neq i \wedge d_{ij}(\mathbf{X}) \neq 0 \\ 0, & \text{for } j \neq i \wedge d_{ij}(\mathbf{X}) = 0 \end{cases}$$

$$b_{ii} = -\Sigma_{j \in n \wedge j \neq i} b_{ij}$$

The algorithm iterates until convergence at $\sigma(\mathbf{X}^{k-1}) - \sigma(\mathbf{X}^k) < \epsilon$, selecting $\epsilon = 1e^{-4}$. Values $x_i \in \mathbf{X}$ are randomly initialised uniformly between $[0, 1]$ and the solution with minimal $\sigma_r(\mathbf{X})$ is selected after 1000 repetitions.

4.5 Embedding Strategy Evaluation

We now wish to find the most concise and accurate approximation to the crowd's consensus, aiming to select the proximity mapping that is best represented in the fewest number of embedded dimensions m, with minimal normed Stress-1 error σ_1 and maximal Spearman's rank correlation coefficient ρ between δ_{ij} and d_{ij}.

However, objective evaluation proves challenging, as we are generating entirely new ground-truths without relying on a prior gold-standard. Therefore, as a further measure of overall expressiveness, we only select embeddings that are at least partially cohesive to PETA's original categories. To do this, the silhouette score coefficient $-1 \leq s \leq 1$ measures the consistency of each original categorical cluster projected into the newly embedded space, where a is the mean intra-cluster distance and b the mean nearest, non-member cluster distance over all samples:

$$s = \begin{cases} 1 - a/b, & \text{if } a < b \\ 0, & \text{if } a = b \\ b/a - 1, & \text{if } a > b \end{cases}$$

Fig. 4 reports each metric, applying MDS with $m \in 1, ..., 6$ dimensions across all 7 distance matrices $\boldsymbol{\Delta}$ of the proximity mappings $g^0, g^1, g^2, \lambda \in \{-20, -4, 4, 20\}$ per trait. As m is incremented, each embedding approximation becomes more accurate, consistently minimising σ_1 errors and maximising ρ correlations. Proximity mapping $g^2, \lambda = -20$ exhibits the highest σ_1 embedding error, while its counterpart $g^2, \lambda = 20$ displays the lowest overall σ_1 loss and produces near perfect ρ correlations for age and ethnicity. However, the relatively low silhouette coefficients s of $g^2, \lambda = 20$ indicate it is a poorer interpretation of dissimilarity annotations for gender and age in comparison to other proximity mappings. This suggests that conceptual embeddings benefit from interval scale annotations, questioning the aptitude of binary alternatives emulated with $g^2, \lambda = \{-20, 20\}$.

Interestingly, embeddings of gender are much more closely related to the original categories than those for age, indicating they are more consistently perceived. In fact for age, only four proximity mappings ever produce $s > 0$; g^0, g^1 and $g^2, \lambda \in \{-4, 4\}$. Although not the closest approximation to the original distance matrix, we select g^1 and $m = 2$ for its cohesion to the original age and gender categories.

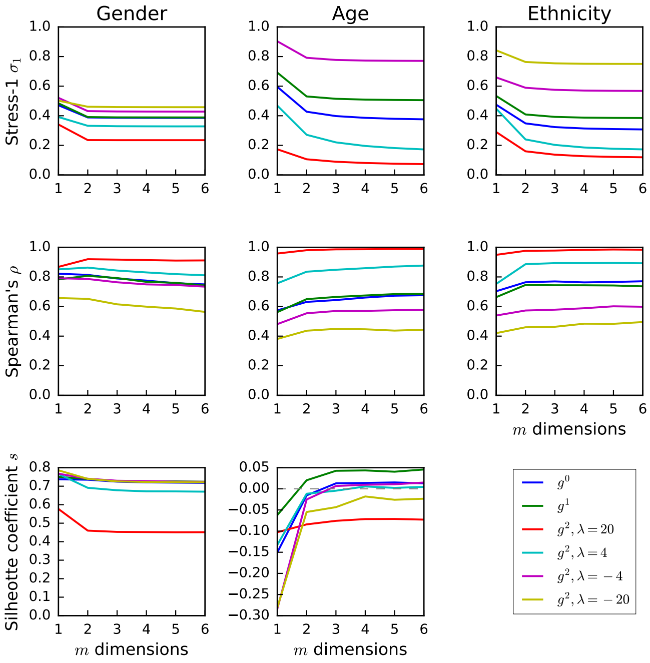


Fig. 4: MDS embedding strategy evaluation of g^0, g^1 and g^2, reporting the normed embedding Stress-1 error σ_1 (upper), Spearman's rank correlation coefficient ρ between δ_{ij} and d_{ij} (mid) and the silhouette score coefficient s of the original categories protected in embedded space (lower).

4.6 Perceptual Space Clustering

The final step of the crowd prototyping process is to draw visual prototypes for each perceptual trait space. We apply agglomerative hierarchical clustering to segment the embedded spaces using an L_1 average linkage criteria:

$$\frac{1}{|A||B|} \sum_{i \in A} \sum_{j \in B} ||x_i - x_j||_1,$$

where $|A|$ and $|B|$ are the cardinalities of any two clusters. Each point starts in its own cluster and at each iteration, the cluster pair with minimum linkage criteria are merged, until a maximum of c clusters remain.

We coarsely cluster each space into $c = 5$ clusters, and draw visual prototypes as the 8 closest images to each centroid. This ensures respondents are not overwhelmed with too many visual prototypes on-screen or indistinguishable options. We also manually name each cluster for ease of discussion in this article and during crowd annotation. New clusters and original categories are compared in Fig. 5 and example prototypes shown in Fig. 6.

Crowd prototyping gender discovers two clear gender classes, with the remaining three prototypes resembling varying levels of uncertainty (Fig. 5 gender upper). In fact, clustering almost perfectly divides the original categories (Fig. 5 gender lower), apart from two images previously labelled "Female", now each clustered into the "Possibly Male" and "Male" prototypes. This demonstrates how the effects of ambiguity and error on the original labels are mitigated with our more objective technique.

Prototyping age finds four clear age characteristics and one central obscured image group (Fig. 5 age upper & Fig. 6a). Unlike gender, the original categories are much more dispersed in the age embedding (Fig. 5 age lower) reflected in the lower overall silhouette scores (Fig. 4). Notably, the "Very young" and "< 15" classes are extracted identically.

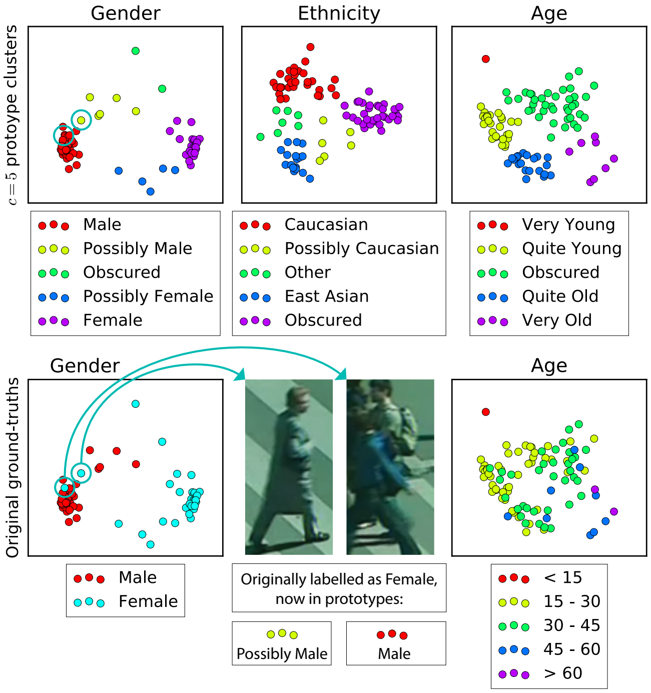
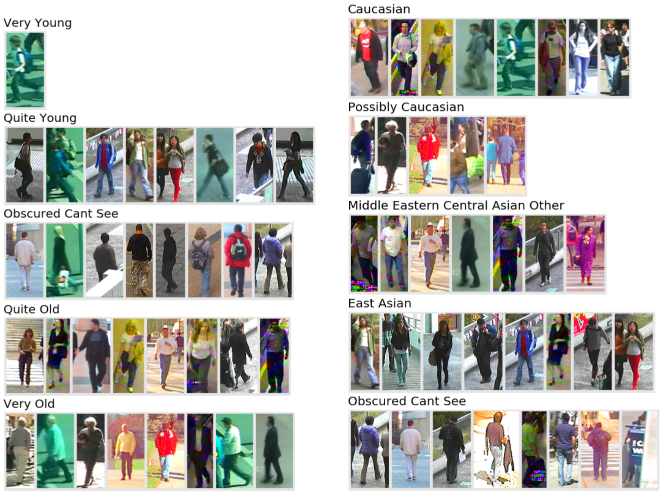


Fig. 5: Comparison of new prototype clusters (upper) against original ground-truth categories (lower) projected in embedded conceptual space. Each point represents one of the $N = 95$ images annotated with similarity. Two examples of conflicts between original and new gender prototypes are highlight.

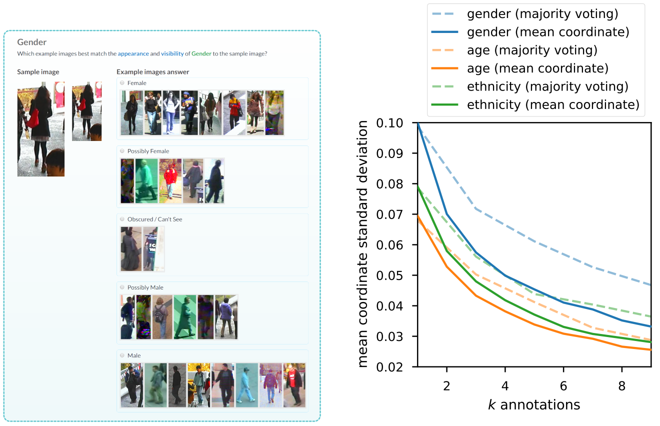


(a) Age. (b) Ethnicity.

Fig. 6: Discovered crowd prototypes, depicting distinct trait concepts from the crowd's perceptual consensus. Prototypes are formed by clustering conceptual spaces, selecting up to 8 images from each region's centroid. Semantic text descriptions are added manually for discussion.

Prototyping ethnicity finds clearly distinguished "Caucasian" and "East Asian" ethnicities to dominate the dataset (Fig. 5 ethnicity). Crowd prototyping also discovers certain concepts that are more specific than others e.g. "Caucasian" vs. "Middle Eastern / Central Asian / Other" (Fig. 6b), representing the prevalence and breadth of each characteristic contained within the dataset.

Prototypes are defined by the region of images they encapsulate, and so abstract away from pre-defined linguistic constraints. Traditionally, categorical lexicons consider only equidistant classes, but we have now also captured the relative distance between each prototype. In contrast to pre-defined categories or relative comparisons, where the precision is fixed, we can now increase labelling precision with further annotations. This is shown to successfully reduce annotation variance in Section 5.2 and subsequently improve overall subject retrieval in Section 7.


Gender

Which example images best match the appearance and visibility of Gender to the sample image?

Sample image

Example images answer

○ Female

○ Possibly Female

○ Obscured / Can't See

○ Possibly Male

○ Male

(a) Example image annotation matching gender prototypes.

(b) Mean coordinate standard deviations at k annotations.

Fig. 7: Large-scale visual prototype matching task.

5 Large-Scale Annotation

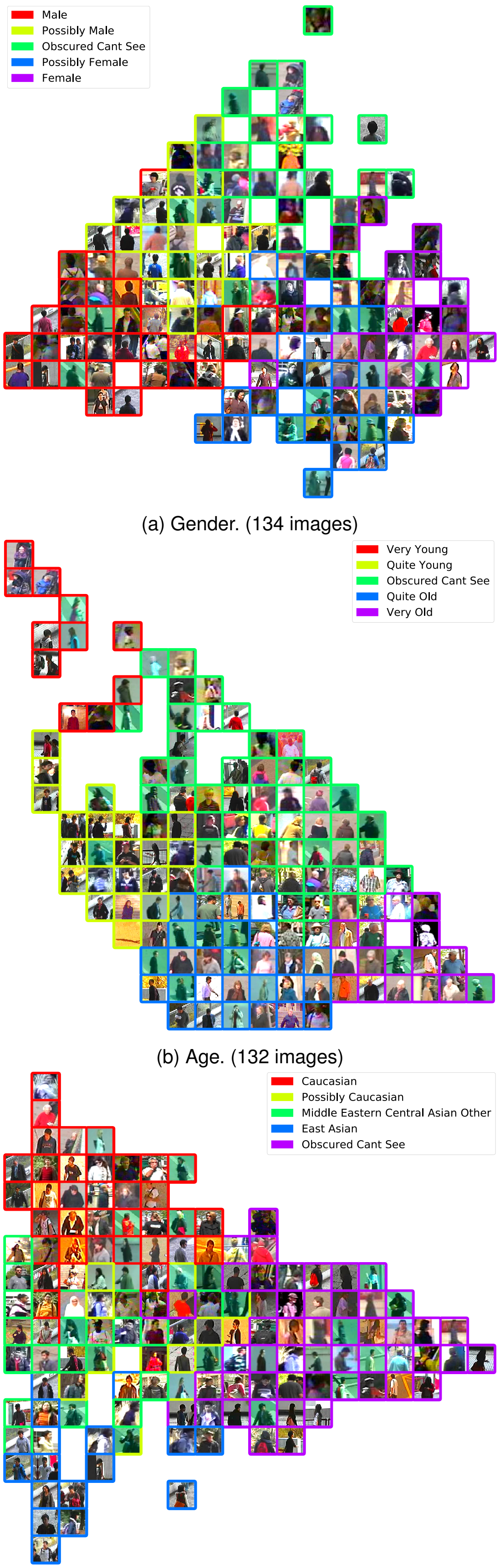
This section discuss our method for collecting super-fine attributes for all 19000 PETA image instances. We propose a prototype matching task that enables fast and efficient annotation at scale. We also evaluate the consistency of super-fine labels, and compare their distributions to conventional binary labels.

5.1 Visual Prototype Matching Task

The prototype matching task asks respondents to match new images to the most visually similar prototype. Each question displays one query image alongside 5 prototypes for one trait (Fig. 7a). Visual annotation enables rapid and intuitive categorical image labelling, requiring minimal effort and expense. The task is run similarly to the pairwise similarity task (Section 4.1), where test questions are drawn from 100 initial responses and respondents are constantly monitored for test question accuracy.

Matching images to visual categories is found to be more objective (and enjoyable) than annotating textual categories. However, it is still expected that repeat image annotations will elicit a number of different responses. This is entirely valid, and is the primary reason for associating visual prototypes with conceptual space coordinates via crowd prototyping. Rather than utilise a majority voting scheme, we can now generate more precise labels as the mean coordinate of matched prototypes. With the increased freedom super-fine labels offer us, we investigate two forms of label generation:

Instance-level labels. Where each label is assigned independently per image instance, calculated as the mean coordinate of image annotations.



(a) Gender. (134 images)

(b) Age. (132 images)

(c) Ethnicity. (140 images)

Fig. 8: Visualisation of large-scale super-fine annotations (subset of all 19K instances). Images are located at their annotated conceptual coordinates and head cropped for clarity. Border colours relate to median super-fine attribute annotation.

Subject-level labels. Where labels are assigned uniformly across each subject, as in traditional attribute-based re-identification. Each subject-level label is calculated as the average of one randomly selected annotation per subject image instance.

5.2 Annotation Consistency Analysis

To investigate the efficacy of the mean coordinate scheme, we collect 10 repeat annotations from 500 randomly selected images and simulate generating coordinates with $k \in 1, .., 9$ annotations per image. Fig 7b reports the mean standard deviation of instance-level simulated coordinates with both majority voting and mean calculation schemes at each k.

Mean coordinate calculations are clearly more consistent than the majority voting scheme. We find subject-level labels to perform similarly, but with uniformly incremented deviations of 0.01 across all k. Unexpectedly, gender has the highest overall variance, while age is lowest. This is in part due to more likely confused age concepts e.g. “Quite Young / Quite Old” existing closer in conceptual space than confused gender concepts e.g. “Possibly Female / Obscured”. Evidently a higher k is always more desirable in producing the most precise labels. With the knowledge of coordinate variance, we select $k = 3$ repeat annotations per image for the main task, balancing precision and cost.

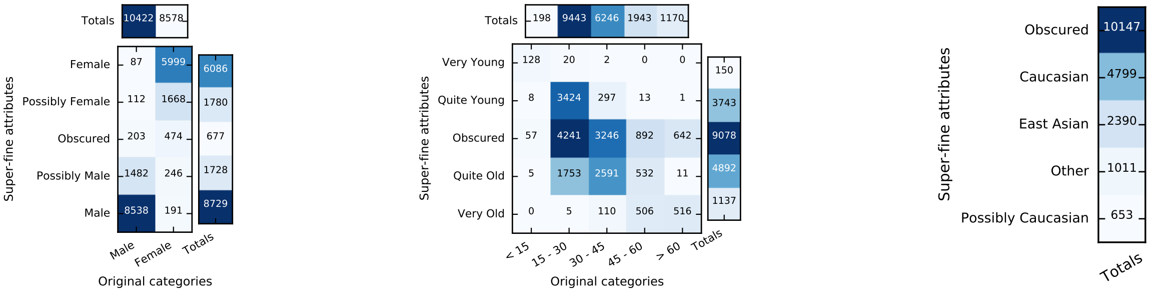
To inspect if our super-fine attribute scheme improves inter-annotator agreement over a traditional binary scheme, we simulate the 4.5% “male” class disagreement reported by [20] on the PRID dataset (which constitutes part of PETA). As with our newly discovered “Male“ and “Female“ prototypes, we measure the binary true-false classes at unit distance apart. Simulating this with mean coordinate calculation over 500 images, we find a standard deviation of 0.21 at $k = 1$, reduced to 0.12 at $k = 3$ annotations. This is substantially higher than any of the three super-fine traits displayed in Fig. 7b, indicating that annotating pre-discovered prototypes efficiently reduces label variation in perceptual space. This is because prototypes that are likely to exhibit conflicting annotations are mapped more closely in perceptual space, generating more certain coordinates.

5.3 Super-Fine Annotation Results

Over 1600 respondents contributed to the annotation tasks, collecting nearly 190000 responses and consuming under $250 per trait. Tasks were also highly rated by respondents, averaging 4.0/5.0, for clarity, ease, fairness and pay.

Fig. 8 visualises a cross-section of each annotated conceptual space, depicting image crops at their relative subject-level coordinates. Clear graduations can be seen between prototypes. For instance, images appear to gradually become older moving from “Very Young” towards “Quite Young” in Fig. 8b. Qualitatively, this demonstrates the rational behind averaging multiple coordinate annotations to generate image labels.

Fig. 9a-c (rhs) reports the annotation distributions of super-fine attributes at subject-level. We observe a large proportion of images labelled as “Obscured” for age and ethnicity in comparison to gender, similarly to the crowd prototypes. This is perhaps explained by the difficulty in distinguishing age and ethnicity without a clear view of



(a) Gender.

(b) Age.

(c) Ethnicity.

Fig. 9: Label distributions and confusion between subject-level super-fine attributes and original binary categories.

the face, whereas gender has many more visual cues to infer from, as can be seen in Fig. 8. In fact, a high number of obscure annotations may even benefit recognition by training estimators to better detect image clarity, discussed in Section 7.3.

5.4 Super-Fine vs. Binary Labels

Fig. 9 also shows the co-occurrence between super-fine and binary labels at subject-level. For gender, an exponential decay can be seen for the confusion of male and female labels, with only a small fraction of complete male-female reversals. This affirms the quality of our new labels in replicating previous schemes, while also aiding precision. On the other hand, age concurrence is much more dispersed, with the disproportionally large original "15-30" category spread across the three central super-fine attributes. Interestingly, the majority of clear images previously labelled "< 15" or "> 60" are classed as "Very Young" or "Very Old" respectively, yet more central ages exhibit weaker correlations.

We also note that original binary forms of gender and age have only 2 and 5 configurations respectively. Yet with super-fine attributes, instance-level labels have 380 gender and 322 age configurations and subject-level labels have 782 gender and 809 age configurations. Critically, images are still being visually assessed for cues pertaining to gender or age but super-fine representations produce several orders of magnitude more individuations. This is a huge step in discriminative power compared to conventional approaches. In the next sections we investigate just how much more effective they are for retrieval once automatically estimated.

6 CNN Attribute Estimation

For automatic label estimation, we select the ResNet-152 CNN model [58], which won the 2015 ImageNet detection, ImageNet localisation, COCO detection and COCO segmentation competitions. Its key advantage over previous image recognition CNNs are residual shortcut connections that enable inputs to skip layers forming multiple data paths inside the network, addressing training degradation of deeper networks. Our ResNet model is pre-trained on the ILSVR2012 challenge containing almost 100 times the number of images in PETA. This means that low-level feature descriptors are already extremely well generalised and provides a significant advantage over previous attribute recognition works that do not employ transfer learning.

The model accepts input image sizes of 224×224 so images are scaled to fit, regardless of their original dimensions. The penultimate layer average pools 2048 features, which are fully connected to the final $|\hat{y}|$ output logits. For fine-tuning the model, we employ the Adam optimiser with hyper-parameters $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$ and a high initial learning rate of $lr = 10^{-2}$, exponentially decayed by a factor of 0.96 per epoch. Models are trained over 24 epochs, at which point training set performance has stabilised and validation set metrics show no evidence for overfitting. Alternative loss functions are used for binary classification and super-fine regression:

Binary classification. The original high-level PETA traits are encoded into one-hot binary form. We model this as a multilabel classification problem, in which each binary class is independent and not mutually exclusive, employing sigmoid cross entropy loss:

$$L_{SCE} = -y * \hat{y} + \log(1 + \exp(\hat{y}))$$

Super-fine regression. This task is modelled as a joint regression problem, in which each perceptual coordinate axis is a continuous value in the range $\mathbb{R} \in [0, 1]$. We use mean squared error to calculate the loss function:

$$L_{MSE} = (y - \hat{y})^2$$

7 Experiments

We present three experiments investigating our estimation model and super-fine label performance. Section 7.1 introduces the estimation and retrieval methodology. Section 7.2 first benchmarks our model against previous works for binary estimation. Sections 7.3 and 7.4 then compare binary and super-fine label retrieval performance in multi-shot and zero-shot scenarios.

7.1 Estimation and Retrieval Methodology

In all experiments, the pre-trained ResNet-152 model is trained, validated and tested on disjoint data sets, reporting test set results averaged over three runs. As binary and super-fine approaches are either classified or regressed, comparing accuracy alone is not sufficient. Therefore, ranked retrieval measures the true discriminative power of each approach, incorporating both attribute precision and recognition accuracy in Sections 7.3 and 7.4.

Set-split criteria. In each experimental run, the dataset is split into three disjoint sets. A common criteria used for PETA is to randomly split the dataset at instance-level into sets of 50% train, 10% validation and 40% test. We call this *multi-shot*, as the training set can contain multiple instances of subjects also in the test set. However, multi-shot splitting can lead to significant estimation bias due to the high number of almost identical images, as mentioned in [27]. To fully contrast performance and highlight the challenges involved, we also experiment with splitting the dataset at subject-level as in *zero-shot identification* [20], [38]. Training and testing on disjoint subject image sets is more characteristic of a real-world operational scenario and as such proves to be far more challenging.

Labelling. The majority of attribute-based re-identification works assume subject-level labels, invariant to pose, viewpoint and environment. In practice, we find this is sometimes sub-optimal, especially with the increased fidelity of super-fine attributes, as significant variation is possible between instances of a subject. We therefore evaluate both instance- and subject-level labelling as discussed in Section 5.1.

Ranked retrieval. This process measures the efficacy of retrieving the corresponding image or subject from the set of all estimated labels, given a ground-truth eyewitness description. The ground-truth label set is ranked by either Hamming (binary) or Euclidean (super-fine) distance to each image label estimation, reporting ROC curves across a number of thresholds. A match is classed as a true positive if the retrieved label is of either the same image (instance-level) or the same subject (subject-level). We report instance- and subject-level labelling and retrieval synonymously.

7.2 Estimation Model Benchmark Results

Our first experiment compares our pre-trained ResNet-152 model to the three most recent works on PETA [21], [22], [23] that employ CNN architectures to jointly classify traditional binary labels in a multi-shot scenario. Table 2 reports the 'mean recognition accuracy' as $mA = (\frac{TP}{P} + \frac{TN}{N})/2$ [30], which accounts for imbalances in attributes' positive to negative sample ratio.

For multi-shot evaluation, MLCNN [21] is most accurate on highly imbalanced traits e.g. Age >60, Carrying Plastic Bag, Upper Jacket. However, DeepMAR [22] gains an average of 0.95% mA over our solution, optimally trading-off the accuracy of the many imbalanced traits against the fewer more balanced traits. To address the substantial class imbalances, these previous works all employ custom loss functions based on sigmoid cross entropy [22], [23] or softmax [21]. Such approaches benefit mean recognition accuracy in a multi-shot scenario, by prioritising under-represented classes that generalise across train and test sets, due to quasi-identical images. However, in a zero-shot scenario their discriminative performance may deteriorate, as under-represented classes no longer generalise well.

In contrast, we employ an unweighted sigmoid cross entropy loss function, which mitigates overfitting to under-represented classes. Crucially, we find this consistently outperforms previous approaches on the most balanced traits in Table 2 e.g. Gender, Age 15-30, Lower Trousers, Upper Other. Importantly, this further enhances our retrieval performance in the next experiments, as more balanced traits are inherently also the most discriminative.

		Multi-shot				**Zero-shot**
		MLCNN	DeepMAR	MAResNet	ResNet-152	ResNet-152
Attribute	ratio	[21]	[22]	[23]	**Ours**	**Ours**
Gender Male	54.9	84.34	89.90	76.60	**93.06±0.33**	83.96±0.56
Age 15-30	49.7	81.05	85.80	78.38	**87.28±0.48**	74.39±1.24
Age 30-45	32.9	79.87	81.80	75.55	**82.98±0.47**	60.60±3.10
Age 45-60	10.2	**92.84**	86.30	80.87	83.95±0.97	55.10±0.71
Age >60	6.2	**97.58**	94.80	86.29	92.74±0.53	56.59±4.08
Acs Hat	10.2	**96.05**	91.80	81.69	91.64±0.26	59.45±3.09
Acs Muffler	8.4	**97.17**	96.10	85.69	94.28±0.70	63.69±1.67
Acs Nothing	74.9	86.11	85.80	74.65	**87.16±0.44**	67.66±1.37
Acs S.Glasses	2.9	-	69.90	**76.18**	60.66±1.16	61.09±1.41
Cry Backpack	19.7	**84.30**	82.60	74.19	83.91±1.13	76.69±0.75
Cry M.Bag	29.6	79.58	82.00	71.99	**83.06±0.21**	64.80±1.38
Cry Nothing	27.6	80.14	83.10	71.31	**84.45±0.38**	68.34±0.56
Cry Other	19.9	**80.91**	77.30	69.71	75.31±0.45	54.31±0.73
Cry P.Bag	7.7	**93.45**	87.00	77.85	82.74±1.19	55.91±1.67
Ftwr Leather	29.6	85.26	87.30	79.11	**87.48±0.81**	66.91±1.64
Ftwr Sandals	2.0	-	67.30	**71.06**	63.56±2.51	63.91±2.45
Ftwr Shoes	36.3	75.78	**80.00**	70.33	79.22±0.49	60.29±0.63
Ftwr Sneakers	21.6	**81.78**	78.70	72.48	78.70±0.61	69.10±1.77
Hair Long	23.8	88.12	88.90	75.94	**89.81±0.22**	78.97±1.45
Lwr Casual	86.1	**90.54**	84.90	77.39	86.90±0.96	63.94±2.16
Lwr Formal	13.8	**90.86**	85.20	77.96	84.54±0.96	60.59±1.94
Lwr Jeans	30.6	83.13	85.70	73.67	**86.05±0.86**	74.62±1.22
Lwr Shorts	3.5	-	80.40	78.19	**80.57±1.50**	80.05±1.11
Lwr Skirt	4.6	-	**82.20**	72.37	78.93±1.52	62.28±0.47
Lwr Trousers	51.5	76.26	84.30	71.51	**85.75±0.70**	50.00±0.83
Upr Casual	85.3	**89.25**	84.40	75.20	86.58±0.37	62.98±1.34
Upr Formal	13.4	**91.12**	85.10	78.13	84.58±0.40	60.28±0.77
Upr Jacket	6.9	**92.34**	79.20	74.32	73.33±1.15	55.52±1.41
Upr Logo	4.0	-	68.40	66.77	**69.42±0.72**	63.93±1.43
Upr Other	45.6	81.97	86.10	78.72	**86.63±0.32**	75.21±0.48
Upr Plaid	2.7	-	81.10	71.87	**82.77±1.86**	76.58±3.44
Upr S.Sleeve	14.2	**88.09**	87.50	77.35	86.62±0.69	85.06±0.77
Upr Stripes	1.7	-	66.50	65.41	**67.26±1.33**	63.99±3.45
Upr T-shirt	8.4	**90.59**	83.00	75.62	77.16±0.73	76.47±0.24
Upr V-Neck	1.2	-	69.80	**75.63**	58.79±1.57	51.19±0.22
Average	-	-	**82.60**	75.43	81.65±0.83	66.45±1.49

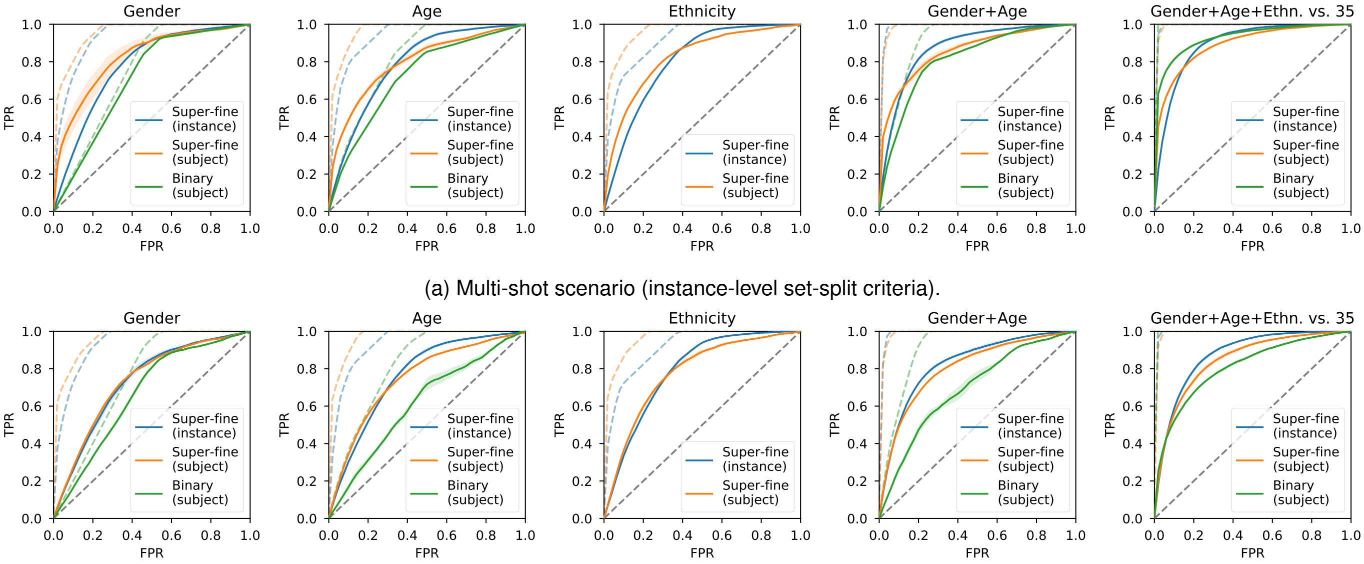
TABLE 2: Comparison of mean recognition accuracy (mA %) of binary attributes across previous PETA studies and our work. (**Bold**) indicates highest accuracy. Our approach initialises ResNet-152 with pre-trained weights from ILSVR2012, while MAResNet [23] initialises ResNet-152 weights randomly, accounting for a significant difference in performance.

Table 2 primarily indicates that our method is competitive under multi-shot evaluation. Naturally, the parity of implementation must be considered when comparing performance. Interestingly, our pre-trained model gains significantly higher results than MAResNet [23], which trains a similar ResNet architecture without pre-trained initialisation. In this scenario our ResNet-152 model achieves an average of 40.12±1.38% totally correct label estimations.

We repeat the experiment with zero-shot subject-level set-split criteria. This scenario attains only 3.31±0.32% totally correct image estimations, highlighting the challenge of training and testing on disjoint subject sets. The average mA is also 15.2% lower, reiterating the findings of [27], where gender accuracy rates dropped over 16% after removing quasi-identical images from PETA. Zero-shot evaluation evidences that much of the measured estimator accuracy is due to training on highly similar images in a multi-shot scenario, and less likely from learning truly generalised descriptors.

7.3 Multi-shot Retrieval Results

Our next experiment investigates the discriminative power of estimated super-fine labels in a multi-shot scenario. Train, validation and test sets are split at instance-level, potentially containing multiple instances of a subject across all sets.



(a) Multi-shot scenario (instance-level set-split criteria).

(b) Zero-shot scenario (subject-level set-split criteria).

Fig. 10: Ranked Retrieval ROC curves. Dotted lines indicate maximum performance assuming perfect label estimation. (TPR) True Positive Rate. (FPR) False Positive Rate. Shaded areas represent standard deviation.

	Attribute modality	Super-fine				Binary	
	Labelling	Instance-level		Subject-level		Subject-level	
Retrieval	Traits	R^2	**mAP**	R^2	**mAP**	Accuracy (%)	**mAP**
Multi-shot	Gender	0.680±0.013	0.776±0.002	0.725±0.073	**0.821±0.031**	92.20±0.54	0.709±0.003
	Age	0.622±0.024	0.788±0.007	0.549±0.025	**0.795±0.012**	79.85±0.46	0.723±0.002
	Ethnicity	0.773±0.001	0.801±0.001	0.696±0.006	**0.821±0.001**	-	-
	Gender+Age	0.663±0.015	**0.871±0.006**	0.631±0.007	0.854±0.010	75.43±0.75	0.813±0.004
	Gender+Age+Ethnicity	0.688±0.011	0.893±0.004	0.666±0.008	0.890±0.002	-	-
	All 35	-	-	-	-	40.23±1.36	**0.927±0.003**
Zero-shot	Gender	0.521±0.039	0.731±0.012	0.514±0.010	**0.733±0.002**	83.31±1.62	0.665±0.007
	Age	0.517±0.029	**0.762±0.007**	0.428±0.020	0.749±0.007	51.27±3.11	0.614±0.015
	Ethnicity	0.713±0.041	**0.789±0.003**	0.418±0.259	0.777±0.008	-	-
	Gender+Age	0.521±0.031	**0.826±0.006**	0.485±0.009	0.804±0.004	44.28±2.91	0.692±0.015
	Gender+Age+Ethnicity	0.610±0.031	**0.872±0.005**	0.456±0.117	0.845±0.007	-	-
	All 35	-	-	-	-	3.31±0.32	0.807±0.006

TABLE 3: Ranked retrieval results. (mAP) mean Average Precision. (R^2) coefficient of determination. (**Bold**) Highest mAP.

Fig 10a visualises the retrieval performance of 5 trait combinations, first comparing sole gender, age and ethnicity. From the outset it can be seen that both instance- and subject-level super-fine labels consistently outperform binary labels by up to 11.2% mAP and 7.2% mAP for gender and age respectively (Table 3).

Dotted lines indicate the maximum possible performance limit given perfect label estimations. Although age is represented with 5 classes, it only slightly enhances its limit in contrast to the 2 more balanced gender classes. In contrast, super-fine performance limits are considerably higher. This is significant, as super-fine labels are generated from the same visually perceived traits, yet enable many more distinctions without greater annotation effort. Remarkably, estimated super-fine labels also outperform the performance limits of binary gender and age, highlighting their discriminative power and capacity for automatic estimation. In other words, automatically estimated super-fine labels supersede the image retrieval performance of even perfectly estimated conventional binary attributes.

Furthermore, ethnicity appears to outperform both gender and age (Table 3), indicating its importance for suspect identification. This could be as a result of the high number of obscured annotations, leading to the remaining images containing more obvious visual features e.g. skin colour, improving automatic label inference. Intuitively, by improving attribute relevance we jointly enhance the accuracy and precision of both ground-truth and estimated labels.

Table 3 recalls the R^2 coefficient of determination for super-fine traits, reporting the proportion of explained variance between predictions and ground-truths. Interestingly, subject-level labelling produces lower R^2 scores for age and ethnicity, likely due to the increased number of configurations making learning more challenging. However, their mAP measures still remain higher than instance-level alternatives, reflecting the importance of label precision, even at the expense of relative estimation accuracy.

In the remaining results, concatenated super-fine gender and age labels outperform the binary equivalent, but all 35 binary attributes comfortably surpass all 3 super-fine traits in this scenario. Notably, in these results the number of instance-level label configurations are substantially increased when combined, therefore outperforming subject-level labels which are less prominently affected. Overall, binary label estimations operate much more closely to their upper limit than super-fine, indicating that an increase in precision is essential for greater retrieval performance. These promising results validate the need for super-fine attributes and highlight the versatility of the ResNet-152 model. The next experiment investigates just how robust each approach is in a more challenging scenario.

7.4 Zero-shot Retrieval Results

Our final experiment emulates a real-world surveillance scenario, where estimators are evaluated without prior exposure to similar instances during training. Train, validation and test sets are split at subject-level, such that no subject instances may exist in multiple sets simultaneously.

In comparison to multi-shot retrieval the overall performance is suppressed, due to the difficulty in mitigating overfitting. Learnt descriptors that overfit to subject identities are now penalised at test time, as they do not generalise to unseen subject instances. As such, subject-level binary labels fair particularly poorly, finding super-fine gender and age to improve mAPs up to 6.8% and 14.8% respectively in the zero-shot scenario.

Retrieval performance limits are not affected by alternative set-split criteria, yet in this experiment, instance- and subject-level super-fine labels achieve more comparable performance (Fig 10b). In contrast to multi-shot, instance-level labels are now more accurately estimated over subject-level labels. This is because they capture intra-subject variation and consequently, estimators can better generalise feature descriptors to totally dissimilar instances, reiterated by higher R^2 scores in Table 3. In particular, ethnicity appears to be the most robust trait, degrading the least in this scenario.

When combining gender and age estimations, we observe a 13.4% increase in mAP for instance-level super-fine labels over binary. Most notably, the retrieval performance of combined gender, age and ethnicity super-fine labels outperforms 35 binary attributes by 6.5% in zero-shot identification. This is testimony to how increasing precision and relevance of a select few traits can improve individuation. In practice, a trade-off must be made between perfectly estimating the exact semantic appearance of every image to learn truly robust descriptors, and learning labels that generalise across all instances of the same subject to enable subject identification. By jointly evaluating both instance- and subject-level labelling and retrieval, we reveal that such a trade-off is not a clear-cut choice, but that significant advantages can be gained from both methods.

8 Conclusions

We introduce *super-fine attributes* as a means of describing perceived visual variation, ambiguity and uncertainty in challenging images. We also propose *crowd prototyping* to pre-discover salient perceptual concepts that facilitate efficient large-scale annotation of super-fine attributes.

By substantially improving label relevance and fidelity over conventional approaches, we demonstrate the proficiency of super-fine labels for discrimination, ranked retrieval and generalisation in zero-shot scenarios. Our incorporation of unsupervised learning techniques with state-of-the-art supervised image recognition not only outperforms conventional approaches, but exceeds the maximum possible retrieval performance of binary gender and age descriptions with estimated super-fine labels in a multi-shot scenario. We also note that although crowd prototyping produces coarse-grained and highly varied categories, such categories are more pertinent than expertly-defined lexicons and visual annotation is more consistent than conventional text-based approaches. This article focuses primarily on employing super-fine attribute descriptions for suspect identification, providing in-depth analysis of gender, age and ethnicity traits from the highly diverse PETA surveillance dataset. However, our methodology is domain agnostic and applicable to a wide range of challenging images.

Looking towards the future, we believe knowledge representation is *the* grand challenge in computer vision and identity science. Although crowd prototyping and image recognition will undoubtedly be refined, the underlying premise of this work is to find higher-dimensional shared embeddings with which to communicate descriptions more effectively. Ultimately, for machine intelligence to truly emulate human behaviour, there must be a drastic increase in not only volume, but complexity of information conveyed between humans and machines to bridge the semantic gap.

References

[1] D. Reid, M. Nixon, and S. Stevenage, “Soft biometrics; human identification using comparative descriptions,” *TPAMI*, vol. 36, no. 6, pp. 1216–1228, 2013.

[2] M. S. Nixon, P. L. Correia, K. Nasrollahi, T. B. Moeslund, A. Hadid, and M. Tistarelli, “On soft biometrics,” *Pattern Recognition Letters*, vol. 68, pp. 218–230, 2015.

[3] M. Q. Hill, S. Streuber, and A. J. OToole, “Creating body shapes from verbal descriptions by linking similarity spaces,” *Psychological Science*, vol. 27, no. 11, pp. 1486–1497, 2016.

[4] “Codes of practice code d identification of persons by police officers,” Home Office, Tech. Rep., 2011, https://www.gov.uk/government/publications/pace-code-d-2011.

[5] “The pnc user manual,” National Policing Improvement Agency, Tech. Rep., 2012, http:/www.levesoninquiry.org.uk/wp-content/uploads/2012/04/Exhibit-KW-NIPA3.pdf.

[6] A. Dantcheva, P. Elia, and A. Ross, “What else does your biometric data reveal? a survey on soft biometrics,” *TIFS*, vol. 11, no. 3, pp. 441–467, 2016.

[7] Y. Sun, M. Zhang, Z. Sun, and T. Tan, “Demographic analysis from biometric data: Achievements, challenges, and new frontiers,” *TPAMI*, 2017.

[8] H. Han, C. Otto, X. Liu, and A. K. Jain, “Demographic estimation from face images: Human vs. machine performance,” *TPAMI*, vol. 37, no. 6, pp. 1148–1161, 2015.

[9] S. Samangooei, B. Guo, and M. S. Nixon, “The use of semantic human description as a soft biometric,” in *BTAS*. IEEE, 2008.

[10] L. Zheng, Y. Yang, and A. G. Hauptmann, “Person re-identification: Past, present and future,” *arXiv preprint arXiv:1610.02984*, 2016.

[11] C. Ng, Y. Tay, and B. Goi, “Recognizing human gender in computer vision: a survey,” in *PRICAI*. Springer, 2012, pp. 335–346.

[12] J. Deng, J. Krause, and L. Fei-Fei, “Fine-grained crowdsourcing for fine-grained recognition,” in *CVPR*. IEEE, 2013.

[13] D. Parikh and K. Grauman, “Relative attributes,” in *ICCV*. IEEE, 2011.

[14] D. Martinho-Corbishley, M. S. Nixon, and J. N. Carter, “On categorising gender in surveillance imagery,” in *BTAS*. IEEE, 2016.

[15] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski, “Sexnet: a neural network identifies sex from human faces.” in *NIPS*, 1990.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *ICCV*. IEEE, 2015.

[17] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *CVPR*. IEEE, 2015.

[18] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *WACV*. IEEE, 2015.

[19] A. Dantcheva, C. Velardo, A. Dangelo, and J.-L. Dugelay, "Bag of soft biometrics for person identification," *MTAS*, vol. 51, no. 2, pp. 739–777, 2011.

[20] R. Layne, T. M. Hospedales, and S. Gong, "Attributes-based re-identification," in *Person Re-Identification*. Springer, 2014, pp. 93–117.

[21] J. Zhu, S. Liao, Z. Lei, and S. Z. Li, "Multi-label convolutional neural network based pedestrian attribute classification," *Image and Vision Computing*, 2016.

[22] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *ACPR*. IEEE, 2015.

[23] E. Bekele, C. Narber, and W. Lawson, "Multi-attribute residual network (maresnet) for soft-biometrics recognition in surveillance scenarios," in *FG*. IEEE, 2017, pp. 386–393.

[24] 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.

[25] Y. Li, C. Huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *ECCV*. Springer, 2016.

[26] P. Sudowe, H. Spitzer, and B. Leibe, "Person attribute recognition with a jointly-trained holistic cnn model," in *ICCV*. IEEE, 2015.

[27] G. Antipov, S.-A. Berrani, N. Ruchaud, and J.-L. Dugelay, "Learned vs. hand-crafted features for pedestrian gender recognition," in *ACMMM*. ACM, 2015.

[28] 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.

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

[30] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *ACMMM*. ACM, 2014.

[31] D. Hall and P. Perona, "Fine-grained classification of pedestrians in video: Benchmark and state of the art," in *CVPR*. IEEE, 2015.

[32] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*. IEEE, 2015.

[33] G. Patterson and J. Hays, "Coco attributes: Attributes for people, animals, and objects," in *ECCV*. Springer, 2016.

[34] A. Kovashka and K. Grauman, "Discovering attribute shades of meaning with the crowd," *International Journal of Computer Vision*, vol. 114, no. 1, pp. 56–73, 2015.

[35] A. Yu and K. Grauman, "Just noticeable differences in visual attributes," in *ICCV*. IEEE, 2015.

[36] B. Qian, X. Wang, N. Cao, Y.-G. Jiang, and I. Davidson, "Learning multiple relative attributes with humans in the loop," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5573–5585, 2014.

[37] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *ICCV*. IEEE, 2009.

[38] D. Martinho-Corbishley, M. Nixon S, and J. Cater N, "Retrieving relative soft biometrics for semantic identification," in *ICPR*. Springer, 2016.

[39] N. Y. Almudhahka, M. S. Nixon, and J. S. Hare, "Unconstrained human identification using comparative facial soft biometrics," in *BTAS*. IEEE, 2016.

[40] E. S. Jaha and M. S. Nixon, "From clothing to identity: Manual and automatic soft biometrics," *TIFS*, vol. 11, no. 10, pp. 2377–2390, 2016.

[41] A. Kovashka and K. Grauman, "Attribute adaptation for personalized image search," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3432–3439.

[42] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*. IEEE, 2009.

[43] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *TPAMI*, vol. 36, no. 3, pp. 453–465, 2014.

[44] R. Gomes, P. Welinder, A. Krause, and P. Perona, "Crowdclustering," in *NIPS*. MIT Press, 2011.

[45] C. Wah, G. Horn, S. Branson, S. Maji, P. Perona, and S. Belongie, "Similarity comparisons for interactive fine-grained categorization," in *CVPR*. IEEE, 2014.

[46] S. Agarwal, J. Wills, L. Cayton, G. R. Lanckriet, D. J. Kriegman, and S. Belongie, "Generalized non-metric multidimensional scaling." in *AISTATS*, 2007, pp. 11–18.

[47] Ç. Demiralp, M. S. Bernstein, and J. Heer, "Learning perceptual kernels for visualization design," *TVCG*, vol. 20, no. 12, pp. 1933–1942, 2014.

[48] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[49] L. L. Thurstone, "A law of comparative judgment." *Psychological review*, vol. 34, no. 4, p. 273, 1927.

[50] S. Edelman and R. Shahbazi, "Renewing the respect for similarity," *Frontiers in computational neuroscience*, vol. 6, 2012.

[51] P. Gärdenfors, *Conceptual spaces: The geometry of thought*. MIT press, 2004.

[52] L. Wittgenstein, *Philosophical investigations*. John Wiley & Sons, 2010.

[53] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Learning to recognize pedestrian attribute," *arXiv preprint arXiv:1501.00901*, 2015.

[54] G. W. Leibniz, "Discourse on metaphysics," in *Philosophical papers and letters*. Springer, 1989, pp. 303–330.

[55] A. Tversky, "Features of similarity." *Psychological review*, vol. 84, no. 4, p. 327, 1977.

[56] R. N. Shepard *et al.*, "Toward a universal law of generalization for psychological science," *Science*, vol. 237, no. 4820, pp. 1317–1323, 1987.

[57] C. L. Krumhansl, "Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density." *Psychological Review*, vol. 85, no. 5, pp. 445–463, 1978.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*. IEEE, 2016.

Daniel Martinho-Corbishley received the MEng degree in Computer Science in 2014 and is now working towards his Ph.D. in human identification with soft biometrics for surveillance at the University of Southampton, U.K. His research interests include computer vision, deep learning and visual perception with an emphasis on semantic image discrimination and retrieval.

Mark S. Nixon is currently a Professor of Computer Vision with the University of Southampton, U.K. His research interests are in image processing and computer vision. His team were early workers in face recognition and later pioneers of gait recognition. He has chaired/program chaired many conferences (BMVC 98, AVBPA '03, FG '06, ICPR '04, ICB '09/15, and BTAS '10). He is a member of IAPR TC4 Biometrics, the IEEE Biometrics Council and Fellow of IET, IAPR, and BMVA.

John N. Carter (M'90) received the B.A. degree in experimental physics from Trinity College, Dublin, Ireland, and the Ph.D. degree in astrophysics from the University of Southampton, U.K. In 1985, he joined the School of Electronics and Computer Science as a Lecturer researching in signal and image processing, where he is currently a Senior Lecturer. His research interest is in the area of 4-D image processing, with applications in object tracking, feature detection, biometrics and automatic gait analysis.