UNIVERSITY OF SOUTHAMPTON

Discovering Human Descriptions for Ubiquitous Visual Identification

by

Daniel Martinho-Corbishley

A thesis submitted in partial fulfilment for the degree of Doctor of Philosophy

in the
Faculty of Physical Sciences and Engineering
School of Electronics and Computer Science

March 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Daniel Martinho-Corbishley

Identifying suspects in surveillance footage is paramount in ensuring public safety, preventing crime, policing and forensic investigation. At present, finding an individual in real-world CCTV footage given only an eye-witness description is near impossible. The vast majority of contemporary research assumes coarse, expertly-defined categories to describe subjects, ineffective in dealing with unconstrained, low quality and obscured images. Such brittle representations hamper semantic image discrimination and the ability to learn robust predictors from challenging subject matter.

This thesis explores human and machine centric techniques for representing and learning semantic human descriptions for suspect identification. By investigating the duality of human-machine communication, we enhance the capabilities of traditional attributes and soft biometric descriptors, expanding their versatility and applicability towards challenging images and large-scale surveillance datasets.

We experiment with crowdsourcing human annotations using ordered and similarity comparisons, and estimating attributes from images employing a variety of state-of-the-art machine learning techniques. Our focus is on utilising a lean lexicon of global and body characteristics that are most pertinent when estimated from stand-alone surveillance footage. Significant improvements in suspect retrieval and identification performance are achieved by discovering enhanced soft biometric descriptions which represent visual trait characteristics with more precision and relevance.

This work evolves the areas of soft biometrics and identity science, drawing ideas from contemporary image attribute recognition, semantic attribute discovery, pedestrian re-identification and perceptual psychology. Our findings indicate that increasing not only the volume, but the complexity of information conveyed between humans and machines is key in deploying soft biometrics ubiquitously.

Contents

Declaration of Authorship

I, Daniel Martinho-Corbishley, declare that the thesis entitled and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published, listed as [**Publications p1-7**] under Section 1.3.

Signed:...

Date:..

Acknowledgements

I would like to thank everyone who dedicated their time and energy to discussing the plethora of topics that helped realise this thesis.

First and foremost I would like to thank Prof. Mark Nixon for his resolute guidance, continued optimism and empowering me to venture off-piste in search of new questions and answers while ensuring a thesis materialised at the end of the journey. I would also like to thank Dr. John Carter for introducing me to biometrics, first suggesting a Ph.D. and for his attentive technical advice over the last five years.

I extend my graditude to all my friends, colleagues and members of the Southampton University Mountaineering Club for their fellowship. To my housemates for their encouragement and countless enlightening conversions. Also, to the hundreds of online annotators for their candid judgements.

Of course, I wish to thank my parents and grandparents Eva, Ray, Manuel and Janina for their enduring love and support. Finally, to Lizzie for her exceptional advice and companionship - best of luck on your own Ph.D. journey!

Chapter 1

Context & Contributions

1.1 Context

Identifying individuals in surveillance imagery is now critical in thwarting terror attacks, preventing crime and ensuring public safety. In 2013 there were an estimated 5.9 million CCTV cameras in the UK[1]. Worldwide, internet video surveillance traffic is projected to increase tenfold between 2015 and 2020[2]. Consequently, the vast quantities of video data being generated are rendering manual approaches to suspect identification infeasible. Novel and pragmatic solutions are now urgently required to tackle such large volumes of data and an ever increasing number of surveillance environments.

The aim of this thesis is to automatically identify suspects in large crowds of people, given only an eyewitness testimony. This involves retrieving relevant images from unconstrained surveillance footage, based on a stand-alone human description. An example could be: *"Find the possibly mixed race, young adult male"*. In this scenario, a human operator generates the suspect query, while the search system narrows the suspect image list by automatically estimating and matching subject descriptions. Notice that the query may be unclear, ambiguous or nuanced, yet both 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.

We propose a novel method to discover improved subject descriptions in large-scale, real-world surveillance footage, utilising the knowledge that stand-alone human descriptions are capable for suspect identification [1] and that comparative descriptions substantially improve identification performance from human vision [2]. Throughout this work we explore issues pertaining to the semantic gap, knowledge representation and human-machine communication in the pursuit of truly ubiquitous soft biometric identification.

[1] http://www.arc24.co.uk/how-many-cctv-cameras-actually-are-there-in-the-uk/

[2] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf

1.2 Contributions

The two primary innovations of this work are:

1. **Attribute discovery for identification**. Building upon previous work in relative and fine-grained attributes, we introduce our new super-fine attributes, enabling the description of any form of trait, in any image context, in unprecedented detail. In conjunction, we propose crowd prototyping as a means of efficiently crowdsourcing super-fine attributes, encapsulating the crowd's perceptual consensus for objective, non-verbal visual description. This extends the applicability of soft biometrics, alleviating the need for restrictive, expertly-defined vocabularies and enabling precise descriptions of challenging surveillance images.
2. **State-of-the-art automatic identification**. Utilising the latest techniques in machine learning and computer vision, we enhance automatic identification performance from unconstrained surveillance footage using stand-alone human descriptions. Experiments in estimating our novel labels establish the necessity for relative and super-fine attributes in identification, which offer increased labelling precision and relevance over traditional categorical and binary representations.

As a result, we highlight a number of contributions as follows:

1. **Describing real-world surveillance imagery**. Cutting-edge research in *global* and *body* soft biometrics, which provides more prevalent information than face-based cues, is transitioned from synthetic, lab-based datasets towards real-world datasets. Visual concepts of obscurity, clarity and ambiguity are addressed for the first time in attribute-based re-identification (Chapters 5, 6).
2. **High fidelity datasets**. Two of the most detailed and comprehensive pedestrian attribute datasets are made available, in terms of ground-truth granularity (SoBiR relative attributes, Chapter 4) and scale (PETA super-fine attributes, Chapter 6).
3. **Harnessing human visual perception**. A paradigm shift in the notion of attribute labelling is presented. Moving away from attempting to record exact biological ground-truths, towards eliciting consistent labels from the crowd's consensus, based only on human visual perception (Chapters 2, 3, 5).
4. **Deep learning regression for identification**. One of the first approaches to experiment with regressing relative, continuous and multi-dimensional attributes with deep learning algorithms for attribute-based identification (Chapters 4, 6).
5. **Similarity comparisons**. Introducing an innovative form of soft biometric annotation utilising pairwise similarity comparisons, providing an unconstrained way

to discover salient concepts. Abstracting away from fixed textual or ordered representations enables more objective image-based descriptions (Chapters 5, 6).

6. **Unconstrained trait investigations**. The first time *gender* has been annotated comparatively from body with ordered and similarity comparisons (Chapters 3, 5). Furthermore, the first time typically binary, ordinal and nominal, *gender*, *age* and *ethnicity* traits have been annotated with a universal process, without imposing prior definitions or constraints (Chapters 6).

7. **Labelling modality utilisation**. We highlight the objectivity of ordered comparisons (Chapter 3), the versatility of similarity comparisons (Chapter 5) and the efficiency of categorical annotation (Chapter 6). Harnessing the advantages of each modality, we facilitate accurate and precise labelling methods, applicable to large-scale and a broad ranging datasets, constrained only by annotation cost.

8. **Fewer traits in more detail**. By focussing on refining already well established soft biometric traits we alleviate the need for ever more attribute descriptions, instead exploring innovative semantic spaces to improve automatic identification and reduce annotation costs (Chapters 3, 5, 6).

9. **Instance- and subject-level labelling**. In contrast to traditional subject-level labelling, we experiment with instance-level labelling, demonstrating its benefits in zero-shot identification. Findings suggest that individual image annotation may be more suitable for real-world deployment (Chapter 6).

1.3 Publications

The follow publications are based on this research:

[p1] D. Martinho-Corbishley, M.S. Nixon, J.N. Carter, "Soft biometric recognition from comparative crowdsourced annotations," in the International Conference on Imaging for Crime Prevention and Detection (ICDP), IEEE, 2015. doi:10.1049/ic.2015.0101.

[p2] D. Martinho-Corbishley, M.S. Nixon, J.N. Carter, "Soft biometric retrieval to describe and identify surveillance images," in the International Conference on Identity, Security and Behavior Analysis (ISBA), IEEE, 2016. doi:10.1109/ISBA.2016.7477240.

[p3] D. Martinho-Corbishley, M.S. Nixon, J.N. Carter, "Analysing comparative soft biometrics from crowdsourced annotations," in IET Biometrics, 2016. doi:10.1049/iet-bmt.2015.0118.

[p4] D. Martinho-Corbishley, M.S. Nixon, J.N. Carter, "On categorising gender in surveillance imagery," in Biometrics Theory, Applications and Systems (BTAS), IEEE, 2016. doi:10.1109/BTAS.2016.7791189.

[p5] D. Martinho-Corbishley, M.S. Nixon, J.N. Carter, "Retrieving relative soft biometrics for semantic identification," in International Conference on Pattern Recognition (ICPR), IEEE, 2016. doi:10.1109/ICPR.2016.7900105.

[p6] M.S. Nixon, B.H. Guo, S.V. Stevenage, E.S. Jaha, N. Almudhahka, D. Martinho-Corbishley, "Towards automated eyewitness descriptions: describing the face, body and clothing for recognition," in Visual Cognition, 2016. doi:10.1080/13506285.2016.1266426

[p7] D. Martinho-Corbishley, M.S. Nixon, J.N. Carter, "Super-fine attributes with crowd prototyping," in submission to Transactions on Pattern Analysis and Machine Intelligence (TPAMI), IEEE, 2017.

1.4 Soft Biometric Processes & Definitions

Soft biometric identification requires humans and machines to communicate visual descriptions of subject image identities via a set *semantic attributes*. This involves two distinct activities, whereby 1) humans and 2) machines independently translate high-dimensional images to lower-dimensional semantic labels. The first step, 1) *soft biometric annotation*, collects objective subject descriptions from human annotators. The second step, 2) *automatic soft biometric recognition*, focuses on machine learning techniques for accurately estimating ground-truth labels from images. Both steps jointly affect label precision, *soft biometric retrieval* accuracy and overall soft biometric identification performance. As such, dual themes of human perception and machine learning are presented through this report. Figure 1.1 visualises the intersection of the four key topic areas; human perception, machine learning, pedestrian surveillance images and soft biometrics.

Soft Biometrics; visual human descriptions composed of semantic attributes.

Soft Biometric Trait; a single human characteristic e.g. *gender*, described with a semantic attribute e.g. 'female' (a.k.a. soft trait)[3].

Soft Biometric Annotation; objective visual descriptions of subject images by human annotators, generating semantic labels (a.k.a. crowdsourced labelling).

Soft Biometric Recognition; automatically predicting semantic attribute labels from subject images, estimating target ground-truth descriptions (a.k.a. image attribute recognition with machine learning and computer vision techniques).

Soft Biometric Retrieval; matching target ground-truth and estimated labels to automatically retrieve subjects based on their soft biometric description (a.k.a. Content-Based Image Retrieval (CBIR)).

Soft Biometric Identification; the end-to-end process of annotation, recognition and retrieval, to automatically identify a suspect from a database of subjects, given only an eyewitness description.

[3] Words pertaining to soft traits are *italicised* for clarity throughout this thesis e.g. *gender*, *age*, *ethnicity*.

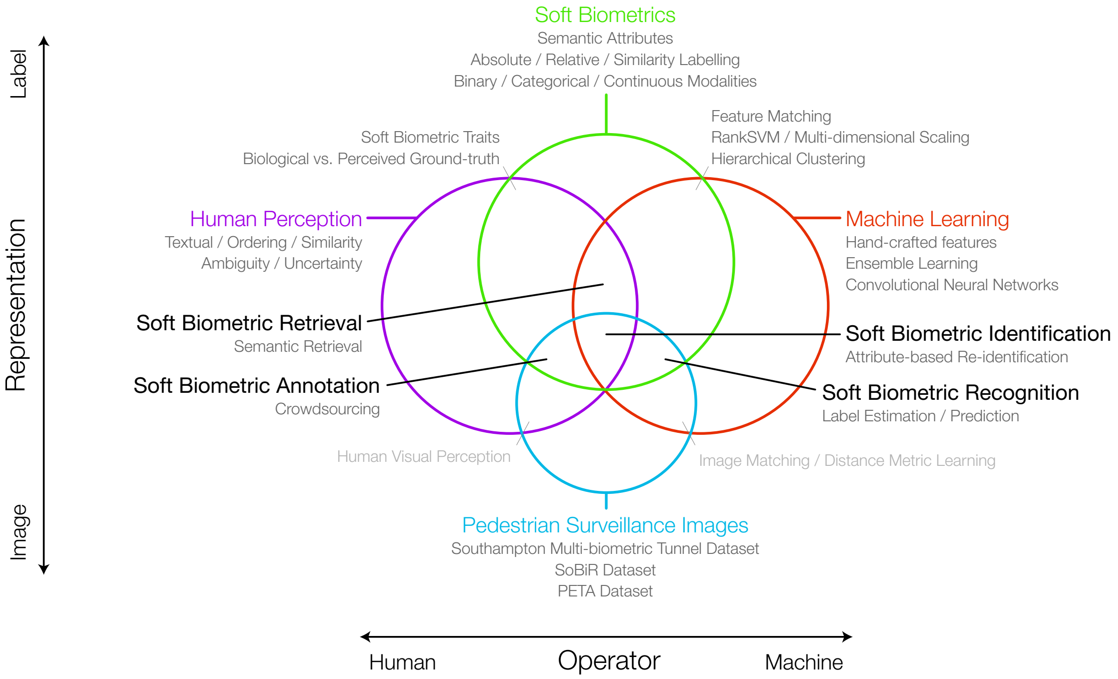

Soft Biometrics
Semantic Attributes
Absolute / Relative / Similarity Labelling
Binary / Categorical / Continuous Modalities
Soft Biometric Traits
Biological vs. Perceived Ground-truth
Feature Matching
RankSVM / Multi-dimensional Scaling
Hierarchical Clustering
Human Perception
Textual / Ordering / Similarity
Ambiguity / Uncertainty
Machine Learning
Hand-crafted features
Ensemble Learning
Convolutional Neural Networks
Soft Biometric Retrieval
Semantic Retrieval
Soft Biometric Identification
Attribute-based Re-identification
Soft Biometric Annotation
Crowdsourcing
Soft Biometric Recognition
Label Estimation / Prediction
Human Visual Perception
Image Matching / Distance Metric Learning
Pedestrian Surveillance Images
Southampton Multi-biometric Tunnel Dataset
SoBiR Dataset
PETA Dataset
Label
Representation
Image
Human
Operator
Machine

Figure 1.1: An overview of how humans and machines can communicate subject descriptions from images to soft biometric labels. Each sphere represents one of the four topic areas; human perception, machine learning, pedestrian surveillance images and soft biometrics. Overlaps between spheres represent the intersection between several topic areas. The horizontal axis represents communication between human and machine operators, while the vertical axis represents translation between high-dimensional images and low-dimensional labels. Soft biometric annotation is the human process of perceiving images and translating them to soft biometric ground-truths. Soft biometric recognition is the machine learning process of estimating soft biometrics labels given an image input. Soft biometric retrieval is the process of searching for a subject, given only a soft biometric description. Soft biometric identification is the intersection of all four areas, whereby retrieval is subsequently applied to human annotations and machine estimations of subject images. Importantly, human-machine subject communication can not occur through the image space alone, but images must first be independently translated to low-dimensional soft biometric representations. In later Chapters 5 and 6 we address this gap by discovering higher-dimensional soft biometric descriptions, more closely aligned to the original subject images.

1.5 Thesis Overview

Chapter 2. We start by discussing influential literature pertaining to soft biometrics, human image attribute recognition, pedestrian re-identification and image attribute discovery. We attempt to intertwine traditionally disjoint human and machine centric studies, highlighting their overlapping affects on this work.

The remaining body of this report continues the dual themes of human-machine communication, alternating between semantic image discrimination techniques in Chapters 3 and 5 and automatic label estimation in Chapters 4 and 6. By transitioning back and forth between the two, we progressively iterate upon our findings, starting with relative attributes in Chapters 3 and 4 and evolving super-fine attributes in Chapters 5 and 6.

Chapter 3 [Publications p1, p3]. This chapter looks at what can be achieved by human vision, before mimicking its abilities by automated means. It investigates soft biometric annotation, building upon previous, cutting-edge works by crowdsourcing a dataset of ordered comparative annotations [1, 2]. We utilise a number of visual cues to describe person images captured in a controlled environment [3], ranking subjects through relative-continuous labels. For the first time gender is included as a comparative trait and we confirm that relative-continuous labels characteristically contain additional discriminative information over traditional categorical annotations.

Chapter 4 [Publications p2, p5]. Next, we investigate automatic soft biometric identification with relative attributes from stand-alone images for the first time. We introduce a novel, publicly accessible dataset, SoBiR, comprising four modalities of soft biometric ground-truth. To tackle the challenges of large intra-class variation and inter-class ambiguity we investigate two label estimation approaches; classical feature-based learning and state-of-the-art convolutional neural networks. Our findings indicate that relative-continuous labels are not only more discriminative, but consistently outperform labels derived from widely used absolute annotation methods. Furthermore, we demonstrate the power of deep learning in enhancing label estimation accuracy and subsequent semantic identification performance across three scenarios.

Chapters 3 and 4 definitively show that relative comparisons outperform absolute categorical annotations, when identifying a limited number of subjects captured in a controlled environment. However, by generating such precise labels from pairwise comparisons, difficulties arise when attempting to describe abstract concepts and scaling to very large datasets that contain visual uncertainty resulting from very low quality images.

Chapter 5 [Publication p4]. We take the first steps in addressing these newly presented challenges. The chapter discusses categorising *gender* from two datasets; SoBiR and PETA, an especially challenging, large-scale benchmark dataset. Our novel approach discovers super fine-grained visual taxonomies of *gender* from pairwise similarity comparisons, annotated via crowdsourcing. We demonstrate the ability to describe multiple, integral concepts, including ambiguity and uncertainty, that go beyond biological binary male-female designators. By forming a perceptual consensus from the crowd, our method generates more discriminative, reproducible and flexible labels over coarse-grained, expertly-defined categorical vocabularies.

Chapter 6 [Publication p7]. Lastly, we demonstrate the effectiveness of our radically new super-fine labels for automatic identification, eliciting *gender*, *age* and *ethnicity* annotations from the PETA dataset. We propose crowd prototyping to discover salient visual concepts and generate a discrete set of image prototypes per trait, prior to large-scale annotation. Previewing the crowd's perception of each trait ensures newly annotated labels pertain to the dataset, enhancing relevance, precision and accuracy over expertly-defined alternatives. Large-scale annotation then matches new images to corresponding prototypes, efficiently representing attributes as coordinate labels. Utilising a deep transfer learning approach, we outperform previous approaches on PETA for conventional binary classification. Better still, regressed super-fine labels outperform automatically classified binary labels across all identification scenarios, demonstrating that accurately modelling subjectivity and uncertainty is key to learning robust estimators. Conclusively, in the challenging zero-shot scenario we highlight that our 3 super-fine attributes outperform 35 conventional attributes for ranked retrieval and note the potential advantages of instance-level over subject-level labelling.

Chapter 7. Reviews the work and proposes a number of avenues for future research.

Chapter 2

On Human Identification from Visual Descriptions

2.1 Introduction

Visual characteristics are intrinsic to the human identity [1, 2, 4–8], making them essential in non-contact surveillance, forensic investigation and for communicating suspect identities via eye-witness testimony. This chapter investigates how such descriptions are currently discovered and utilised for automatic subject identification and ways in which they may be generalised for universal application.

Section 2.2 starts with anthropometry, concerning the use of salient, precise and accurate physical characteristics for traditional *biological* subject identification. When operating in the *visual* domain, human descriptions become dependent on perceptible features, intrinsic to each image context. Section 2.3 grounds our work in soft biometrics, cataloguing the transition from biological, to perceived visual descriptions for identification.

Facilitating *automatic* identification from such descriptions requires joint human perception and machine vision interpretations to be objective and consistent. This problem is known as the semantic gap [9, 10], which lies at the intersection of human-centric psychology and machine-centric image attribute recognition and attribute-based re-identification studies. Furthermore, enabling *ubiquitous* visual identification requires annotation methods to scale efficiently and adapt to broad surveillance contexts.

As such, we intertwine prominent ideas presented across soft biometrics, human- and machine-centric literature, discussing the succession of categorical to ordinal and multi-dimensional attribute representations, covering works in human attribute recognition (Section 2.4), crowdsourcing (Section 2.5), fine-grained and relative attributes (Section 2.6), and semantic attribute discovery (Section 2.7). Lastly, Section 2.8 summarises our findings.

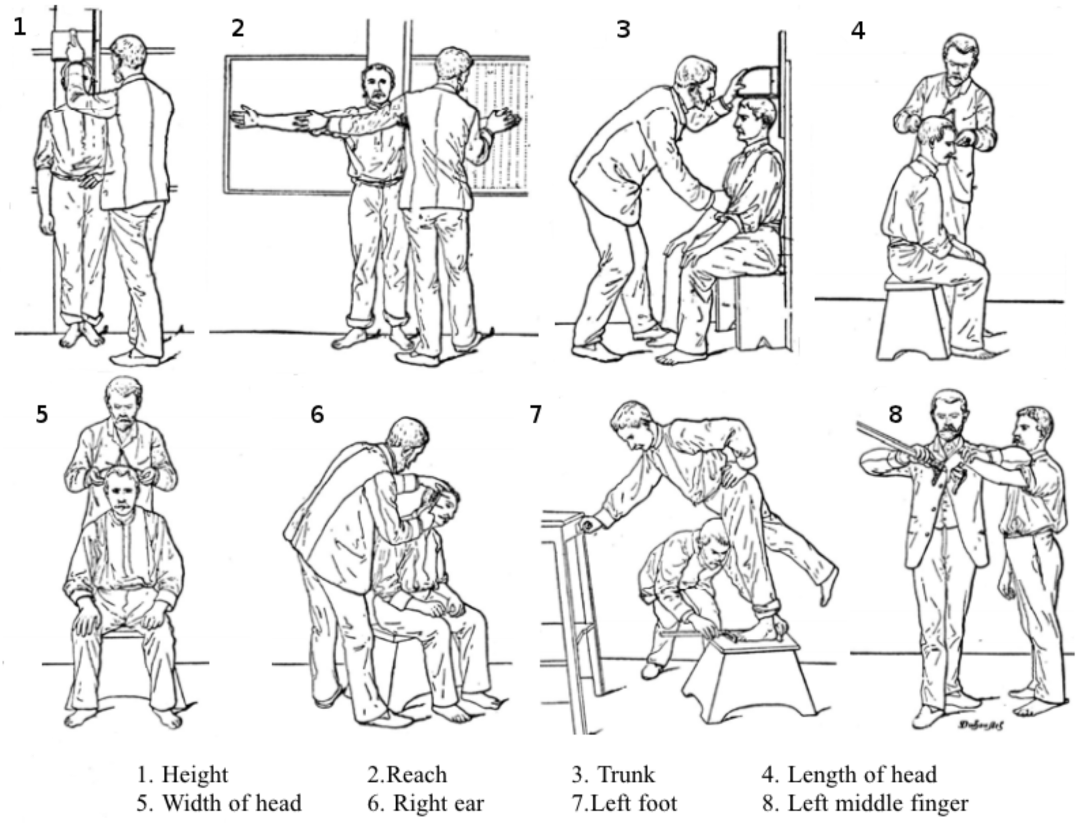


Figure 2.1: Illustration extract from the Bertillonage police reference manual for gathering anthropometric measurements [11]

2.2 Anthropometry

In 1809, working as a clerk in the Perfecture of Police in Paris, Alphonse Bertillion witnessed at first-hand the failings of the identification and cataloguing system intended to supersede the brutal practice of criminal branding. Following his father's anthropological work, Bertillion developed a method for identifying people based on physiological traits, becoming the first pioneer of systematic person identification. Serving as Chief of the Judicial Identification Service of France in 1896, he published 'The Bertillon System of Identification' [11], proposing a formal procedure for recording such features to identify repeat offenders, aptly named 'Bertillonage'. The system included measurements of standing height, right ear and nose dimensions, eye and hair colour, and indelible marks, illustrated in Figure 2.1.

The success of Bertillonage came from its ability to reduce the probability of false positives, as the chance of two individuals possessing similar measurements for all 13 features is very unlikely. It was later displaced by the discovery of unique fingerprints, which are able to identify suspects from marks left at the crime scene, needing no prior criminal offence for system enrolment. Nevertheless, with the proliferation of Closed Circuit Television (CCTV) in modern day society, Lucas et al. argue that suspects' anatomical characteristics can often be gleaned from crime scene video footage [12].

Lucas et al. find that body measurements are more variable than the face, and therefore more pertinent to identification [12], influencing our focus on *global* and *body* descriptions. Importantly, larger anthropometric dimensions also make them less likely to be occluded and easier to locate in low quality images. The study finds that using a total of 8 distances between skeletal points (disregarding *sex* and *ethnicity* measures), identification is possible with a 10^{-20} chance of collision, comparable to fingerprint identification.

Further to this, strong correlations are shown to exist between groups of body measurements [13] e.g. lengths and heights, ankle and hand circumferences, hip and thigh circumferences etc. Adjeroh et al. exploit this information to accurately estimate *weight* and *gender* traits, employing a two-step logistic regression and Principle Component Analysis (PCA) approach [13]. Recently, Kakadiaris et al. classify *gender* from whole-body measurements, applying an SVM+ algorithm to anthropometric ratios, mitigating noise from computer vision-based measurements [14]. Crucially, these studies demonstrate that discriminative measurements can be inferred from human metrology at-a-distance, even when subjects are partially occluded, or faces are unobservable, as in common in real-world surveillance footage.

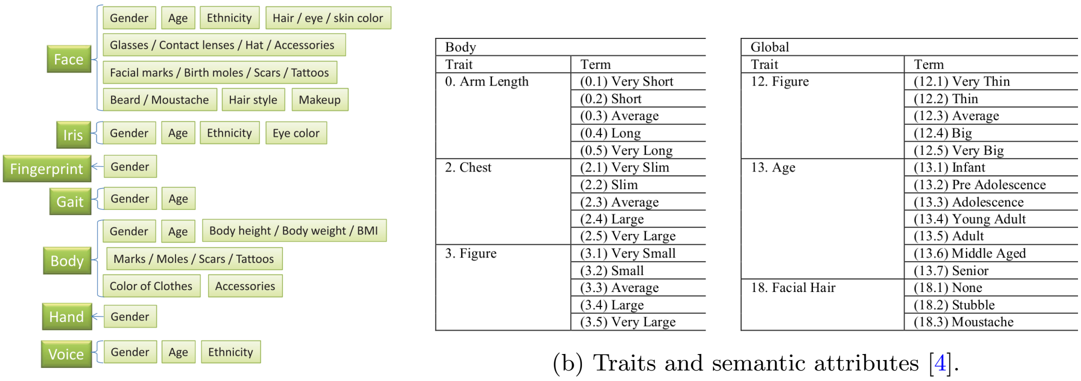
2.3 Soft Biometrics

2.3.1 Hard to Soft Biometrics

Biometrics is the science of identifying individuals based on physical characteristics [15]. Typically, these characteristics are select human features that convey sufficient distinguishing information to perform identification. Principally, biometrics establish an individual's identity based on *who they are* rather than by *what they possess* [16]. Classical *hard* biometrics, including face, fingerprint, iris, voice, gait and DNA, are based on measuring highly discriminative features, unique to every individual.

Although hugely powerful, hard biometrics are restricted to situations in which the subject's biometric signature is pre-enrolled in a controlled environment. This makes them unsuitable for use in unconstrained surveillance scenarios, remote non-cooperative identification 'at-a-distance', or from images captured 'in-the-wild'. Like anthropometric measurements, traditional biometrics also lead to a semantic gap between a human's perception and the established biometric signature, limiting their applicability for subject search based on eye-witness testimony.

Soft biometrics study the use of human describable characteristics for person identification. They are a new branch of biometric in identity science, relying only on human perception to systematically describe and identify subjects, addressing many of the issues pertaining to remote surveillance. The power of soft biometrics is in utilising visual cues



Body	
Trait	Term
0. Arm Length	(0.1) Very Short
	(0.2) Short
	(0.3) Average
	(0.4) Long
	(0.5) Very Long
2. Chest	(2.1) Very Slim
	(2.2) Slim
	(2.3) Average
	(2.4) Large
	(2.5) Very Large
3. Figure	(3.1) Very Small
	(3.2) Small
	(3.3) Average
	(3.4) Large
	(3.5) Very Large

Global	
Trait	Term
12. Figure	(12.1) Very Thin
	(12.2) Thin
	(12.3) Average
	(12.4) Big
	(12.5) Very Big
13. Age	(13.1) Infant
	(13.2) Pre Adolescence
	(13.3) Adolescence
	(13.4) Young Adult
	(13.5) Adult
	(13.6) Middle Aged
	(13.7) Senior
18. Facial Hair	(18.1) None
	(18.2) Stubble
	(18.3) Moustache

(b) Traits and semantic attributes [4].

(a) Modalities and traits [5].

Figure 2.2: Overview of soft biometric modalities, traits and semantic attribute hierarchy.

that can be perceived in partially obscured, occluded and very low quality surveillance footage, and when hard biometrics prove unobtainable.

2.3.2 In Literature

Jain et al. first coined the term 'soft biometrics' in 2004, noting that ancillary information was not utilised in many automatic identification systems [17]. However, when a system falsely rejects a user, human operators often verify individuals using visual cues alone. By incorporating characteristics such as *gender*, *ethnicity* and *height* into a fingerprint system's decision making process, its performance was found to improve, reducing the need for manual intervention [18].

Two recent surveys encapsulate the wide range of modalities [5] and traits [4] employed in the field, reviewing techniques for demographic estimation from face, clothing, iris and even hand or ear, Figure 2.2. Dantcheva et al. introduce a taxonomy for organising soft biometrics attributes encapsulating demographic, anthropometric, medical, material and behavioural attributes. The review highlights the benefits of soft biometrics as a human understandable interpretation for video surveillance, providing robustness to low quality images and consent-free acquisition [5]. Nixon et al. discuss the evolution of soft biometrics from the original Bertillonage, to the search for discriminative measures and their estimation from surveillance imagery. The survey defines the area as *"the estimation or use of personal characteristics describable by humans that can be used to aid or effect person recognition"* [4].

Earlier definitions claim that soft biometric traits *"lack distinctiveness and permanence"* but are *"easily distinguished at a distance"* [17, 19, 20]. Later works state that soft biometrics have *"stronger invariance properties than vision-based (hard) biometrics"* [21]. While a concrete definition is yet to be settled upon, most agree that soft biometrics are *"human describable physical traits"* [1, 2, 4, 22]. We provide our own definition

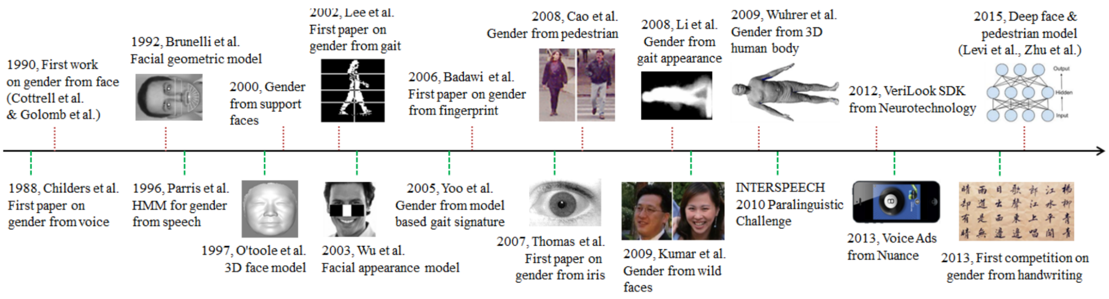

1990, First work on gender from face (Cottrell et al. & Golomb et al.)
1992, Brunelli et al. Facial geometric model
2000, Gender from support faces
2002, Lee et al. First paper on gender from gait
2006, Badawi et al. First paper on gender from fingerprint
2008, Cao et al. Gender from pedestrian
2008, Li et al. Gender from gait appearance
2009, Wuhrer et al. Gender from 3D human body
2012, VeriLook SDK from Neurotechnology
2015, Deep face & pedestrian model (Levi et al., Zhu et al.)
Output
Hidden
Input
1988, Childers et al. First paper on gender from voice
1996, Parris et al. HMM for gender from speech
1997, O'toole et al. 3D face model
2003, Wu et al. Facial appearance model
2005, Yoo et al. Gender from model based gait signature
2007, Thomas et al. First paper on gender from iris
2009, Kumar et al. Gender from wild faces
INTERSPEECH 2010 Paralinguistic Challenge
2013, Voice Ads from Nuance
2013, First competition on gender from handwriting

Figure 2.3: Major milestones in the history of automatic gender estimation from biometric data [6].

for this thesis, emphasising stand-alone identification and the transition from physical, biological descriptions, to perceived, visual semantics:

> *Soft biometrics are visual descriptions composed of semantic attributes which enable stand-alone human identification.*

2.3.3 Applications

Originally, biometric identification studies focused on fusion techniques, combining *ancillary* soft biometric information with hard biometrics like gait [1, 2, 13], face [19, 23] or fingerprints [18, 24]. Although still relying upon hard biometric pre-enrolment, these earlier works gradually uncover the importance of supplementary soft biometric information.

Now, the recent branch of *stand-alone* soft biometrics investigates identification with sole human descriptions, as in studies dealing exclusively with body [22], face [25] and clothing [26]. This means subjects need not be pre-enrolled, opening up exciting opportunities such as Content-Based Image Retrieval (CBIR), human accessible search queries with only verbal description, and circumventing the need to manually comb large banks of archived video footage for forensic investigation. This is a compelling premise for our work; to investigate the power of stand-alone soft biometrics in performing identification, showing they provide more than just subsidiary information.

Person descriptions are already employed in critical policing applications such as the Police National Computer (PNC) [27]. It encapsulates an enormous variety of identifiers, including documentation e.g. passport numbers and DNA reports, through to personal characteristics e.g. *marks*, *scars* and *shoe size*, denoted with pre-defined nominal attributes. The 1984 Police and Criminal Evidence Act (Code D) sets out the code of practice for the identification of persons by police officers [28]. The code details *age*, *sex*, *race* and *clothing style* as common traits mentioned in eye-witness procedures when suspect identities are unknown. For *race*, 7 Identity Codes (ICs) are defined to describe police officers' perceived views of an individual's ethnicity, as opposed to the individual's self-definition. However, stand-alone descriptive identification is yet to be totally ubiquitous. Feris et al. address the practical implications of attribute-based people search in live surveillance environments, stating that automated facial identification would have been valuable in apprehending the 2013 Boston Marathon bombing suspects [8].

However, even facial recognition is not suitable for totally ubiquitous identification in surveillance. In fact, we ran the entire PETA dataset of CCTV captured pedestrian images through DeepFace, a start-of-the-art face detection algorithm [29]. Of 19000 body images, DeepFace detected just 28 faces, 2 of which were false detections, highlighting the stark limitations of relying on the face for real-world identification.

Very recently, Sun et al. reviews the estimation of *gender*, *age* and *race* from body and face [6], noting that identification from demography is fast becoming a primary motivation of study, superseding hard biometrics for its pervasive and universal qualities. Figure 2.3 chronicles the timeline of gender recognition research across several modalities, demonstrating its progression from small scale, constrained datasets to real-world pedestrian imagery. The biometric survey notes that *race* and *ethnicity*, like *sex* and *gender*, are related to independent biological and sociological factors respectively, hence our focus on sociologically perceived traits e.g. *gender*, *figure* and *ethnicity*.

Another recent and broad ranging survey encapsulates up to the minute developments and emergent topics in biometrics [16]. Akhtar et al. suggest "*the future is almost here*", with regards to the ubiquity of biometric authentication in modern society, owing to the proliferation of wearable, mobile sensors e.g. GPS, fingerprint, rf-capture, gyrometers, ECG and the accessibility of big data e.g. online social profiling [16]. With such advances, the survey notes that behavioural identification and prediction is becoming vital in surveillance, giving us a glimpse into the near future of biometric applications.

The following sections delve into the techniques used to recognise and describe human attributes, incorporating the first categorical soft biometric approaches by Dantcheva and Samangooei [1, 22] in Section 2.4.2 and the subsequent relative attributes approach by Reid [2] in Section 2.6.1.

2.4 Categorical Human Attribute Recognition

2.4.1 From Face

Human attribute recognition is vast area of research, originating in facial recognition as a prominent area of modern computer vision. One of the earliest approaches is Golomb et al.'s 1990's SexNet, a neural network to recognise gender from faces [30]. Even at this early stage, the paper suggests forming a special category for difficult to classify samples, finding a human annotation error rate of 11.6%. A 2012 survey in gender recognition compares face, gait and body modalities [31], reporting upwards of 99% accuracy from face but only 82.4% from body. Since then, Convolutional Neural Networks (CNNs) have become pervasive in facial demographic recognition. Liu et al. attain up to 98% and 94% *gender* accuracy from challenging 'in-the-wild' datasets CelebA and LFW [32].

For *age* recognition, Levi et al. achieve 84.7% (off-by-one) accuracy on Adience [33] and Wang et al. find that modelling *age* via support-vector regression improves upon traditional classification [34]. Lately, Han et al. outperform the online crowd's estimation of *age* by incorporating a quality-assessment method with hierarchical classification, meanwhile estimating *gender* and *race* as binary attributes [35].

2.4.2 From Body

In 2008, Samangooei et al. first investigate the potential for identifying people from the *body*, using soft biometric descriptions [1]. The study selects 23 traits for their universality, distinctiveness, permanence and collectability, represented with mutually exclusive categorical semantic attributes. For *race*, the most prominently mentioned categories are discovered from psychology, witness analysis and human identification literature. Overall, it is found that *race* and *sex* are the most pertinent traits, concluding that humans have the ability to consistently identify individuals at-a-distance and under varying conditions using higher level semantic concepts. Since then, human metrology research has demonstrated how the body offers improved recognition over the face in less constrained scenarios [12–14].

Subsequently, Dantcheva et al. take the next step by performing human identification using only automatically estimated soft biometric traits from face and body [22]. Semantic attributes are recognised using a variety of computer vision techniques including facial landmark localisation, colour histogram extraction and thresholding alongside Gaussian mixture models and fuzzy clustering methods. Experimenting on the most challenging dataset at the time, VIPeR [36], finds colour histograms the most robust descriptors. These works lead into a corresponding area now refereed to as *attribute-based re-identification*.

2.4.3 Attribute-based Re-identification

Pedestrian re-identification (re-id) has received huge attention from the computer vision community, centred on matching individuals across multi-camera networks. Approaches that claim high identification rates often to do so by matching images directly to one another through metric learning [38, 39], or by fusing auxiliary semantic attributes with low-level image features [37, 40, 41]. Although these provide state-of-the-art re-identification performance on prevalent benchmark datasets, their scope is limited to

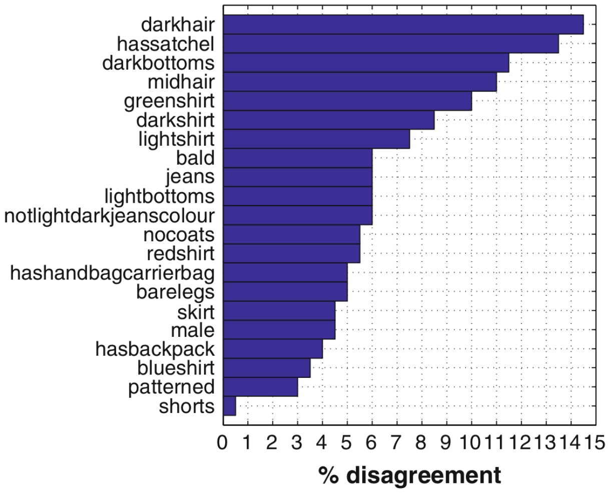


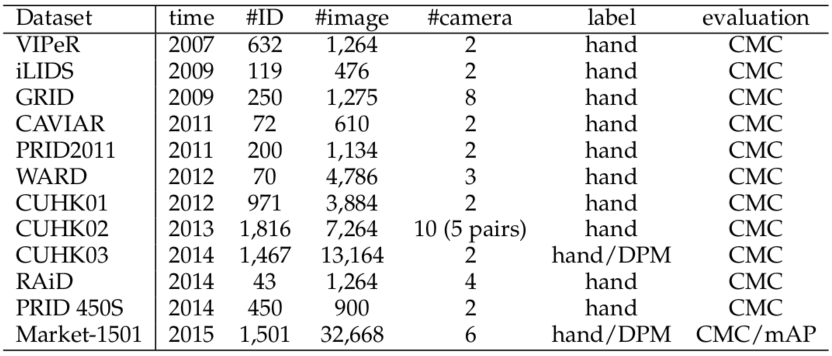
Figure 2.4: Annotation disagreement error frequencies for two annotators on PRID [37].

situations where an image of the probe subject is available from a directly corresponding camera and environment. However, surveillance operators often need to search video footage given only a human description of a suspect. This motivates our investigation into novel, stand-alone identification solutions that are non-reliant on image matching techniques or hard biometrics.

Two reviews portray the challenges [42] and milestones [43] of the area, focusing exclusively on pattern recognition, machine learning and instance retrieval approaches in the image domain. Gong et al. highlight the fundamental challenge of dealing with enormous inter- and intra-class variation, affecting observations under different camera views and conditions. The paper also defines ‘closed-world’ e.g. one-shot (VIPeR) and ‘open-world’ e.g. multi-shot and zero-shot (SoBiR, PETA) dataset evaluations [42]. Zheng et al. also chronicle the recent history and performance improvements of re-id, from hand-crafted approaches into the present age of deep learning [43]. Surprisingly, neither review dwells on attribute-based approaches. In fact, Gong et al. suggest that it is unclear whether a universal salient feature set exists to identify individuals between any camera configuration [42]. This is precisely the goal of soft biometrics, to learn salient features reported by humans and emulate their discriminative success with machines.

In 2014, Layne et al. introduce attributes-based re-identification [44], with 21 binary body descriptions on the popular VIPeR and PRID datasets, and later disclose 4.5% annotator disagreement for the ‘male’ attribute and up to 15% for ‘darkhair’ from PRID [37], Figure 2.4. The paper also proposes the challenging zero-shot identification scenario, where images of the target suspect are previously unseen by the learning algorithm.

There has since been a growing trend to solve re-identification using human describable attributes. The advent of deep learning has produced a number of CNN approaches for estimating body attributes, proving necessary in tackling such high degrees of class variation. These employ established models, such as AlexNet [45, 46], CaffeNet pre-trained



Dataset	time	#ID	#image	#camera	label	evaluation
VIPeR	2007	632	1,264	2	hand	CMC
iLIDS	2009	119	476	2	hand	CMC
GRID	2009	250	1,275	8	hand	CMC
CAVIAR	2011	72	610	2	hand	CMC
PRID2011	2011	200	1,134	2	hand	CMC
WARD	2012	70	4,786	3	hand	CMC
CUHK01	2012	971	3,884	2	hand	CMC
CUHK02	2013	1,816	7,264	10 (5 pairs)	hand	CMC
CUHK03	2014	1,467	13,164	2	hand/DPM	CMC
RAiD	2014	43	1,264	4	hand	CMC
PRID 450S	2014	450	900	2	hand	CMC
Market-1501	2015	1,501	32,668	6	hand/DPM	CMC/mAP

Figure 2.5: Popular benchmark re-id dataset statistics [43].

on ILSVRC-2012 [47] or with custom loss layers [48], R-CNN like architectures [38, 49] and particularly recently ResNet [50]. Combining several soft biometrics modalities, especially clothing, has also proven important in improving subject recognition rates [51, 52] and can be estimated for surveillance tracking and search [53, 54]. However, these works focus on the machine aspects of the problem, giving little to no consideration surrounding annotation methods, opting commonly for coarse-grained, binary or multi-class categorical attribute representations.

Pan et al. survey transfer learning in 2010, which has recently become a major topic in advancing deep learning approaches [55]. Concurrently, the bottleneck of acquiring domain specific labels has also become a focus, with two studies transferring clothing attributes from large, 'very fine-grained' fashion datasets to less constrained surveillance scenarios using an R-CNN detector [38] and MRF-IBP model [56]. These approaches re-purpose attributes from disjoint domains, focussing on domain adaptation with transductive transfer learning [55]. We are instead concerned with finding more relevant, innovative image descriptions within the surveillance domain, and perform feature-representation-transfer with unsupervised transfer learning in Chapter 6.

Consequently, attribute-based re-identification encapsulates two sets of challenges. First, the traditional re-identification 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.

2.4.4 Datasets & Attributes

Since 2014 there has been an explosion in large-scale, in-the-wild surveillance and people datasets [47, 48, 57–60], some of which contain up to 57K bounding boxes [49] and 66K people images [61]. Zheng et al. list prominent re-id datasets with up to 1816 unique subject identities in Figure 2.5 [43]. In Chapters 5 and 6, we opt for the PETA dataset

[57] which amalgamates 10 such benchmark datasets, encapsulating 19K instances and 8.9K unique subjects, making it the most diverse dataset of its kind. Similarly to Antipov et al. [46], we take an in-depth look at PETA, exploring challenging evaluation scenarios to inspect model robustness, discussed in Chapter 6.

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 disclose annotator disagreement [1, 37] (Figure 2.4) or improve recognition for difficult to classify images with special categories [30], automatic detection [35] and manual removal [46]. Smeulders et al. define this as the *sensory gap* between physical and recorded subject representations, which forms part of the semantic gap problem [9]. We therefore believe that attempting to recover biological ground-truth from such challenging images is impractical with current techniques, especially considering many annotation processes are undocumented. Instead, throughout this thesis, we explicitly aim to quantify and estimate the *crowd's perception* of such data in a more objective approach, opening up a plethora of possibilities.

2.5 Crowdsourcing

The attribute-based approaches mentioned so far all require human ground-truth labels. Although extensively investigated in the context of soft biometrics, collection methods are infrequently mentioned in re-identification studies. As such, a contemporary topic to emerge within the field of computer vision is *crowdsourcing*, which now plays an pivotal role in ascertaining large volumes of visual perception data for deep learning approaches.

Kovashka et al. survey the extent of image attributes that may be crowdsourced, techniques for their collection and strategies to select appropriate images for annotation, with the aim of encouraging annotators to provide high-quality data [62]. The study overviews a number of topics, from fine-grained and subjective attributes to concept embeddings, as well as effective user interface designs, quality assurance methods and interactive learning approaches. Very recently, O'Toole & Phillips propose five principles for crowdsourcing experiments in face recognition, arguing that the majority of annotation errors can be avoided with minor changes [63]. One principle is the awareness of the 'other-race effect', where people recognise faces from their own race more accurately than others. A second crucially notes that face recognition algorithms dependent on facial localisation will not consider body features, whereas human annotators are shown to rely on the body when the face is not visible or distinctive [64]. Chapters 3 and 5 investigate two novel crowdsourcing approaches in the quest for improved subjects descriptions, considering the topics presented in these works and the following sections.

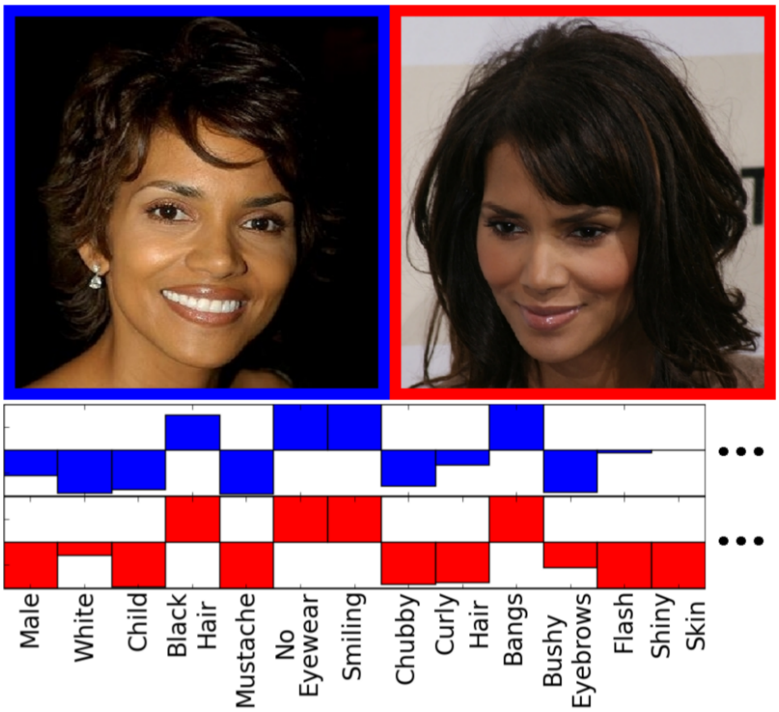


Figure 2.6: Describing visual appearance with simile attributes [69].

2.6 Fine-Grained Attributes

In the field of image classification 'fine-grained attributes' has become a catch-all term encapsulating relative attributes [65], just noticeable differences [66], large scale categorisation [58] and categorization at sub-ordinate levels [67]. Qian et al. comment that continuous and relative attributes are now widely accepted over traditional binary and multi-class annotations [68], as the field treads ever closer towards discerning the boundary between subtle and indistinguishable differences.

2.6.1 Relative Attributes

In 2009, Kumar et al. outperform traditional binary classifiers by describing face regions with pairwise 'simile attributes' to a set of reference appearances e.g. "*a mouth like Barak Obama's*" [69], Figure 2.6. In 2011, Parikh et al. introduce 'relative attributes' as a means of ranking images by attribute strength using class-level ordered pairwise comparisons e.g. "*bears are furrier than giraffes*" [65]. The paper demonstrates how unseen, zero-shot object classes can be inferred by combining relative attributes with a similarity constrained RankSVM. Extending this idea, Kovashka et al. propose iteratively whittling away irrelevant results, empowering users to better communicate their preferences via comparative descriptions to exemplar images [70].

In 2013, Reid et al. present a psychologically grounded justification for using subject-level ordered comparisons as soft biometric descriptors, labelling 19 *body* and *global* human characteristics [2]. The study demonstrates accurate retrieval of subjects using the Elo rating system and reveals that comparative labels are more objective than absolute measures which are commonly estimated unreliably in eyewitness testimonies [71–73]. Importantly, Reid demonstrates how continuous relative measurements can be inferred using a limited number of subject-to-subject comparisons.

Contemporary soft biometrics indicate 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 [2], face [25] and clothing [26]. Unfortunately, none of these studies are able to describe abstract traits like *ethnicity*, due to the restriction of single concept, one-dimensional representations. This motivates us to find a versatile solution that can represent descriptions given any manner of trait in Chapters 5 and 6.

2.6.2 Objective & Subjective Attributes

In comparison to dealing with mostly objective comparisons, Kiapour et al. collect a large-scale consensus across 5 subjective fashion categories, gamifying crowdsourcing and ranking images with the TrueSkill algorithm [74]. Importantly, Yu et al. highlight the psychophysics phenomena of 'just noticeable differences', concluding it is inappropriate to force a total order for any given image attribute [66], influencing our investigation into 'super-fine attributes' in Chapter 5. The proposed solution infers when two images are indistinguishable using a Bayesian approach to non-uniformly map low-level feature space and mid-level attribute spaces, aiming to find a perceptual consensus from a number of different respondents.

Rather than requiring majority voting and large numbers of annotations, Fu et al. propose an outlier detection approach when dealing with subjective visual properties, collected as pairwise comparisons from the crowd, stating that outliers may be caused by lazy, malicious, careless or simply incorrect workers [75]. Unlike other crowdsourcing studies, we monitor respondents throughout each task and require successful completion of an initial quiz. In fact, we find the majority of workers to be motivated, conscientious and sincere, crucially enabling us to quantify subjectivity and ambiguity, rather than discard it.

2.7 Semantic Attribute Discovery

Attribute discovery surrounds bridging the semantic gap from both ends of the spectrum. Hare et al. discuss how this is often tackled 'bottom up' from low-level feature descriptors to object labels, rather than 'top down' from the full semantic content to object labels [10], as in Figure 2.7. The work argues for hierarchical labelling ontologies and the use

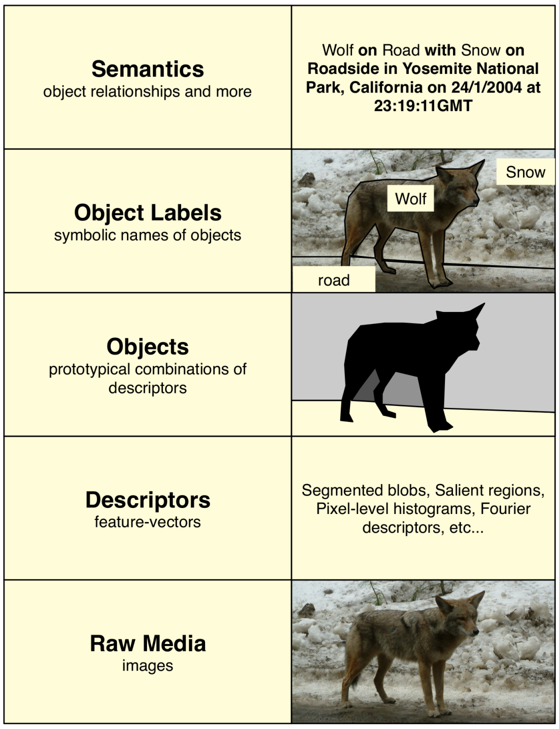

Semantics
object relationships and more
Wolf on Road with Snow on Roadside in Yosemite National Park, California on 24/1/2004 at 23:19:11GMT
Object Labels
symbolic names of objects
Snow
Wolf
road
Objects
prototypical combinations of descriptors
Descriptors
feature-vectors
Segmented blobs, Salient regions, Pixel-level histograms, Fourier descriptors, etc...
Raw Media
images

Figure 2.7: The Semantic Gap: Hierarchy of levels between the raw media and full semantics [10].

of *soft* annotations, such that similar documents share closer locations in the modelled 'semantic space', mirrored by our approach in Chapters 5 and 6.

Farhadi et al. suggest *"shifting the goal of recognition from naming to describing"* by inferring descriptive classes e.g. "*fury with four legs*" from class-level names e.g. "*dog*" to alleviate expensive manual labelling [76]. By putting semantic attribute inference at the heart of the problem, objects can be accurately described in detail, aiding the representation of new and unseen objects. Many 'bottom up' approaches find new semantic attributes by extracting co-occurrence patterns from low-level image features e.g. to enable zero-shot recognition [77]. In contrast, the following works all utilise *human perception* to discover improved semantic attributes from the 'top down'.

As such, semantic attribute discovery extends the possibilities of expertly-defined 'fine-grained' attributes, mitigating the need for domain specific engineering and linguistically constrained descriptions. The area encapsulates the machine learning topics of humans-in-the-loop [67, 78] and crowdclustering [79, 80], alongside psychology topics of perception [7, 81], similarity measures [82] and prototype theory [83].

2.7.1 Interactive Human Approaches

'Human-in-the-loop' approaches are interactive systems that incorporate human feedback in a tasks' decision making loop. Such 'active learning' algorithms alleviate laborious labelling by updating annotation tasks for respondents in real-time. Parikh et al.

employ such a system to name images and reduce the proportion of unnameable queries posed to the user [78], aiming to answer the primary question of *"which visual attributes should be learned?"*. In a similar manner, Deng et al. learn fine-grained categorisation at sub-ordinate levels, gamifying crowdsourcing [67]. The proposed BubbleGame elicits joint class and localisation annotations, guiding algorithms to inspect salient areas of the image, thereby learning *how* humans perceive differences in images.

In a similar manner, 'crowdclustering' approaches attempt to discover categories common among images. Two key studies annotate query images by asking respondents to select the most visually similar examples from a grid of images. In this way, the studies learn which salient concepts are perceived by the crowd [79, 80]. We discuss these further in Chapter 5.

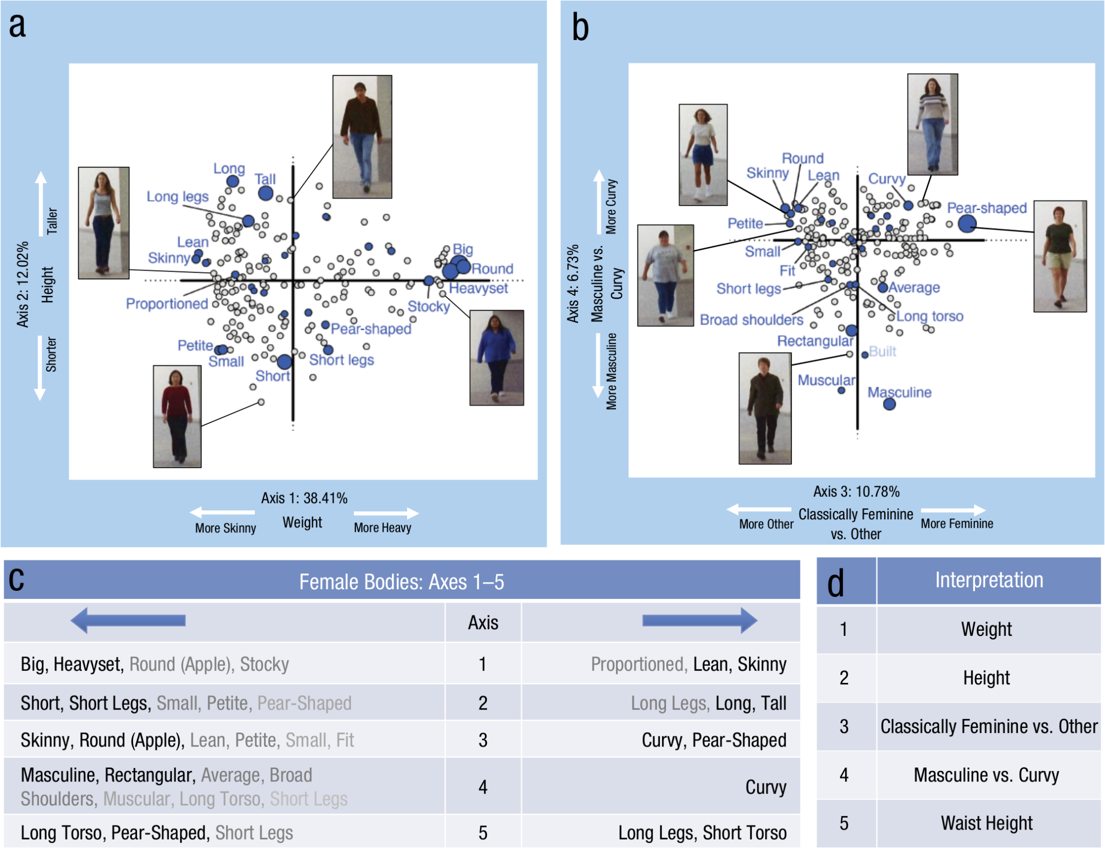
A novel approach to attribute discovery is taken by Maji, eliciting pairwise textual descriptions of image differences e.g. *"propeller plane vs. passenger plane"*, generating a corpus of parts and attributes through Natural Language Processing (NLP) techniques [84]. In this case, 'fine-grained' attributes are taken as multiple classes which describe the differences between image pairs. Importantly, the paper states that an optimal lexicon should achieve twin goals of *communication* and *discrimination* i.e. it should be easy to describe instances, as well as sufficient to distinguish instances from one another.

Crucially, all mentioned approaches evaluate their algorithms against a known 'gold-standard' ground-truth. In contrast, to evaluate 'subjective ground-truths' in the absence of a gold-standard Ellis & Whitman investigate regularisation techniques [85] and Sheng at el. show that repeat labelling is often a good approach [86]. Likewise, rather than emulate a previous method or synthetic target ground-truth, we differentiate our approach by primarily evaluating the discriminative ability of newly discovered labels.

2.7.2 Similarity in Perceptual Psychology

In 1927, Thurstone introduced the concept of a 'psychological continuum', whereby perceptual, rather than physical, discriminative measurements of object qualities are obtained from pairwise comparisons [81]. This kind of measurement is the focus of psychometrics and psychophysics, enabling the measurement of abstract human concepts.

Accordingly, the notions of *similarity* and *difference* are key in discovering new semantic concepts with which to discriminate images. Edelman & Shahbaz offer an in-depth appraisal of similarity as an explanatory concept in the context of neuroscience, demonstrating a similarity-based framework for visual object representation and dimensionality reduction [82]. Furthermore, Gärdenfors provides the basis for prototype theory, equating conceptual similarity to geometric spaces with a number of *quality* dimensions, stating that *"natural categories are convex regions"* in conceptual space [83].



Female Bodies: Axes 1–5		
←	Axis	→
Big, Heavyset, Round (Apple), Stocky	1	Proportioned, Lean, Skinny
Short, Short Legs, Small, Petite, Pear-Shaped	2	Long Legs, Long, Tall
Skinny, Round (Apple), Lean, Petite, Small, Fit	3	Curvy, Pear-Shaped
Masculine, Rectangular, Average, Broad Shoulders, Muscular, Long Torso, Short Legs	4	Curvy
Long Torso, Pear-Shaped, Short Legs	5	Long Legs, Short Torso

	Interpretation
1	Weight
2	Height
3	Classically Feminine vs. Other
4	Masculine vs. Curvy
5	Waist Height

Figure 2.8: Axes of the language space describing female body types [7].

Rice et al. investigate if humans use body features in identifying pairs of images at-a-distance, classifying them as 'same' or 'different' on a 5-point similarity Likert scale [64]. When purposefully selecting person images with indiscriminate faces, facial identification performs at near chance level, but importantly, whole-body identification is still accurate. In this scenario, respondents' eye movements are found to shift towards the body, indicating that humans subconsciously use body cues for visual person identification when other sources of information are not available.

Recently, a novel psychology study investigates the similarity of body descriptions, closely related to our work. Hill et al. explore the relationship between physical body shapes and their linguistic descriptions, to generate realistic 3D avatars [7], Figure 2.8. The study discovers multi-dimensional linguistic 'similarity spaces' from perceptually salient global and local body features. The dimensions of body-shape variation are related to major axis of the language-based similarity space, correlating human descriptions with laser body scans.

In computer vision, images are often compared directly to one another, measuring their similarity as part of distance metric learning techniques. Deselaers & Ferrari investigate semantic and visual similarity within the ImageNet dataset [87], determining visual prototypes for every category and confirming that visual classes are separable across semantic boundaries [88]. Lastly, Scheirer measures human vision to improve computer vision, creating an improved 'perceptual annotation' face detector over traditional Viola-Jones approach, employing an online psychometric testing platform to measure exemplar difficulty [89]. The paper mentions the *"obvious gap between current state-of-the-art computer vision applications and human performance"* commenting on the fact that *"humans have access to more extensive training data through a lifetime of unbiased experience with the visual world"*. Although some computer vision approaches are stated as obtaining 'super-human' performance, such successes are still restricted to highly constrained application scenarios, while a vast range of human cognitive processes remain to be tackled automatically. In contrast to traditional facial recognition, pedestrian identification is far less constrained and therefore still proves highly challenging.

2.8 Conclusions

In this chapter we present a wide range of literature surrounding soft biometrics and human- and machine-centric approaches. We highlight the transition from face to body, from biological to perceived ground-truths and from categorical to relative attributes, in the quest to narrow the semantic gap and facilitate more ubiquitous identification techniques. In essence, the question is not *if*, but *how* soft biometric labels can be discerned and utilised for subject identification.

The background in soft biometrics, attribute-based re-identification and relative attributes influences the design of our preliminary annotation task and dataset in the Chapters 3 and 4, exploring the discriminative characteristics of comparative labelling. Leading on from this, the review of semantic attribute discovery discusses the integration of abstract conepts from human knowledge in machine learning algorithms. Subsequently, Chapters 5 and 6 tackle pertinent questions regarding the applicability of contemporary approaches on real-world datasets, by attempting to emulate the crowd's perception, largely inspired by works in perceptual psychology, discussed further in situ.

Chapter 3

Crowdsourcing Ordered Comparisons

3.1 Introduction

This chapter starts our search for ubiquitous identification techniques by introducing relative soft biometrics. We introduce a novel methodology for crowdsourcing *ordered comparative* annotations of *global* and *body* soft biometric traits and investigate their potential for human description and retrieval. Crowdsourcing enables data collection from a diverse, global pool of annotators, simulating the typical response variation that can be expected in a surveillance scenario. This work follows previous studies in soft biometrics and relative attributes, discussed in Sections 2.3 and 2.6.

Section 3.2 presents an image set capturing 100 subject images in a synthetic environment, enabling the coherent appraisal of crowdsourced relative annotations. Section 3.3 introduces the crowdsourcing annotation task and Section 3.4 evaluates the crowdsourcing suitability of selected soft biometric traits, analysing their distributions and discerning the degree of uncertainty for each trait. Section 3.5 explains the RankSVM algorithm to infer precise relative subject signatures from our new comparative annotations. Section 3.6 then analyses the inferred relative measurements for stability, confusion and discrepancy when viewed in an exemplary scenario. To further ascertain their correspondence to traditional methods, comparative annotation characteristics are interpreted against categorical annotations collected by [1] in Section 3.7. Lastly, a semantic retrieval experiment evaluates the discriminative properties of the inferred relative labels, contrasted to previous categorical [1] and comparative [2] labelling techniques in Section 3.8. Our findings are summarised in Section 3.9.

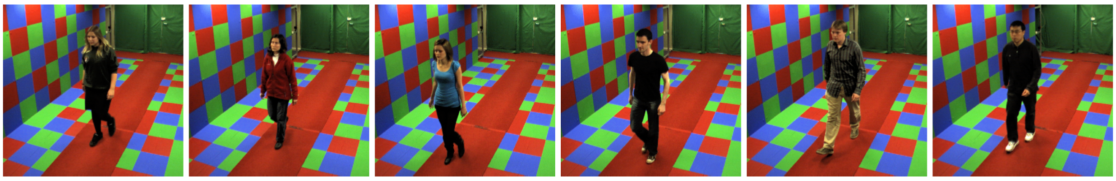


Figure 3.1: Example subject images for comparative annotation.

3.2 Multi-Biometric Tunnel (MBT) Dataset

A multitude of large-scale pedestrian surveillance and re-identification datasets exist that are captured 'in-the-wild' as discussed in Section 2.4. However, at this preliminary stage, we are interested in investigating the characteristics of crowdsourced ordered comparisons in a controlled environment, to mitigate issues pertaining to severe occlusion and large variations in pose and illumination.

For this reason we construct a more constrained dataset, reducing the complexities associated with large-scale datasets and challenging images, as in [1, 2, 19, 22]. We opt for the University of Southampton Multi-Biometric Tunnel (MBT) dataset [3]. Originally intended for gait analysis, MBT captures subjects walking through a purpose built biometric tunnel across 12 camera viewpoints simultaneously. The dataset enables direct comparison of our approach to two previous studies using similar synthetic image sets [1, 2]. The range of captured viewpoints also emulate realistic CCTV camera placements, facilitating our viewpoint invariant experiments in Chapter 4.

From MBT we extract a *gender balanced* subset of 100 randomly selected subject images. Images from a single forward-facing camera are aligned to a similar position along the tunnel and cropped to equal size, as seen in Figure 3.1. In Chapters 5 and 6 we explore annotating PETA, a very large-scale, real-world image set containing approximately 90 times more subjects. There we tackle a number of additional challenges, such as; how to efficiently collect very large-scale comparative annotations, how to interpret severely obscured low quality images and whether to label subjects uniformly or as individual image instances.

3.3 Crowdsourcing Task

In this section we detail the design considerations in constructing a crowdsourcing task to facilitate the collection of a large number of high quality comparative annotations. We employ the CrowdFlower[1] platform to build and run the crowdsourced annotation task. The platform provides comprehensive data analysis and quality control tools, allowing end-users to accept a range of responses while rejecting non-genuine answers.

[1] http://www.crowdflower.com/

	Response labels (5-point bi-polar scale and "Can't see")					
Soft traits	5	4	3	2	1	0
Gender	Much more Feminine	More Feminine	Same	More Masculine	Much more Masculine	Can't see
Age	Much more Old	More Old	Same	More Young	Much more Young	Can't see
Height	Much more Tall	More Tall	Same	More Short	Much more Short	Can't see
Weight	Much more Heavy	More Heavy	Same	More Light	Much more Light	Can't see
Figure	Much more Fat	More Fat	Same	More Thin	Much more Thin	Can't see
Chest size	Much more Big	More Big	Same	More Small	Much more Small	Can't see
Arm thickness	Much more Thick	More Thick	Same	More Thin	Much more Thin	Can't see
Leg thickness	Much more Thick	More Thick	Same	More Thin	Much more Thin	Can't see
Skin colour	Much more Dark	More Dark	Same	More Light	Much more Light	Can't see
Hair colour	Much more Dark	More Dark	Same	More Light	Much more Light	Can't see
Hair length	Much more Long	More Long	Same	More Short	Much more Short	Can't see
Muscle build	Much more Muscle	More Muscle	Same	More Lean	Much more Lean	Can't see

Table 3.1: Lexicon of soft traits and possible response labels.

As it connects to global pools of contributors, unambiguous and decisive questions must be presented.

We aim to improve upon the crowdsourcing work of Han et al. [35], who spent a significant sum of money collecting a large number of human intelligence tasks (HITs), gaining only a few valid responses. Additionally, the goal is to collect geographically unconstrained data to better model the average human perception and description of others, compared to more restrictive annotation tasks as in [2]. Although individual annotator perceptions are undoubtedly affected by their own personal traits, we choose to consider the average human perception. This is because the characteristics of users interacting with a subject identification system may not be previously known.

3.3.1 Trait & Label Derivation

In 2015, Dantcheva et al. defined soft biometrics as *"physical, behavioural or adhered human characteristics, classifiable in predefined human compliant categories"* [5]. However, human descriptions of visual attributes are inherently unreliable, especially when dealing with absolute or continuous demographic value estimations from a diverse group of people such as with crowdsourcing [35]. Several studies have concluded that estimates of absolute body measurements are often very inaccurate [73], subject to anchoring and cross-race effects [63, 72] and avoid extreme judgements [71]. In order to avoid such issues, we follow Reid in comparatively labelling attributes, mitigating the influence of a respondents' self perception on their annotations [2]. In this section we explain our labelling method and lexicon selection, resulting in more objective annotation questions and avoiding the limitations of eliciting absolute labels.

In 1994, MacLeod et al. set out the first system to record body attributes, founded on psychological observations of perception and memory [90]. The study deduces the 13 most reliably interpreted body 'scales', measured on a 5-point scale. These 13 body traits have in turn been assessed for categorical annotation variance [1], stability and discriminative power at-a-distance [19] and imputation accuracy [2]. By collectively reviewing the most significant, prevalent and stable traits from [1, 2, 19, 90], the final soft biometric trait lexicon of 2 global and 10 body soft traits is deduced, seen in Table 3.1.

Trait and label nomenclature is also simplified, to preserve the question and response objectivity with a global crowdsourcing audience.

Lucas et al. argue against separating by *ethnicity*, stating that it is often misinterpreted when describing low quality images [12]. Although known to be one of the most distinctive traits in policing [27, 28], there is no apparent method to represent multiple, abstract ethnicities through a single set of binary polar labels. As such, *ethnicity* is excluded from our list of global soft traits. In Chapter 6 we successfully tackle the issue of objectively comparing and categorising *ethnicity*, demonstrating its priority for automatic suspect identification.

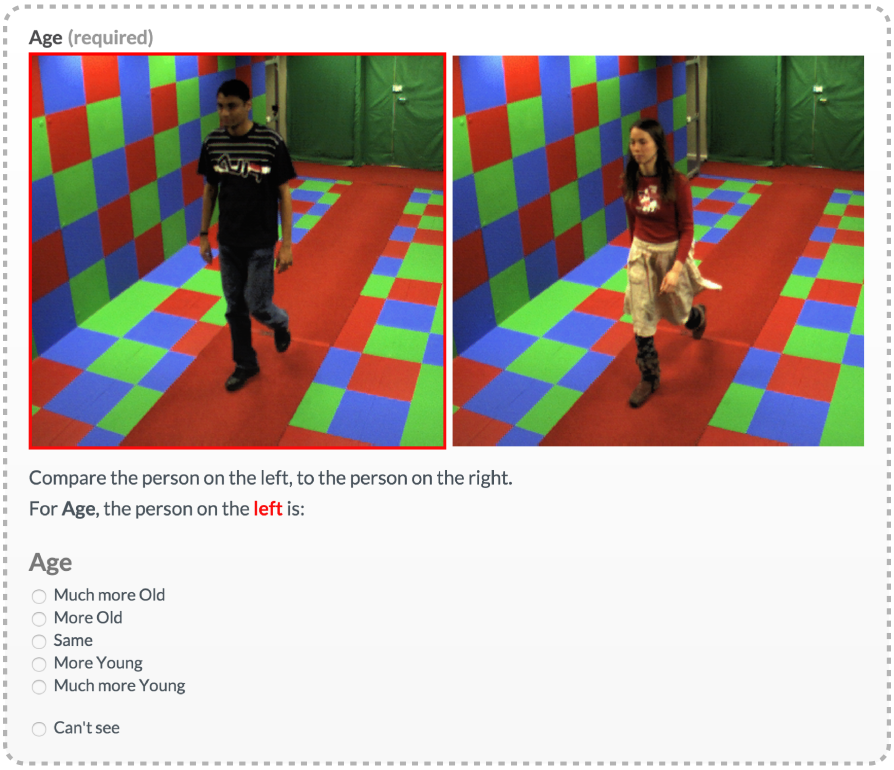
Finally, it is important to note that *gender* and *skin colour* are also collected as comparative traits in this study. Previously, Reid et al. was influenced by police witness evidence forms, opting to instead describe these traits through traditional binary and multi-class labels [2]. However, this can lead to near homogeneous labelling if the dataset is unbalanced, or annotation suffers from the 'cross race effect' [4]. Comparative annotation aims to mitigate these effects, by objectively describing differences between pairs of subjects, rather than subjectively categorising their similarities. As far as we know, this is the first time *gender* has been measured in this way and on such a scale, being most commonly described in a binary fashion. This also follows the trend of global awareness regarding the LBGT community and alternative gender classifications.

3.3.2 Question & Response Design

Each annotation question is essentially a psychometric procedure, whereby the respondent is shown two stimuli images and asked to "compare the person on the left, to the person on the right", for the 12 traits defined in our new lexicon. In total $12 \times \binom{100}{2}$ unique annotations were asked, comparing each subject to every other subject for each trait. A 5-point bi-polar Likert-type scale was used for all annotations as in [1, 2, 90], following a consistent format: "Much more `A`", "More `A`", "Same", "More `B`", "Much more `B`", Table 3.1. The 5-point Likert-type scale is commonly used in psychometric studies and was chosen to balance response granularity and annotation speed.

Reid et al. collects an optional 'certainty' rating for each annotation [2]. Empirically we find crowdsourcing respondents often ignored this time consuming option with no extra cash incentive. Instead, we provide an additional "Can't see" option as an acceptable response for hard to distinguish questions. This choice is imperative, as it reduces the chance of collecting feigned and inaccurate responses for very low confidence answers.

Respondents are permitted to answer up to 20 pages, each containing 10 image annotation tasks. The crowdsourcing platform enables the creation of pre-defined 'gold-standard' test questions, in order to measure respondents' accuracy and minimise the number of spurious responses. The first page consists entirely of test questions, which



Age (required)

Compare the person on the left, to the person on the right.

For **Age**, the person on the **left** is:

Age

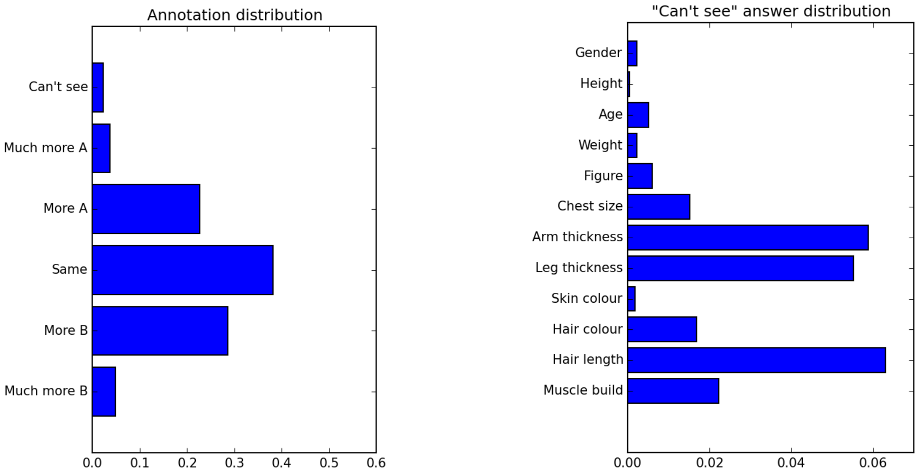
○ Much more Old
○ More Old
○ Same
○ More Young
○ Much more Young

○ Can't see

Figure 3.2: Screenshot of one annotation task question.

must be passed in order to proceed and be paid. Subsequent pages contain 1 covert test question and 9 genuine annotations, therefore continuously monitoring respondents' reliability. Several subsets of questions were trialled, to measure the acceptability of the predefined test questions. As a result, the following considerations were made:

- To make test questions fair, they are sampled from more obvious comparisons and only the most fundamentally incorrect responses are rejected. However, respondents must score at least 80% test question accuracy to proceed.
- "Can't see" was marked as an acceptable response for all annotations, but secretly capped at a maximum response rate of 20% per respondent. The very few respondents straying outside this limit were excluded from the process.
- Respondents were also rejected if their response distribution varied largely from the average response distribution formed during the initial trials.
- In addition to a large number of introductory examples, each question included text and highlighting, reiterating the task question to "compare the person on the left, to the person on the right".
- The response form was formatted using vertically aligned radio buttons, enabling quick and instinctive responses to incentivise respondents further. Initial answers were left blank to avoid anchoring [2]. Figure 3.2 illustrates final question layout and accompanying text.



(a) Overall annotation distribution.

(b) "Can't see" answer distribution.

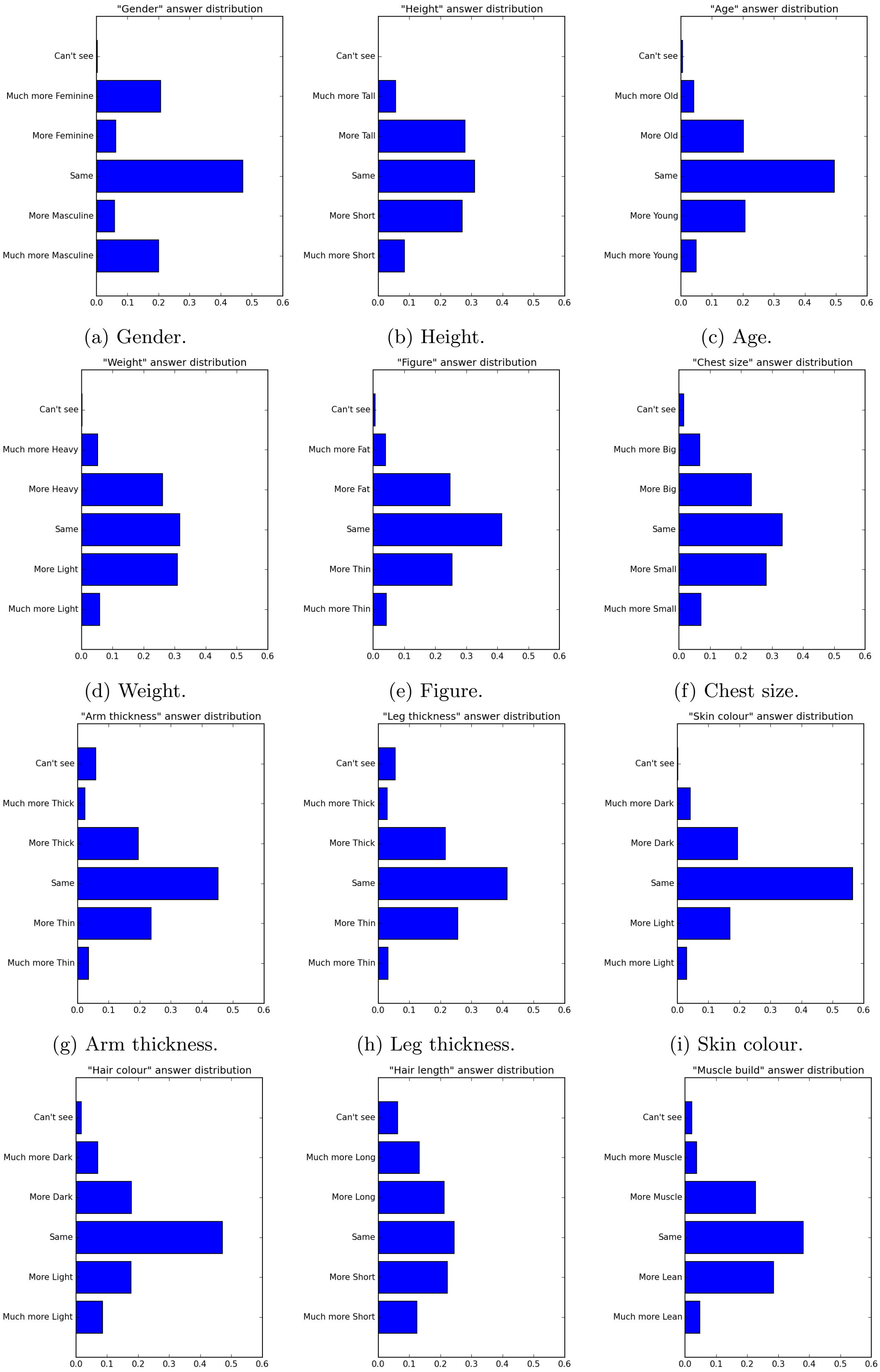
Figure 3.3: Collated answer distributions.

3.4 Annotation Analysis

The annotation task collected 59400 unique annotations collected from 892 trusted respondents (124 untrusted respondents were flagged, and 4383 responses rejected). Including trail runs, the final task cost only $303. Clear instructional text and objective test questions meant our task was more economic compared to Han et al.'s study, that spent $3000 on 112,519 HITs [35]. Furthermore, 179 trusted respondents rated our task favourably, giving it an average of 4.4 out of 5.

Figure 3.3a details the overall annotation distribution for the task. Although "Can't see" is always an acceptable response, only 2.4% of answers are marked as such. Figure 3.3b compares the distribution of "Can't see" responses, forming a measure of uncertainty for each trait. As expected *arm thickness* and *leg thickness* are very uncertain, being the least distinctive traits chosen from previous work [1, 2]. Interestingly, *hair length* is the most uncertain, due to one subject wearing a head scarf, and many others with long hair obscured by their body and camera angle. This suggests a future task to collect annotations from alternative viewpoints, where the visibility and interpretation of physical features may vary, as discussed in Chapter 6.

Figure 3.4 encapsulates the annotation distributions, observing how polarised or indecisive respondents were when comparing subjects for each trait. For instance, *gender* appears to be exceptionally decisive, with a high number of responses for "Much more Feminine/Masculine" and near 50% "Same". Meanwhile, annotators shied away from answering "Much more Fat/Thin" for *figure*, similarly to *arm thickness*, *leg thickness* and *muscle build*. Other distributions exhibit a high response rate for "Same", reflecting distributions within the dataset, as many subjects are of similar *age* and *skin colour*.



(a) Gender. (b) Height. (c) Age.

(d) Weight. (e) Figure. (f) Chest size.

(g) Arm thickness. (h) Leg thickness. (i) Skin colour.

(j) Hair colour. (k) Hair length. (l) Muscle build.

Figure 3.4: Answer distributions per trait.

This informs us that differing trait distribution characteristics are not due to respondent confusion and that annotation uncertainty is trait specific and independent. Furthermore, *gender* and *skin colour* are two of the least uncertain traits, demonstrating their suitability for comparative collection, contradicting the assumptions of MacLeod and Reid [2, 90].

3.5 Inferring Semantic Ranks

To interpret the annotated pairwise comparisons, we wish to infer the semantic *score* for each soft biometric trait associated with every subject. Relative scores enable us to rank subjects, forming an ordered list for each trait. In this section we introduce the soft-margin RankSVM formulation to infer relative attributes, by applying one model to each set of ordered pairwise trait comparisons independently.

The RankSVM algorithm has been successfully applied to web page ranking [91], ordering images from low-level features [65] and also producing continuous ground-truth measures of texture [92] and clothing appearance [52] from noisy pairwise comparisons.

Given a two sets of pairwise constraints O_q and S_q for each trait $q \in 1, ..., Q$. Each image instance $i, j \in 1, ..., N$ is represented as a unique and mutually exclusive binary feature vector $I = [\boldsymbol{x_i}] \in \{0, 1\}^N$, forming an $N \times N$ dimensional identity matrix.

The set of ordered pairs is defined as $(i, j) \in O_q \implies i \succ j$, i.e. instance i is described to be more like trait q's polarity `A` than instance j. To reduce the effects of annotation discrepancy, "Much more" and "More" responses are represented equally in each set O_q. The set of unordered pairs is defined as $(i, j) \in S_q \implies i \sim j$, i.e. instance i and j possess similar qualities for trait q, constituting all "Same" responses. "Can't See" responses are ignored, such that they do not impose pairwise constraints. Our goal is to derive Q trait target ranking vectors $\boldsymbol{w_q}$, such that the maximum number of the following constraints are satisfied:

$$\begin{aligned} \text{for ordered images: } & \forall(i,j) \in O_q,\ \boldsymbol{w_q^T x_i} > \boldsymbol{w_q^T x_j} \\ \text{for similar images: } & \forall(i,j) \in S_q,\ \boldsymbol{w_q^T x_i} = \boldsymbol{w_q^T x_j} \end{aligned} \tag{3.1}$$

To approximate the solution, we employ Joachims' popular soft-margin RankSVM [91] and Parikh's later extension to support similarity constraints [65] as follows:

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \frac{1}{2}||\boldsymbol{w_q^T}||_2{}^2 + C\sum \xi_{ij}^2 \\ \text{subject to} \quad & \boldsymbol{w_q^T}(\boldsymbol{x_i} - \boldsymbol{x_j}) \geq 1 - \xi_{ij},\ \forall(i,j) \in O_q, \\ & |\boldsymbol{w_q^T}(\boldsymbol{x_i} - \boldsymbol{x_j})| \leq \xi_{ij},\ \forall(i,j) \in S_q, \\ & \xi_{ij} \geq 0, \end{aligned} \tag{3.2}$$

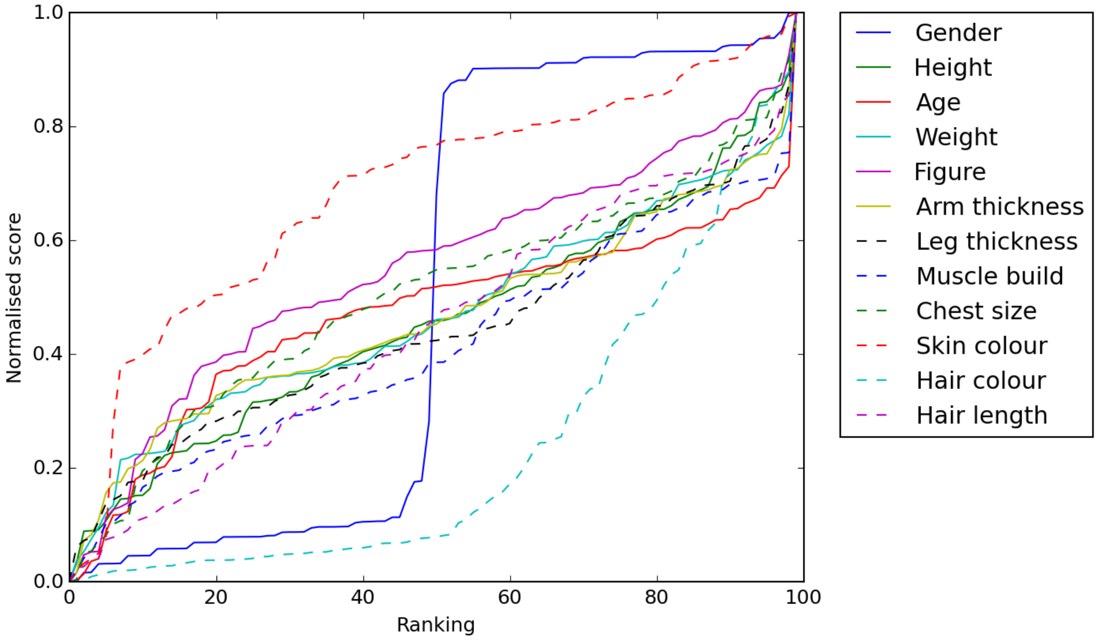


Figure 3.5: Normalised relative scores against ranks for each trait.

where ξ_{ij} is the slack variable representing the misclassification ranking error between instances i and j and C is trade-off constant between maximising the margin and satisfying the pairwise constraints [65]. As this is an SVM formulation, it is also possible to be extended to learn rankings from any given feature space, e.g. automatically generated image features.

3.6 Inferred Score & Rank Analysis

3.6.1 Trait Characteristics

By applying the RankSVM algorithm to each trait set independently, we find $q \in Q$ separate target ranking vectors $\boldsymbol{w_q}$. Each element of $\boldsymbol{w_q}$ represents a relative continuous score describing the appearance of a single subjects' trait within an image. These scores are then also used to rank subjects. Figure 3.5 contrasts the normalised relative scores against their ordered rankings, illustrating the different characteristics of each soft trait's inferred values (using $C = 1$). *Gender* comparisons produce a highly binary distribution between 'feminine' and 'masculine' polarities. However, the gender response is not a perfect step function, and there are several subjects whose gender is not as pronounced as others, visualised between ranks 48-52. *Hair colour* also exhibits a number of similar low scores, representing subjects with dark shades of hair (ranks 0-50), whereas lighter shades are more easily distinguished (ranks 50-100). In contrast, traits like *height* display an almost linear correlation between score and rank, as even subtle differences are clearly observed.

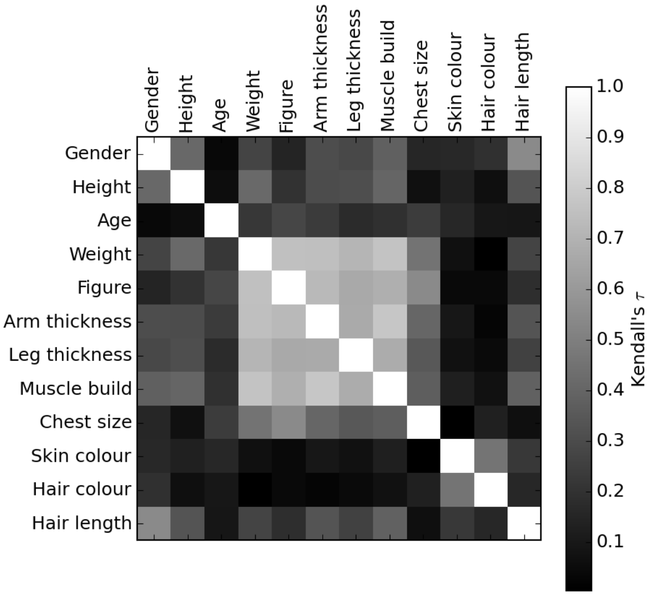


Figure 3.6: Kendall's tau correlation between trait ranks.

Kendall's τ coefficient is used to measure rank correlations between traits in Figure 3.6. Similarly to [13], there is a correlation cluster between build characteristics e.g. *weight*, *figure*, *arm* and *leg thickness* and *muscle build*. A strong correlation pair was found between *skin colour* and *hair colour*, as darker skinned subjects tend to have darker hair. *Gender*, *height* and *hair length* also possess strong correlations, while *age* varies most independently.

3.6.2 Trait Stability

In order to discern the stability and consistency of the inferred measurements, we vary the number of comparisons, $n \in 1, ..., \frac{N}{2}$, used in the ranking inference process. Annotations are sampled at random, maintaining n comparisons per subject trait, reported as an average of 50 iterations per n. This provides insight into the underlying subjectivity, confusion and discrepancy for each soft trait.

Figure 3.7 depicts a measure of inferred rank stability for each trait, applying Kendall's τ across all iterations of n. We find *gender* to be the most stable at lower n values, while *chest size* and *skin colour* are the least stable overall. Only slight improvements to stability coefficients are provided by $n > 3$ comparisons, suggesting suitable orders are found by $n = 4$.

Next we inspect the standard deviation of both ranks and scores at each n, Figure 3.8. We find *gender* produces the most consistent soft trait scores at lower n, closely followed by *height*, Figure 3.8a. However, the same is not true for rank, where *gender* is the least consistent, but *height* remains the most stable measure, Figure 3.8b.

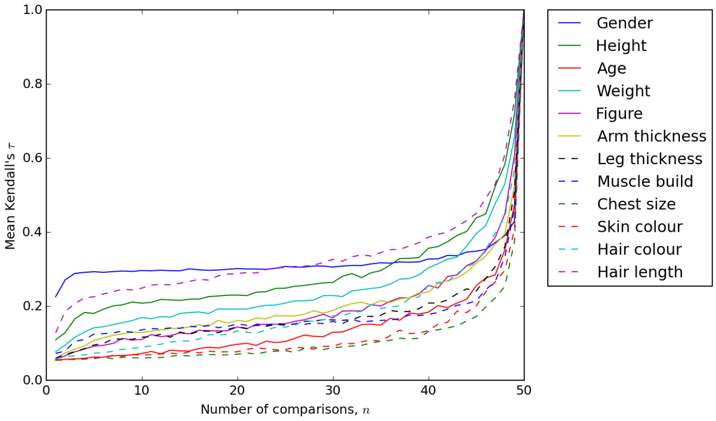
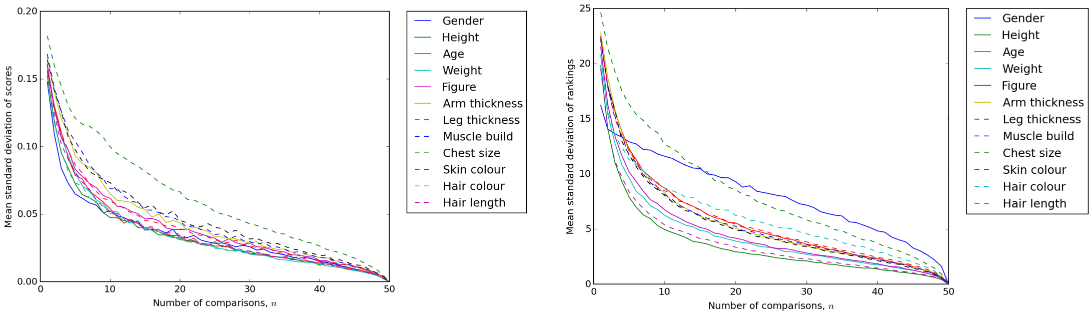


Figure 3.7: Kendall's *tau* correlation of traits ranks, varying number of comparisons, n.



(a) Scores.

(b) Rankings.

Figure 3.8: Mean standard deviation of traits, varying number of comparisons, n.

Although this appears to directly contradict the findings in Figure 3.7, rank variance is not necessarily related to overall rank discordance. For example, the binary separation of *gender* in Figure 3.5 is maintained at low n, due to similar relative scores, resulting in overall rank concordance, portrayed in Figure 3.7 and 3.8a. However, within each binary region, subject ranks vary widely, due to lack of distinction within the perceived 'feminine' and 'masculine' groups, increasing the standard deviation reported in Figure 3.8b. Traits like *hair colour* also follow a similar pattern.

From this analysis we can discern that highly distinctive and separable traits like *gender* and *hair colour* produce stronger overall rank correlations and more consistent inferred scores, at the expense of confusing subject ranks with similar trait qualities. Our findings also indicate that *height*, *gender*, *hair length* and *weight* are overall more salient subject descriptions, while *chest size* is the least clearly discerned.

3.7 Relative & Absolute Measurement Correspondence

In this section we compare the correspondence of our comparative annotations to Samangooei's categorical Multi-Biometric Tunnel (MBT) dataset annotations [1]. We do this in two ways, comparing both the equivalent subject labels and inferred rankings.

3.7.1 Annotation Interpretation Methodology

We start by defining the methodology for interpreting the two sets of annotations and then define two measures of correspondence between labels and ranks.

Comparative annotations for each subject pair are mapped to relative integers, $\boldsymbol{R} = [R_{ij}] \in \{5, 4, 3, 2, 1\}$. "Can't see" responses are ignored for these measurements. Categorical annotations are quantised to absolute integers for each label class, $\boldsymbol{A} = [A_i] \in \mathbb{Z}$, with corresponding semantic orders, e.g. "Female" to 1 and "Male" to 2 for *gender*, "Very Short" to 1 and "Very Tall" to 5 for *height* etc.

Label correspondence measures how similar the two sets of annotations are when comparing the semantics of each pair of labels, counted per subject pair as follows:

$$\text{label correspondence}(i, j) = \begin{cases} 1 & R_{ij} < 3 \text{ and } A_i < A_j \\ 1 & R_{ij} > 3 \text{ and } A_i > A_j \\ 1 & R_{ij} = 3 \text{ and } A_i = A_j \\ 0 & \mathit{otherwise} \end{cases} \tag{3.3}$$

The overall label correspondence is averaged over all n pairwise annotations per trait:

$$\text{overall label correspondence} = \frac{1}{n} \sum_{i \in N} \sum_{j \in N \wedge j \neq i} \text{label correspondence}(i, j) \tag{3.4}$$

Next we compare the inferred comparative ranks, $\boldsymbol{w}$, to derived categorical ranks, $\boldsymbol{a}$. Subjects are ordered based on the absolute value assigned to each trait category in $\boldsymbol{A}$. A subjects' position in this order defines their categorical rank, a_i. Rank correspondence for each subject $i \in 1, ..., N$ is therefore expressed as follows:

$$\text{rank correspondence}(i) = \begin{cases} 1 & w_i < a_i \\ 0 & \mathit{otherwise} \end{cases} \tag{3.5}$$

Finally, the overall rank correspondence is averaged over all N subjects per trait:

$$\text{overall rank correspondence} = \frac{1}{N} \sum_{i \in N} \text{rank correspondence}(i) \tag{3.6}$$

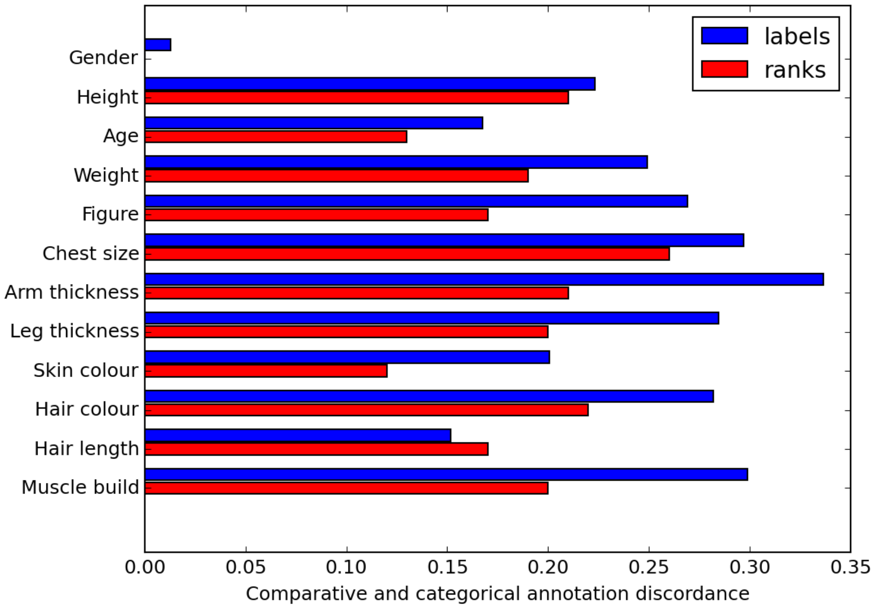


Figure 3.9: Overall label and rank discordance.

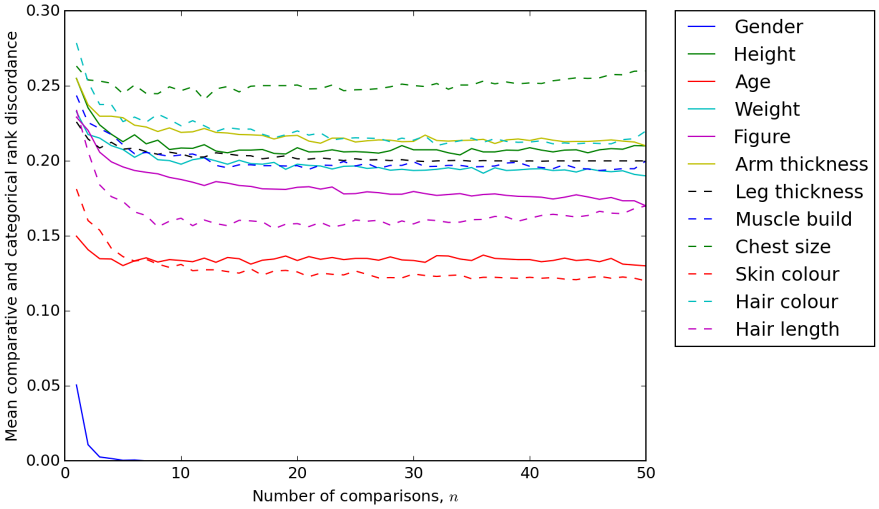


Figure 3.10: Mean rank discordance, while varying number of comparisons n.

3.7.2 Label & Rank Correspondence Analysis

From the final inferred relative ranks (where $n = N$), we display the fraction of discordance between comparative and categorical labels and ranks in Figure 3.9. With the exception of *gender*, traits vary between 12% and 34%. *Gender*, being extremely easy to discern and categorise shows perfect correspondence for rank. The discordance of *gender* labels is much lower than expected, suggesting annotators avoided describing two subjects of the same gender as only ‘more feminine / masculine’ even though this would be a valid response. Perhaps surprisingly, labels are overall more discordant than ranks, indicating that comparative annotations contradict pairs of categorical annotations more than the final ranks, possibly as they offer more precision in description.

We also inspect the correspondence of comparative and categorical ranks while varying n, Figure 3.10. Again, *gender* requires only 3 comparisons per subject to rank them in perfect correspondence to the level of detail offered by the inferred categorical ranks. For the remaining traits, rank discordance remains consistent when $n > 5$. This includes the least uncertain traits (Section 3.4) and most stable and salient traits (Section 3.6.2). Therefore, we can imply that differences in final rankings are due to comparative annotations containing additional discriminative information over categorical annotations.

3.8 Soft Biometric Retrieval

This section demonstrates how it is possible to perform soft biometric retrieval using pre-interpreted relative scores. Biometric retrieval is the process of identifying an unknown observation (the probe or suspect), by matching it to a set of known subjects (the gallery). This process is the intersection of human perception, soft biometric annotation and machine learning applied to labelling datasets, as seen in Figure 1.1. Performing Content-Based Image Retrieval (CBIR) in this way is suited to forensic investigation where suspects must be identified from a gallery of people captured across a video surveillance network given only an eyewitness description.

3.8.1 Methodology

We aim to recognise a previously unknown suspect description from a gallery of the 100 known subjects. By varying the number of comparisons supplied to generate the suspect's signature, we simulate an eye witness testimony that compares the suspect to n known subjects. Our retrieval methodology is inspired by [2].

The experiment chooses probe subject $i \in 1, ..., N$ from the annotated dataset and excludes $n \in 1, ..., \frac{N}{2}$ sets of randomly sampled comparisons between the probe and n other subjects. The excluded comparisons are used to form a new suspect query, inserted into the dataset. Biometric signatures are generated for each gallery subject and suspect, represented as a vector of Q target values for subject i, such that $\boldsymbol{x_i} = \{w_q^i\}, \forall q \in Q$, using the RankSVM technique described in Section 3.5.

To perform retrieval, a Euclidean distance Nearest Neighbour operator is applied between the probe signature and the gallery subject signatures as in [1, 2]. The outcome is classed as successful if the closest match to the suspect is the original probe subject (rank-1 retrieval accuracy).

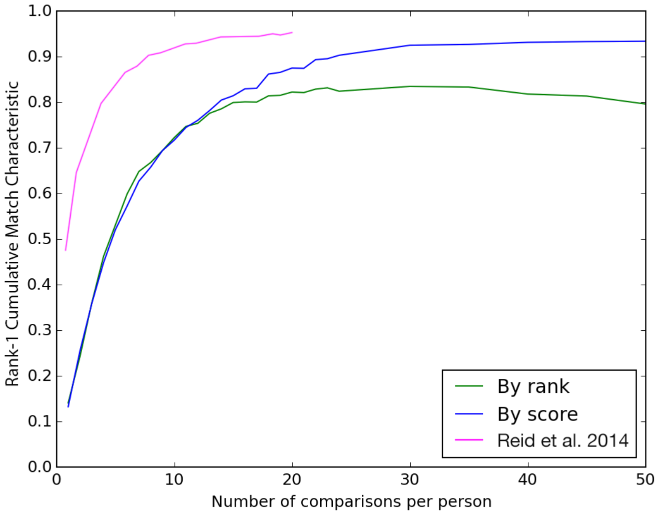


Figure 3.11: Rank-1 retrieval accuracy, for $n \in [1, 50]$, compared to [2].

3.8.2 Performance Analysis

For each subject and set of n comparisons, 50 iterations are run. Results are recorded using signatures built from both relative normalised scores and ranking positions of each trait.

A direct comparison to [2] is made, who performed retrieval with annotations gathered from a smaller image dataset of 80 subjects, using 7 additional traits (4 comparative, 3 categorical). Therefore, the annotation workload for 12 sets of comparisons is equivalent to 19 sets from our lexicon. The study collected 558 annotations from 57 annotators, with the remaining comparisons synthetically inferred [2]. We also compare our results to Samangeooei et al.'s original categorical labels of 23 soft traits, collected on the multi-biometric tunnel dataset [1].

Our goal is to emulate a realistic response environment by using crowdsourced data. Therefore, we treat all annotations as equal, including "Can't see" responses, and we do not synthetically infer labels. Drawing annotations from a deeper and wider pool of online annotators than Reid and avoiding label inference mitigates the effects of inflated self-correlation within our data. For this reason, we find lower retrieval accuracies, compared to [2], at lower n values when performing rank-1 retrieval, Figure 3.11. Rank-1 retrieval rates also climb more slowly when increasing n, suggesting our data is slightly more inconsistent when labelling extra subjects. Even so, score-based retrieval still attains a maximum match rate of 93% at rank-1, compared to Reid et al.'s 95% [2].

Increasing the number of comparisons n to form the probe query means that relatively fewer $N - n$ comparisons remain for the subject's gallery signature. A side effect of this is a decrease in rank-based rank-1 retrieval rates at $n > 25$. However, score-based signatures remain consistent, even when generated from significantly different

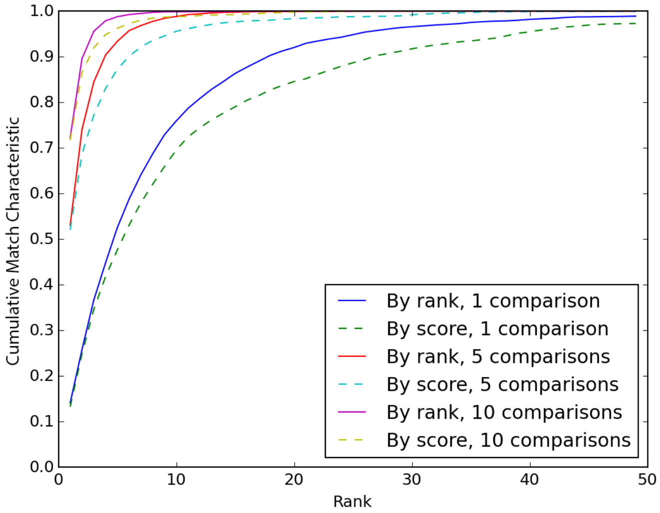


Figure 3.12: CMC for $n \in (1, 5, 10)$.

sets of comparisons. This can be attributed to traits like *gender* and *hair colour*, which have regions of similar relative scores (Figure 3.5) that are not adversely affected when retrieved by a score-based method. Therefore, relative scores describe the possessed quality of a trait better than relative ranks, which tend to diverge between gallery and probe queries as n increases.

A second experiment assesses retrieval accuracy while varying the acceptance rank, reported as the the Cumulative Match Characteristic (CMC). This reproduces a surveillance scenario in which the operator can rapidly eliminate irrelevant subjects, leaving only the most likely correct matches to manual intervention.

With only $n = 1$ comparison the system obtains 75% accuracy at rank 10, while with $n = 10$ comparisons it achieves 100% retrieval accuracy at rank 7. In these cases, rank based signatures outperform score based signatures, as increasing the acceptance rank improves cases where correct matches have small rank errors but proportionally larger score errors. These promising results show that with only $n = 5$ sets of comparisons, a surveillance operator would be guaranteed to find the correct identity in the top 13% of results.

Figure 3.13 illustrates our approach against [2] at $n = 10$ and the original absolute labelling scheme of [1]. At $n = 10$ our approach's retrieval rates clearly surpass [1] and actually converge to 100% earlier than [2], even when using far fewer soft traits and including "Can't see" and discrepant annotations collected via crowdsourcing.

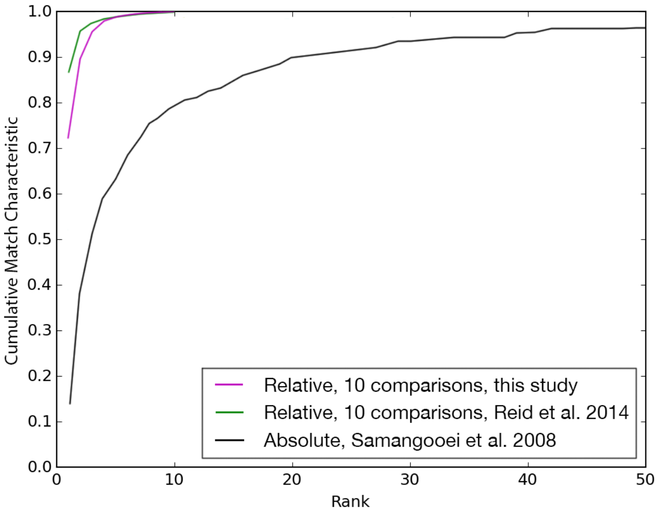


Figure 3.13: CMC for $n = 10$, compared to [2] and absolute labels from [1].

3.9 Conclusions

In this chapter we demonstrate that crowdsourcing comparative soft biometrics provides enhanced precision and retrieval performance over categorical alternatives, when describing individuals from images. By applying a RankSVM algorithm to interpret human comparisons, we build precise, relative soft biometric signatures from objectively crowdsourced comparisons.

With this technique and a lean lexicon of soft traits, our experiments perform retrieval almost as well as, and in some cases better than a previous relative attributes study which uses more traits and fewer subjects [2]. We also establish which soft traits are most suited to comparative description, finding *height* to be the most stable overall and *gender* to exhibit very stable scores, but less stable ranks. We also report strong correlations between *body shape* characteristics, *gender* and *height*, and *skin colour* and *hair colour*, while traits like *hair length* display the most visual uncertainty, due to camera viewpoint occlusion. Furthermore, to our knowledge this is the first study to annotate *gender* comparatively.

Lastly, we reiterate Reid's findings [2], showing that comparative annotations contain more discriminative information than categorical labels. In Chapter 4 we employ this preliminary dataset for automatic label prediction and identification.

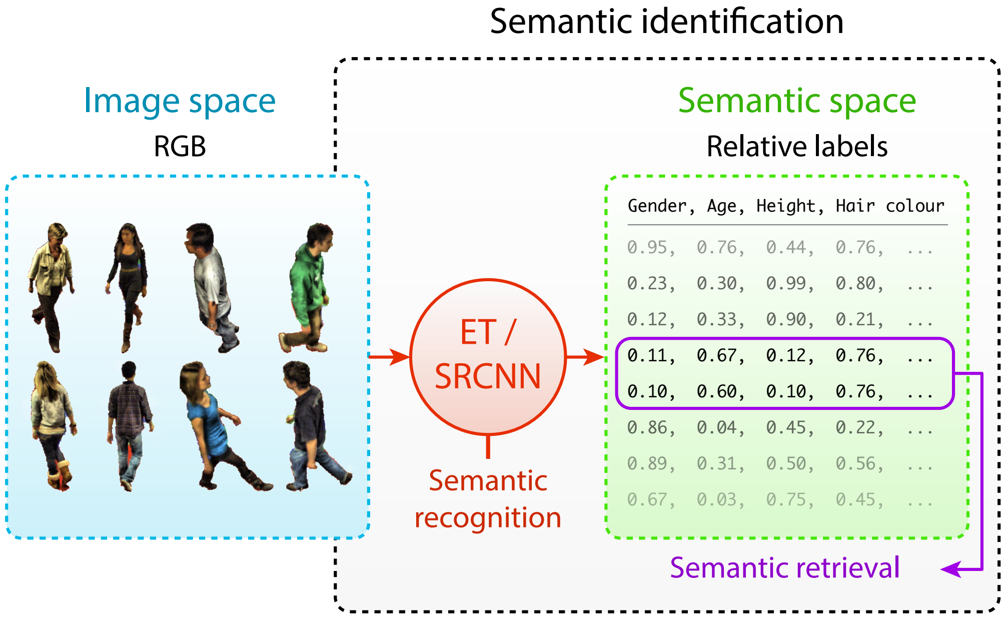
Chapter 4

Automatic Identification with Relative Attributes

4.1 Introduction

In the previous chapter we proposed a methodology to successfully crowdsource precise, relative soft biometric descriptions from humans. We now wish to perform automatic subject identification from these novel ground-truths. This chapter investigates the automatic recognition of traditional categorical and new relative soft biometric labels, contrasting their semantic identification performance. Throughout the chapter, we progress through a broad range of computer vision and machine learning techniques, uncovering methods to optimally recognise soft biometric descriptions from stand-alone subject images.

Section 4.2 considers attribute-based re-identification and image attribute recognition works, introducing the three identification scenarios central to our evaluation methodology. Section 4.3 then presents a novel dataset which enables experimentation with categorical and relative attributes for each identification scenario. Section 4.4 introduces a number of hand-crafted feature descriptor and supervised learning components to facilitate label recognition, analysing the correspondence of each combination in Section 4.5. Subsequently, the most suitable hand-crafted approach is expanded upon in Section 4.6 and an alternative deep learning approach is proposed in Section 4.7, with a comprehensive model training strategy outlined in Section 4.8. Lastly, the three identification experiments are evaluated and extensively analysed in Section 4.9, contrasting the semantic identification performance of each recognition approach. Our findings are summarised in Section 4.10.



```
Gender, Age, Height, Hair colour
0.95,  0.76,  0.44,  0.76,  ...
0.23,  0.30,  0.99,  0.80,  ...
0.12,  0.33,  0.90,  0.21,  ...
0.11,  0.67,  0.12,  0.76,  ...
0.10,  0.60,  0.10,  0.76,  ...
0.86,  0.04,  0.45,  0.22,  ...
0.89,  0.31,  0.50,  0.56,  ...
0.67,  0.03,  0.75,  0.45,  ...
```

Figure 4.1: Visual overview of semantic identification, illustrating semantic recognition and retrieval of estimated relative attributes.

4.2 Approach Considerations

4.2.1 Attribute-based Re-identification

Expanding upon our discussion in Section 2.4.3, attribute-based re-identification is the process of re-identifying a previously unknown subject by matching only automatically recognised attributes. Almost all works alluding to attribute-based re-identification assume binary or categorical ground-truth labels [41, 44, 45, 47, 50, 57, 93–95], limiting their discriminative power in attempts to recover biological ground-truth from subject images. However, we have already demonstrated that comparative soft biometrics outperform categorical counterparts for subject retrieval. Therefore, this chapter focuses on contrasting the identification performance of categorical and relative labels. For the first time we confirm if automatically estimated relative labels also pragmatically benefit automatic identification.

4.2.2 Soft Biometric Recognition

We propose two methods for recognising binary and relative attributes from subject images. The first investigation forms a baseline approach, exploring the use of a number of traditional hand-crafted feature extractors in combination with popular supervised learning algorithms. Very recently, Convolutional Neural Networks (CNNs) and deep learning techniques have become the trend for re-identification metric learning [96–98] and pedestrian attribute prediction [38, 41, 50, 93, 95]. Our second approach therefore investigates a deep learning Semantic Recognition Convolutional Neural Network

(SRCNN), to jointly predict image attributes, contrasting performances between approaches.

Once image attributes are recognised, subjects are then retrieved in the resulting 'semantic space', matching predicted and ground-truth subject descriptions to one another to perform semantic identification, illustrated in Figure 4.1. Specifically, we investigate *absolute binary* (abs-bin), *relative binary* (rel-bin) and *relative continuous* (rel-con) labelling techniques.

4.2.3 Identification Scenarios

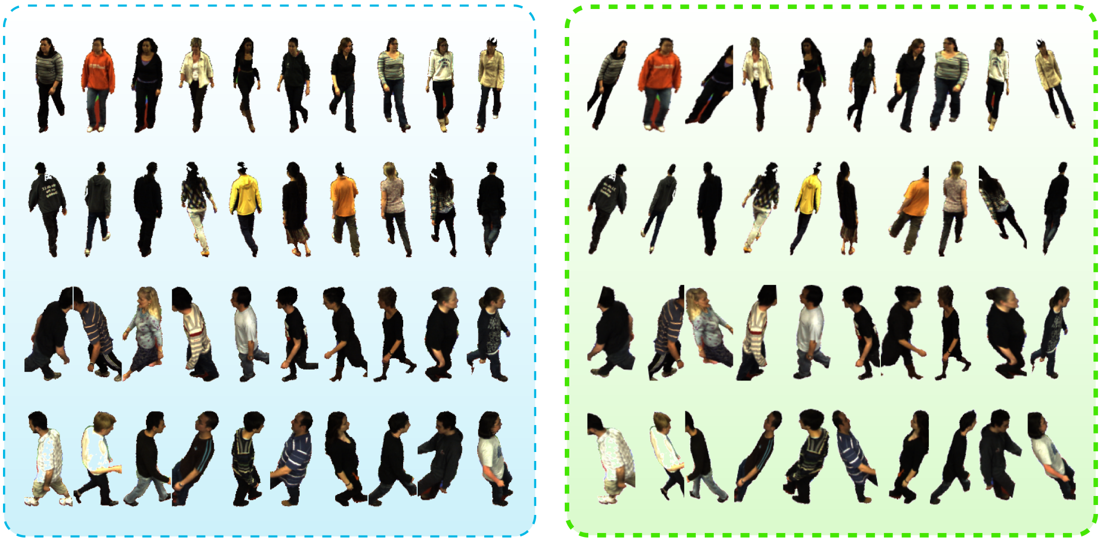
As described in Figure 1.1, soft biometric identification is the culmination of automatic retrieval given human annotated and machine recognised labels from images. As such, we evaluate a number of identification measures, informing us how well each labelling modality communicates subject image identities between humans and machines.

Several methodologies and scenarios exist for evaluating the performance of any given identification solution. As surveillance datasets have become ever larger and more complex, so have the evaluation criteria, in response to real-world needs and requirements. Historically, the most widely reported benchmark re-id datasets (e.g. VIPeR [36] and GRID [99]) only allow for highly constrained, closed-world, one-shot re-identification results to be reported from a single pair of cameras. We therefore introduce a new dataset with which to investigate several re-identification scenarios in conjunction to our new relative attributes. As such, evaluating a broad range of scenarios better informs us on the generalising capability and discriminative power of our techniques. The three identification scenarios are as follows:

Single-shot re-identification. The classic 'biometric identification' scenario, whereby one image of each probe subject is contained in the gallery *training* set and probe *test* set. Probe images are re-identified from a similar viewpoint and environment.

Multi-shot re-identification. The re-identification of probe subjects from non overlapping cameras. Multiple images of the probe maybe present in both the training and test sets, but from disjoint viewpoints and varied environments and poses.

Zero-shot identification. The identification of a previously unseen probe subject. This is the toughest such scenario, whereby the probe subject is not observed during training, therefore an original soft biometric signature must be automatically recognised. This amounts to subject identification, rather than subject re-identification that relies upon perquisite images of the probe subject.



(a) Pre-processed originals.

(b) Augmented variations

Figure 4.2: SoBiR camera views top-to-bottom; front, back, top and side.

4.3 Soft Biometric Retrieval (SoBiR) Dataset

In this section we introduce the Soft Biometric Retrieval (SoBiR) dataset[1], designed to be a pragmatic, flexible and challenging framework with which to investigate automatic semantic retrieval. It comprises a subset of images from the MBT dataset [3], discussed in Section 3.2, drawing labels from MBT's original categorical annotations and our final inferred continuous trait scores, generated in Section 3.5 (Chapter 3).

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 4.2a.

4.3.1 Image Set

We select 2 images for all 100 subjects, sampled at random distance from 4 pairs of cameras; *front*, *back*, *side* and *top* views, exhibiting variations in pose and image quality. Camera placements aim to replicate a high number of surveillance viewpoints as subjects move through the tunnel, capturing both left and right-hand sides in a constrained environment. An *all* view set combines image samples from each camera, enabling multi-shot re-identification and zero-shot identification scenarios. Camera pairs also enable traditional one-shot re-identification, similarly to VIPeR and GRID. In comparison,

[1] SoBiR can be found at http://users.ecs.soton.ac.uk/dmc1g14/#isba-16.

	Comparative (Ours)	**Categorical** [101]
Trait	*[More A, More B]*	*[0, 1, 2, 3, 4, 5, 6, ..]*
Gender	[Feminine, Masculine]	[Female, Male]
Age	[Old, Young]	[Infant, Pre Adolescence, Adolescence, Young Adult, Adult, Middle Aged, Senior]
Height	[Tall, Short]	[Very Short, Short, Average, Tall, Very Tall]
Weight	[Heavy, Light]	[Very Thin, Thin, Average, Big, Very Big]
Figure	[Fat, Thin]	[Very Small, Small, Average, Large, Very Large]
Chest size	[Big, Small]	[Very Slim, Slim, Average, Large, Very Large]
Arm thickness	[Thick, Thin]	[Very Thin, Thin, Average, Thick, Very Thick]
Leg thickness	[Thick, Thin]	[Very Thin, Thin, Average, Thick, Very Thick]
Skin colour	[Dark, Light]	[White, Tanned, Oriental, Black]
Hair colour	[Dark, Light]	[Black, Brown, Red, Blond, Grey, Dyed]
Hair length	[Long, Short]	[Shaven, Short, Medium, Long]
Muscle build	[Muscle, Lean]	[Very Lean, Lean, Average, Muscly, Very Muscly]

Table 4.1: Lexicon of comparative and categorical traits and labels included in SoBiR.

SoBiR images have more consistent lighting and higher resolution, pre-conditioned with basic background subtraction. As the tunnel background consists of a regular grid, seen in Figure 3.1, we chose to segment subject images to avoid unwanted bias associated with the background pattern. However, this approach may be less suited to real-world scenarios, where backgrounds are less regular and segmentation is not applicable.

SoBiR images are pre-processed before performing retrieval, as in Figure 4.2a. Maintaining their aspect ratio, images are first scaled to a height of 256 pixels. A horizontal mid-point is calculated for each image, taking the median value of each row's mean non-white pixel location. Scaled images are then placed in a 256x256 white square, aligning the horizontal mid-point to 128 pixels across. Any image pixels falling outside the 256x256 area are cropped. Lastly, to adjust for low lighting levels in some regions of the tunnel, Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied [100]. This procedure improves an image's local contrast by redistributing lightness values. To mitigate the amplification of noise in homogeneous image regions, histograms are computed across several image neighbourhoods and bin values are clipped and redistributed at a limit of 0.01.

4.3.2 Soft Biometric Labels

SoBiR comprises a compact lexicon of 12 soft traits, drawn from two sources of categorical and comparative ground-truth annotations, recorded in Table 4.1. Samangooei first presented the set of absolute-categorical annotations [101], collecting a number of visually assessable, global and body features. As discussed in Section 3.3.1 we select the 12 most pertinent traits, of the original 23 descriptions. Subjects are described in an absolute sense, using pre-defined categories e.g. 'Very Short', 'Short', 'Average', 'Tall', 'Very Tall' for *height*. We derive absolute-binary (*abs-bin*) representations 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.

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

Table 4.2: Semantic space characteristics of SoBiR labels.

Chapter 3 discusses the crowdsourcing of relative labels, expressed as ordered relations between subject pairs e.g. 'Much More Feminine', 'More Feminine', 'Same', 'More Masculine', 'Much More Masculine' for *gender*. In Section 3.5 relative-continuous (*rel-con*) labels are derived from these responses, by applying a similarity constrained RankSVM to all pairwise comparisons. By ranking subjects on a bi-polar scale, a continuous value is attained to describe every subject's possession of each soft trait, e.g. from 'Most Feminine' to 'Most Masculine' and 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, yet are still relative measurements.

Table 4.2 summarises the characteristics of each labelling modality, noting that 4096 unique combinations are possible for both binary semantic spaces, although only rel-bin labels are perfectly balanced across the subject population. In contrast, the continuous rel-con semantic space offers potentially infinite descriptive combinations.

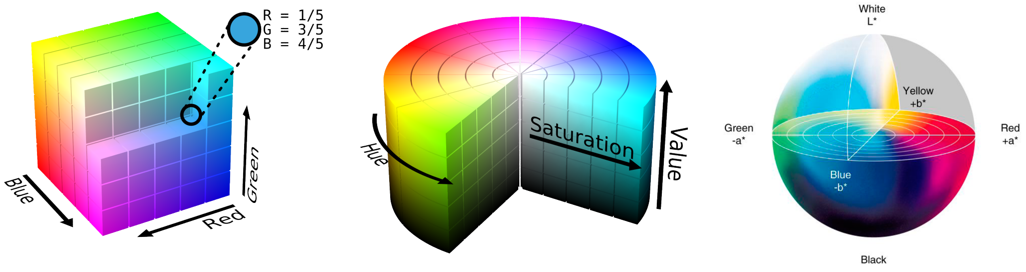
As discussed Section 3.2, associated image labels are an-notated from MBT dataset's *front* camera viewpoint, similarly to Samangooei's and Reid's approaches [1, 2]. As such, we defer tackling issues pertaining to viewpoint specific annotation and label variation until Chapter 6.

4.4 Hand-crafted Recognition Components

Our first approach towards performing comprehensive automatic soft biometric recognition comprises a two-stage, computer vision and machine learning approach. Images are first decomposed through a variety of hand-crafted visual descriptors reducing their dimensionality, in the hope of maintaining descriptive robustness to viewpoint, pose and illumination. Next, a number of supervised regression algorithms are employed to learn a prediction model from the visual features, estimating target soft biometric labels automatically. In this section we introduce several feature descriptor and learning algorithm components, analysing each combination in Section 4.5 and selecting the optimal combination in Section 4.9.

4.4.1 Feature Descriptors

We take inspiration from related literature to decompose images into feature descriptors, ready for predictor training. Our visual feature selection comprises of appearance-based techniques, split into four types; *colour*, *shape & spatial*, *texture* and *compositional*.

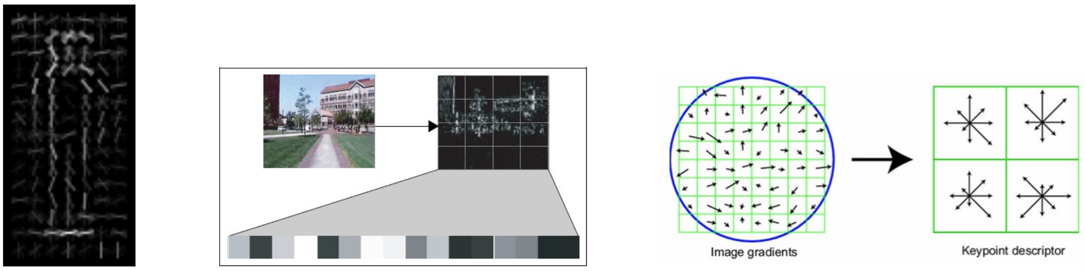


(a) RGB. (b) HSV. (c) CIELAB.

Figure 4.3: Colour space visualisations [102].

Colour histograms. The simplest of the visual feature descriptors. Images are first converted into a particular *colour space*, and values for each *colour channel* are accumulated and binned in a representative histogram.

The colour spaces we investigate are discussed in an early CBIR survey [9] and used in the following works, they include; `RGB` [10, 57, 103, 104], `HSV` (similar to HSL, HSB) [57, 105–108] and `CIELAB` [65, 76, 109]. Our colour histograms use 16 bins for all colour space channels following Layne et al. [37].



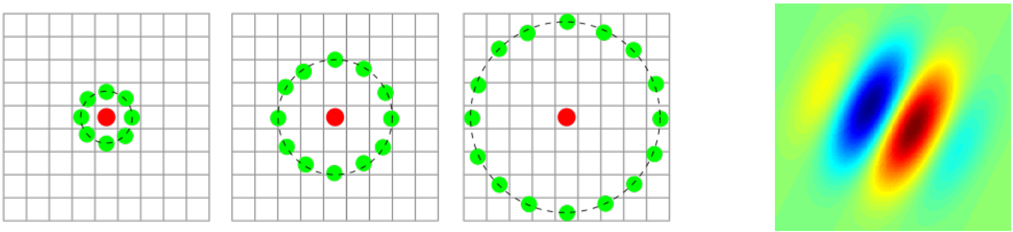
(a) HOG [110]. (b) GIST [111]. (c) SIFT [112].

Figure 4.4: Shape and spatial descriptor visualisations.

Shape & spatial descriptors. The most popular shape descriptor is the Histograms of Oriented Gradients (`HOG`), first described by Dalal and Triggs [110] and implemented in many gender prediction and attribute description studies [76, 103, 108, 109, 113]. `HOG` applies Canny edge detection, splitting the resulting edges into a grid of histograms that locally bin the edge orientations in each grid cell. Our utilisation first resizes images to 128 × 256 pixels, applying `HOG` with a non-overlapping grid of cell size 16 × 16 pixels, each cell described by a histogram of 9 bins, producing 2304 features.

We also include the Holistic Representation of the Spatial Envelope (`GIST`) descriptor, which bypasses segmentation to describe the scene in a fixed length vector [111]. Parikh et al. successfully applied GIST feature extraction to both outdoor scene recognition and face description problems [65].

Lastly, we investigate Local Scale-Invariant Features (SIFT), to decompose the image into a set of key-points representing shape [112], used in [103, 109]. We represent 25 of the most distinguished key-points in our SIFT descriptor.



(a) Uniform LBP [114].

(b) Gabor [115].

Figure 4.5: Texture descriptor visualisations.

Texture descriptors. We apply two texture descriptors, both of which are investigated by Matthews et al. in the search for automatic semantic texture description [92].

Uniform Local Binary Patterns (LBP) first originated for the purpose of texture analysis [114, 116] but have later been found to be applicable to tasks such as gender classification [117]. Similarly to HOG, LBP decomposes the image into a grid, building a histogram for each cell by comparing each pixel to its radial neighbours' intensity values. We select radii of $r = \{2, 4, 8\}$ pixels with 8 bins per histogram.

We also experiment with a set of Gabor filter decompositions (Gabor), first proposed in 1989 [118]. Each Gabor filter is a modulated Gaussian kernel function, used to convolve the image. Similarly to [36, 92], we use a combination of 16 filter parameters; $\theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$, $\sigma \in \{1, 3\}$, $\lambda \in \{0.05, 0.25\}$.

Compositional descriptors. These concatenate a set of descriptors into one vector representation. An example of this is our Colour descriptor, an amalgamation of RGB, HSV and Lab histograms.

We also take inspiration from the Ensemble of Localised Features (ELF) feature representation, where the image is split into 6 horizontal strips, each described by a battery of features [36]. This approach has been used successfully in many related re-identification papers [37, 44, 119, 120]. Our version of ELF combines Colour and Gabor descriptors, splitting images into six equally sized horizontal strips, each described by 9 colour channels; RGB, HSV, CIELAB, and 16 Gabor luminance channel texture filters. Channels are represented using 16 bins, producing a total of 400 features per strip.

Lastly we also experiment with concatenations of HOG with Colour and Gabor features (HOGCB) and HOG with ELF (HOGELF).

4.4.2 Supervised Learning Algorithms

Alongside image feature descriptors, we also investigate a series of supervised learning algorithms. For each learning algorithm, a separate model is independently learnt for each soft trait

In general, these algorithms learn a model to fit a set of training data $\boldsymbol{X} = [\boldsymbol{x}_i] \in \mathbb{R}^p$ (image features) to a set of observed target values $\boldsymbol{y}$ (inferred soft biometric scores). We investigate *regression* predictors, able to estimate continuous values $\hat{\boldsymbol{y}}$ (relative scores), rather than more common *classification* predictors, often utilised with traditionally categorical data, to output a binary class for each attribute.

The most basic regression algorithm employed is linear regression (`LR`), which fits a linear model of coefficients $\boldsymbol{w} = (w_1, ..., w_p)$, to all p features of $\boldsymbol{X}$ to minimise the sum of squares between observations and predictions:

$$\min_{\boldsymbol{w}} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2{}^2$$

Nearest Neighbour algorithm. This widely used technique is conceptually quite simple; test samples are classified as the median class of the k nearest training samples. To output a continuous value, our KNearest Neighbour Regression (`KNN`) method chooses the median value of the nearest neighbouring samples' values, where $k = 5$.

Support Vector algorithm. Support vector models (SVMs) are a popular choice of machine learning algorithm, effective in high-dimensional spaces and easily adaptable to many problems. Support vector machines attempt to separate data-points by constructing a hyper-plane to attain the largest classification margin, while minimising the misclassification error. This trade-off is controlled by parameter $C = 1$. We choose a linear Kernel Ridge Regression formulation (`LKRR`), which regularises coefficients $\boldsymbol{w}$ as follows:

$$\min_{\boldsymbol{w}} ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2{}^2 + \Gamma||\boldsymbol{w}||_2{}^2$$

where $\Gamma = 1$. `LKRR`. This learns a standard support vector model, but combines ridge regression regularisation with the kernel trick, enabling an approximate fitting to be performed faster and more robustly.

Ensemble learning algorithms. Ensemble methods improve generalisability and robustness, by combining several predictors learnt on subsets of the dataset to avoid overfitting. We employ two popular boosting algorithms that combine weak classifiers into a more powerful ensemble of predictors; AdaBoost Regressor (`ABR`) and Gradient Boosting Regression Trees (`GBRT`). Both methods iteratively learn weak classifiers, targeted at learning more difficult samples over already correctly classified predictions.

Lastly, we experiment with two decision tree based predictors; RandomForrest Regressor (`RFR`) and Extremely Randomised Trees Regressor (`ERTR`). Traditionally, decision tree methods learn simple decision rules to predict target values, but can easily become overcomplicated, overfitting training data. Both `RFR` and `ERTR` mitigate this by splitting decision nodes among random subsets of features, averaging values to decrease variance and improve generalisation.

4.5 Hand-crafted Recognition Correspondence

In this section we investigate the correspondences between each visual feature and supervised learning algorithm for estimating relative ground-truth labels. This gives us an initial sense of the recognition performance for each combination, such that we can confidently take forward the optimal choice for further comparative experimentation in Section 4.9.

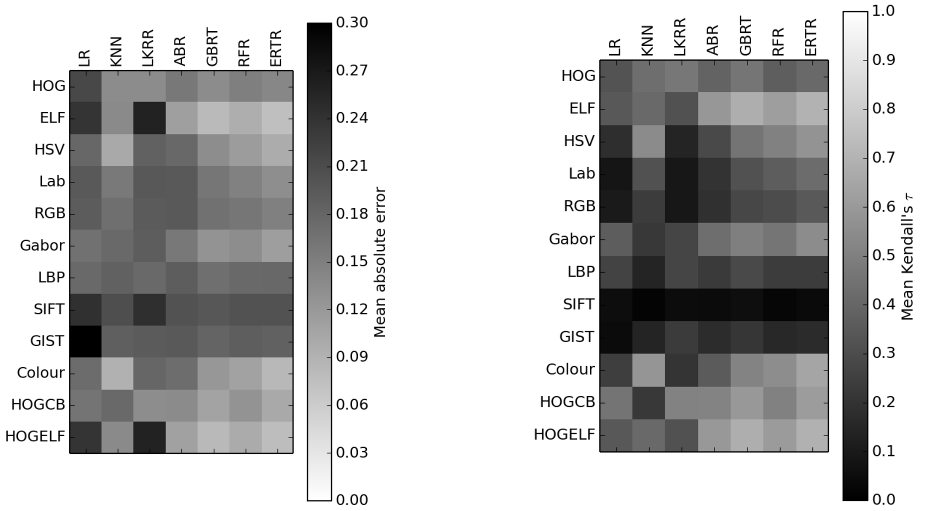
4.5.1 Methodology

We experiment with one-shot and zero-shot identification scenarios, performing 5 cross-fold validation with results reported over an average of 3 iterations. The image dataset is sampled from a mirrored pair of forward-facing cameras, containing 5 images per camera subject, selected at a close position to each camera (a total sample of 1000 images). A predictor is trained for each soft trait and estimated values are compared by Euclidean distance to the original inferred relative trait scores.

For each experiment, two sets of averaged results are reported. The absolute ℓ_1 norm error and Kendall's τ rank correlation coefficient between estimated and original scores. Images are mirrored vertically at run-time, generating a second descriptor to provide higher viewpoint invariance.

4.5.2 Analysis

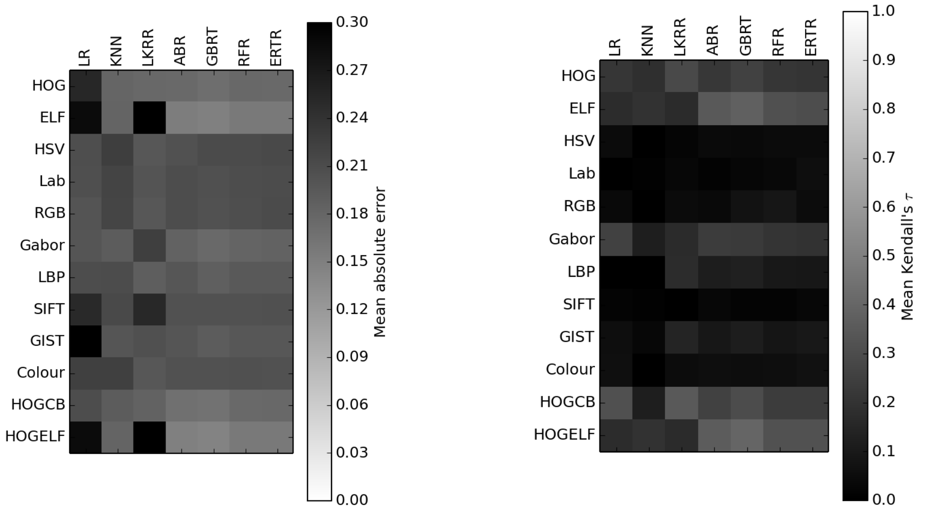
Results are visualised as confusion matrices in Figures 4.6 and 4.7, contrasting all descriptor and predictor combinations. As expected, one-shot re-identification, Figure 4.6, reveals lower absolute errors and higher rank correspondences compared to zero-shot identification, Figure 4.7. This demonstrates that certain feature and predictor combinations are able to learn models that successfully separate individuals when trained



(a) Mean absolute ℓ_1 error.

(b) Mean Kendall's τ.

Figure 4.6: One-shot recognition scenario.



(a) Mean absolute ℓ_1 error.

(b) Mean Kendall's τ.

Figure 4.7: Zero-shot recognition scenario - lighter shades represent lower error and higher correlation.

and tested on highly similar image sets, but do not generalise as well when estimating unseen subjects in the zero-shot scenario.

More basic predictors tend to perform adequately with simpler visual features in an one-shot, but are likely to overfit when tested in zero-shot or with more complex features. An example of this is `KNN` performing very well with the `Colour` descriptor in one-shot testing, but failing to attain much correspondence in the zero-shot scenario.

Additionally, certain visual features such as `SIFT` and `GIST` are able to be estimated to some degree of accuracy when measuring absolute error. However, on inspecting rank correlations, we find these descriptors perform very poorly overall, indicating weak correspondence between soft traits and the underlying visual features, with models poorly explaining any covariance.

The best performing combinations are effective at generalising in both scenarios, demonstrating that models are not simply overfitting to a particular feature set. Ensemble methods prove very powerful, performing best when applied with more complex features such as `HOG`, `ELF`, `HOGCB` and `HOGELF`, showing the ability of ensemble algorithms to successfully harness extra informative features and ignore redundant information. Evidently, the localised colour and texture features in `ELF` provide some of the most successful descriptors, proving even more powerful when combined with the `HOG` shape descriptor. We therefore select the `HOGELF` descriptor for future experimentation, as it provides the best results in both scenarios when combined with the ensemble learning methods.

The `GBRT` predictor is the strongest performer in both scenarios, offering the best bias-variance trade-off. However, we select to take forward the `ERTR` predictor due to its much more efficient computation time, consistently outperforming the next best `ABR` predictor in one-shot recognition.

4.6 ET Approach

Following Section 4.5 we select the `HOGELF` descriptor and the `ERTR` predictor as the optimal trade-off between recognition performance and pragmatic run-time for more intensive experimentation. In this Section we elaborate on our hand-crafted approach, ready for soft biometric identification in Section 4.9.

For each trait $q \in 1, ..., Q$, we train an independent model to learn one soft biometric label. Collectively, the models predict $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_{\boldsymbol{i}}] \in \mathbb{R}^Q$ labels for $i \in 1, ..., N$ subject samples. Models are trained given a set image descriptors with p features, $\boldsymbol{X} = [\boldsymbol{x}_{\boldsymbol{i}}] \in \mathbb{R}^p$, and ground-truth target labels $\boldsymbol{y} = [\boldsymbol{y}_{\boldsymbol{i}}] \in \mathbb{R}^Q$. For binary labelling we use an the Extremely Randomised Trees Classifier (`ERTC`), while for continuous labelling we use the Extremely Randomised Trees Regressor (`ERTR`) variation.

During the first round of training we perform a parameter grid search with 2-fold cross validation. The best parameter set for each soft trait predictor is chosen by minimising the average sample loss $L(q)$. We search for both the number of estimators $n_{est} \in \{1, 5, 10, 50, 100, 200, 300, 400\}$ and maximum features $n_{max} \in \{sqrt, log\}$.

To train models independently, the loss of each attribute q is calculated between each subject's ground-truth y_{iq} and predicted $\hat{y}_{iq}$ labels, averaged over N training subjects:

$$L(q) = \frac{1}{N}\sum_{i \in N} L\prime(y_{iq}, \hat{y}_{iq}). \tag{4.1}$$

For binary labelling, we define $L\prime$ between subject i's ground-truth and subject j's predicted values, such that $L(q)$ is the Hamming loss for attribute q:

$$L\prime_{HAM}(y_{iq}, \hat{y}_{jq}) = 1(y_{iq} \neq \hat{y}_{jq}), \tag{4.2}$$

For continuous labelling, we define $L\prime$ between subject i's ground-truth and subject j's predicted values, such that $L(q)$ is the Mean Squared Error loss for attribute q;

$$L\prime_{MSE}(y_{iq}, \hat{y}_{jq}) = (y_{iq} - \hat{y}_{jq})^2. \tag{4.3}$$

4.7 SRCNN Approach

In our second retrieval approach we propose a deep learning, feed-forward, Semantic Recognition Convolutional Neural Network (SRCNN) architecture to jointly learn and predict a set of semantic attributes from input images, illustrated in Figure 4.8. We explain the overall architecture of our SRCNN and detail the training strategy employed to alleviate overfitting.

By designing the Deep Neural Network (DNN) as a whole, image feature maps and attributes are learnt in conjunction, overcoming many of the challenges associated with empirically matching feature descriptors to machine learning methods. In contrast to the hand-crafted feature approach, DNNs can also jointly model a large number of attribute outputs, capturing additional covariance between traits, not possible when modelling predictors disjointly.

4.7.1 Architecture

To construct our custom SRCNN architecture for joint attribute estimation, we take inspiration from Zhu's Multi-label CNN (MLCNN) [41]. Rather than apply a single, very deep CNN classifier to the entire image, disregarding spatial feature arrangements, Zhu et al. utilise a grid of smaller CNNs to localise trait characteristics to specific image regions. The paper estimates 21 binary attributes, reporting *gender* accuracies of 69.6 ± 2.6% and 68.4 ± 1.8% from the VIPeR and GRID datasets respectively, improving average attribute recognition accuracy by 1.5% and 0.4% over a standard full-body CNN approach [41]. This approach is similar to the traditional `ELF` and `HOG` descriptors, which

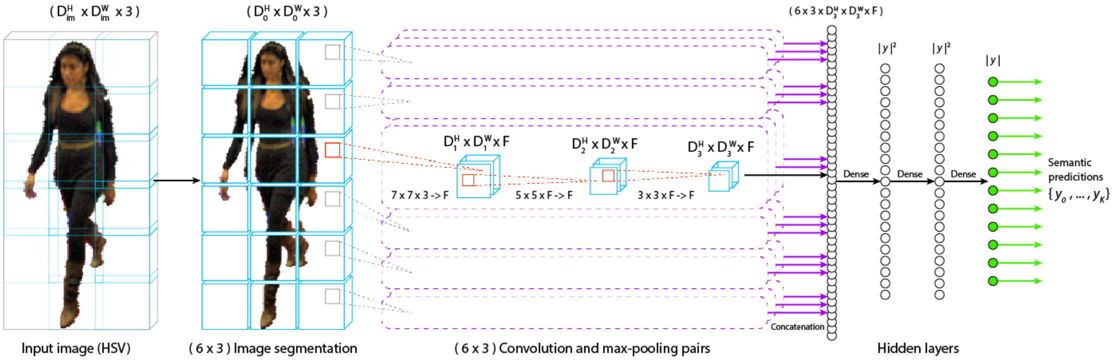


Figure 4.8: Semantic Retrieval Convolutional Neural Network architecture.

concatenate descriptors of multiple, fixed image regions. However, in this instance, convolutional features are subsequently fed into two fully connected layers for joint classification.

Our SRCNN implementation accepts input images of size $D^H_{im} \times D^W_{im} \times 3$, 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, influenced by our findings in Section 4.5.

Therefore, images are divided into a grid of 6×3 overlapping cells in place of body-part detection, similarly to [41]. 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 Q^2. The last fully-connected layer represents the final output, of size Q. In this way, attributes are jointly learnt, exploiting any relationships that occur between labels. As many of our soft traits are global descriptions, unlike [41] we do not predefine connections between image regions and semantic attributes, such that correlations may be learnt automatically during training to generalise our solution. Coincidentally, Zhu et al. mirror this update, opting for fully connected output layers in their later work [95]. We use a sigmoid activation for the final layer and Rectified Linear Unit (ReLU) activations for all other layers.

4.7.2 Loss Functions

Two separate multi-label loss functions are defined for classification and regression formulations, learning joint estimators simultaneously and where all targets are in the range $[0, 1]$. For binary classification we calculate the joint-loss using Binary Cross-Entropy between each subjects' predicted and ground-truth values for all attributes $q \in Q$ per subject i:

$$L_{BCE}(\boldsymbol{y_i}, \boldsymbol{\hat{y}_i}) = -\frac{1}{Q}\sum_{q \in Q}\Big(y_{iq}\log(\hat{y}_{iq}) + (1 - y_{iq})\log(1 - \hat{y}_{iq})\Big). \tag{4.4}$$

For continuous regression, we calculate the joint Mean Squared Error loss for all attributes $q \in Q$ per subject i:

$$L_{MSE}(\boldsymbol{y_i}, \boldsymbol{\hat{y}_i}) = \frac{1}{Q}\sum_{q \in Q}(y_{iq} - \hat{y}_{iq})^2. \tag{4.5}$$

To optimise all 1,259,436 parameters, the SRCNN is trained through back propagation with the ADADELTA stochastic gradient descent method [121] with $lr = 10^{-3}, \beta_1 = 0.9, \beta_2 = 0.999$ and random weight initialisation sampled from a normal distribution truncated at two standard deviations. The solution is implemented in Python using the Theano library and run on a GPU using CUDA and CuDNN.

4.8 Training Strategy

We employ several training strategies to reduce overfitting and help find robust solutions for both the ET and SRCNN approaches. For SRCNN classification tasks, dropout regularisation [122] is applied between convolutional layers, with a dropout ratio of 0.5. It was found that including dropout for certain SRCNN regression experiments excessively prolonged training time, due to the characteristics of the MSE loss function, discussed in Section 4.9.1. No subsequent fine-tuning is required after implementing our training strategy.

4.8.1 Data Augmentation

During ET and SRCNN training steps, 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 $x_s \in [-\frac{D^W_{im}}{5}, \frac{D^W_{im}}{5}]$. Finally, horizontal translation is applied in the range $x_t \in [-\frac{3D^W_{im}}{10}, \frac{3D^W_{im}}{10}]$. 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 can be seen in Figure 4.2b.

4.8.2 Early Stopping

To mitigate overfitting of the deep SRCNN network, we define an early stopping function, based on the semantic retrieval accuracy of the validation set, rather than on its loss value as is common. Section 4.9.1 experimentally shows our reasoning for doing this, as in some cases recognition accuracy continues to increase although loss values cease to be reduced. Therefore, training is halted if $AR(e) < AR(e-w)$ and $e > w$, where $AR(e)$ represents the average semantic retrieval rank of the validation set at epoch e. A trailing window of $w = 30$ epochs is chosen, balancing premature stopping against responsiveness.

4.8.3 Attribute Retrieval Weighting

Once the recognition prediction models are learnt and soft biometric labels have been estimated, subjects are retrieved by matching their predicted labels to ground-truth signatures. However, soft traits do not have equal discriminative ability, affected by label distributions, recognition accuracy and ground-truth annotation methods. Therefore we discover a set of attribute weightings to optimise semantic retrieval performance post-training.

The objective function finds a weighting vector of coefficients $\boldsymbol{w} \in \mathbb{R}^Q$, such that the linear combination of attribute losses is minimised for same-identities, in relation to different-identities, on the validation set. Decomposing the problem, we wish to attain a lower loss value between probe i's estimated $\boldsymbol{\hat{y}_i}$ and ground-truth target $\boldsymbol{y_i}$ labels, $\boldsymbol{L}(\boldsymbol{y_i}, \boldsymbol{\hat{y}_i})$ (same-identity loss), than between probe i's estimated label $\boldsymbol{\hat{y}_i}$ and gallery subject j's ground-truth target $\boldsymbol{y_j}$ labels, $\boldsymbol{L}(\boldsymbol{y_j}, \boldsymbol{\hat{y}_i})$ (different-identity loss), for all probe

	No. Samples		**No. Subjects**		**No. Cameras**	
Experiment	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 4.3: Non-overlapping train (Tr.), validation (Va.) and test (Te.) set criteria.

and gallery subjects $i, j \in N$:

$$\min_{\boldsymbol{w}} \sum_{i\in N} \left(\sum_{j\in N \wedge j\neq i} \begin{cases} 1, & \text{if } \boldsymbol{w}^T\boldsymbol{L}(\boldsymbol{y_j}, \hat{\boldsymbol{y}}_{\boldsymbol{i}}) < \boldsymbol{w}^T\boldsymbol{L}(\boldsymbol{y_i}, \hat{\boldsymbol{y}}_{\boldsymbol{i}}) \\ 0, & \text{otherwise} \end{cases} \right)^{\lambda} + \Gamma||\boldsymbol{w} - 1||_2{}^2 \tag{4.6}$$

where $\boldsymbol{L}$ is the vector loss between subject i's predicted and subject j's ground-truth values for all Q attributes:

$$\boldsymbol{L}(\boldsymbol{y_i}, \hat{\boldsymbol{y}}_{\boldsymbol{j}}) = L\prime(y_{iq}, \hat{y}_{jq}) : q \in Q \tag{4.7}$$

and $L\prime$ is either the Hamming $L\prime_{HAM}$ loss for binary labels or Mean Squared Error $L\prime_{MSE}$ loss for continuous labels.

Choosing the term $0 < \lambda < 1$ gives precedence to improving already low ranks over higher ones. We empirically choose $\lambda = 0.8$ to prioritise the number of low-end ranks, while still minimising the average overall rank. A regularisation constant of $\Gamma = 10^4$ is chosen to appropriately scale the ℓ_2 distance term relative to sum of loss differences. To prevent early overfitting, the coefficients of $\boldsymbol{w}$ are randomly initialised in the range of $[0.99, 1.01]$.

We model the rank optimisation problem as a black box without known derivates. Derivate-free optimisation is an extensive area of research and a large number of algorithms exist to solve our problem. We opt for Powell's popular Constrained Optimization BY Linear Approximations (COBYLA) heuristic search method, which utilises a simplex of $m + 1$ points to construct an approximation to the objective function for minimisation [123]. This is a form of 'trust region' algorithm, whereby the region constrained by the simplex is expanded (trusted) if the objective function is adequately approximated within it, otherwise the region is contracted to formulate a better approximation within the area. The algorithm terminates when the trust region volume reaches a predefined lower bound of 10^{-4}.

Method	Label	r=1	r=5	r=10	r=25	nAUC
One-shot re-identification (average)						
SRCNN	abs-bin	29.2±7.3	50.8±8.9	58.1±8.1	71.0±6.0	80.7±4.6
SRCNN	rel-bin	30.9±7.7	53.4±9.6	61.5±9.6	74.3±7.63	82.0±5.5
SRCNN	rel-con	**35.7±12.2**	**56.8±11.3**	**67.2±9.6**	84.1±6.6	**88.1±4.5**
ET	abs-bin	13.3±2.9	37.0±5.4	53.9±5.2	75.2±3.5	83.2±2.1
ET	rel-bin	20.1±3.4	49.5±4.3	67.1±4.9	**86.3±1.5**	**88.1±1.7**
ET	rel-con	10.3±2.8	32.4±3.1	50.5±4.0	79.2±3.3	84.2±1.9

Table 4.4: SoBiR one-shot retrieval with SRCNN and ET, reporting CMC% and normalised Area Under Curve (nAUC). (**Bold**) highest match rate.

4.9 Soft Biometric Identification

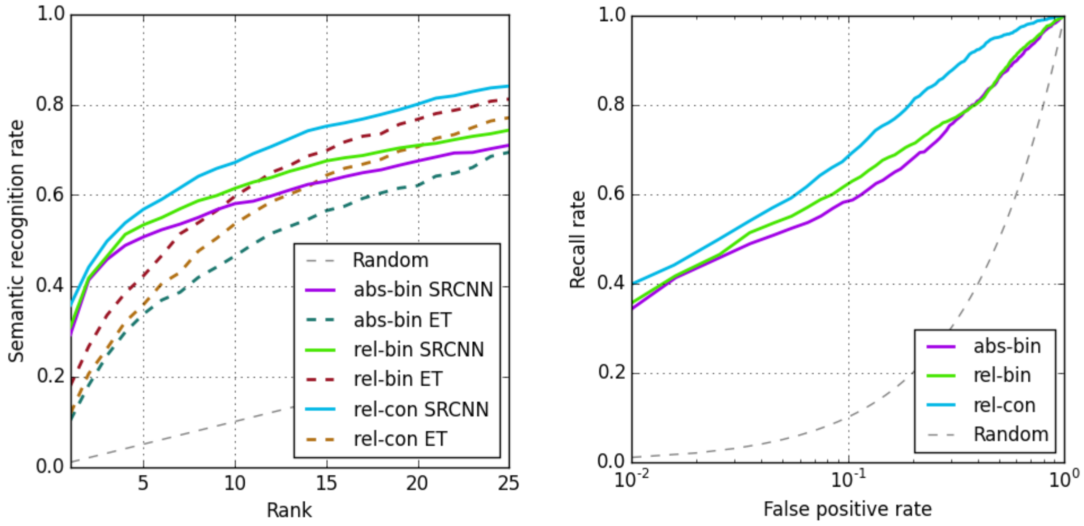
We present three experiments evaluating *one-shot*, *multi-shot* and *zero-shot* identification, each comparing abs-bin, rel-bin and rel-con soft biometric label modalities described in Section 4.3.2. Experiments follow the non-overlapping, train-validation-test split criteria in Table 4.3.

One-shot re-identification. Semantic recognition is performed with one camera pair at a time, randomly selecting alternative train-test viewpoints per subject. Therefore, subject images exhibit minimal variation between instances, replicating an idealised re-identification scenario. The challenge with this situation is lack of training data, as only one camera set is supplied at a time. The process is repeated across all four camera pairs from the SoBiR dataset.

Multi-shot re-identification. One camera image per subject is sampled for the test set and the remaining 7 camera images are allocated to the training set. While more variation is found between subject image instances, the increased volume of training data enables predictors to learn more robust models over one-shot re-identification.

Zero-shot identification. The most challenging evaluation, simulating real-world operation. The scenario performs identification of previously unseen suspects, given only an eye-witness description. 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. As evaluation is performed on totally unseen subjects, trained models are penalised at test time for any form of overfitting.

Cumulative Match Characteristic (CMC) and Receiver Operating Characteristic (ROC) 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 recognising semantic labels from a sample image, and then performing semantic retrieval against a gallery of known descriptions.



(a) One-shot CMC (average).

(b) One-shot ROC (average).

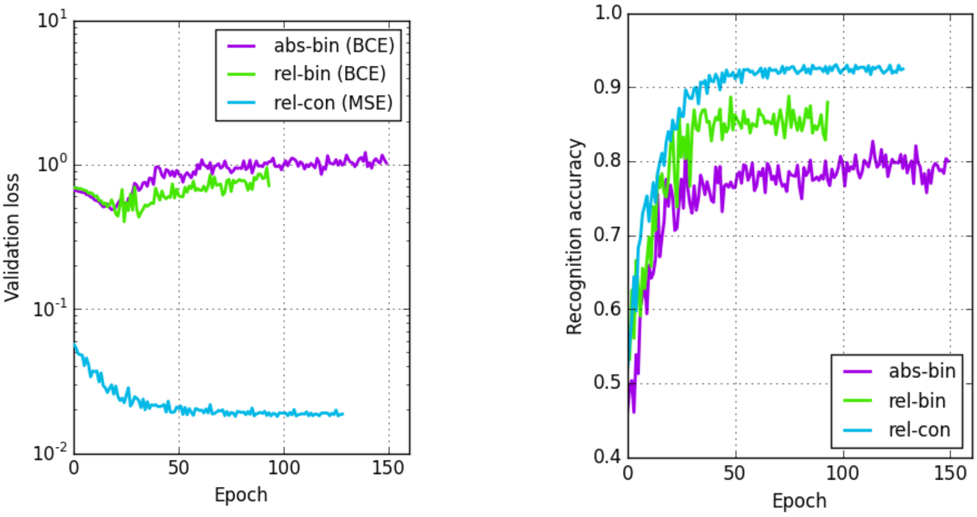
Figure 4.9: SoBiR one-shot identification with SRCNN and ET.

4.9.1 One-shot Re-identification

Table 4.4 and Figure 4.9a and 4.9b summarise the average one-shot semantic identification results achieved by applying ET and SRCNN to *front*, *back*, *top* and *side* camera views from SoBiR. In all three labelling modes SRCNN clearly surpasses the ET semantic recognition approach at lower ranks, with a top rank-1 increase of 23.2%. At higher ranks, SRCNN is overtaken by ET, with the balanced rel-bin labels performing particularly well. This may indicate that the deep learning method is overfitting to the limited number of samples in this scenario, improving lower ranks at the expense of generalisation and the ET model is more successful in mitigating this affect with slightly higher overall nAUC scores.

Critically, SRCNN's rel-con retrieval accuracy outperforms both abs-bin and rel-bin modes, gaining an average of 6.5% and 4.8% at rank-1 respectively, although exhibiting very high stand deviations. This shows that the deep learning solution is generally able to capture more correspondence with rel-con labels, which may also make this modality more sensitive to variations in camera viewpoint. Furthermore, all forms of relative label outperform absolute labels with both models.

To investigate why rel-con significantly outperforms abs-bin and rel-bin with SRCNN, we plot a side-by-side example of the validation loss and average label prediction accuracy during training time in Figure 4.10. In rel-con recognition mode, the validation loss is consistently minimised, while attribute prediction accuracy steadily increases. Meanwhile, in binary classification modes, validation loss values reverse around epoch 25, as recognition accuracy continues to increase, breaking their monotonicity. This suggests



(a) Validation loss.

(b) Validation recognition accuracy.

Figure 4.10: Example validation set characteristics, computed during training time, contrasting validation loss and average recognition accuracy.

	ET Soft trait weightings mean±std			
Trait	abs-bin	rel-bin	rel-con	Average
Gender	**1.59±0.22**	1.13±0.13	**1.19±0.20**	**1.30±0.18**
Height	1.00±0.22	**1.55±0.50**	**1.22±0.31**	1.26±0.35
Age	0.78±0.26	1.13±0.42	0.85±0.15	0.92±0.27
Weight	0.82±0.22	0.79±0.25	0.87±0.04	0.83±0.17
Figure	0.82±0.07	1.04±0.38	1.09±0.11	0.98±0.19
Chest size	0.82±0.26	1.04±0.13	0.83±0.09	0.90±0.16
Arm thickness	1.18±0.18	1.04±0.08	0.94±0.30	1.06±0.18
Leg thickness	0.85±0.11	0.75±0.42	0.80±0.15	0.80±0.19
Skin colour	**1.26±0.18**	**1.21±0.17**	1.11±0.07	**1.19±0.14**
Hair colour	**2.55±0.22**	**1.59±0.42**	**1.78±0.11**	**1.64±0.25**
Hair length	**1.26±0.15**	**1.17±0.13**	**1.37±0.24**	**1.27±0.17**
Muscle build	1.11±0.11	0.63±0.79	0.80±0.19	0.84±0.36

Table 4.5: ET one-shot re-identification optimised soft trait weightings (top four emboldened).

that SRCNN is still learning important prediction decisions, enhancing overall recognition performance, yet overfitting to a subset of samples, negatively affecting validation loss.

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.

Method	Label	r=1	r=5	r=10	r=25	nAUC
(a) Multi-shot re-identification						
SRCNN	abs-bin	43.0±6.3	68.8±5.9	75.8±5.9	85.2±7.2	89.8±3.0
SRCNN	rel-bin	43.2±6.1	67.6±8.6	76.0±6.5	82.8±7.1	88.1±3.6
SRCNN	rel-con	**46.4±4.5**	**72.2±4.4**	**81.8±3.2**	**90.2±2.4**	**92.8±0.8**
ET	abs-bin	13.3±2.9	37.0±5.4	53.9±5.4	75.2±3.5	83.2±1.6
ET	rel-bin	20.1±3.4	49.5±4.3	67.1±4.9	86.3±1.5	88.1±0.7
ET	rel-con	10.3±2.8	32.4±3.1	50.5±4.90	79.2±3.3	84.2±1.4
(b) Zero-shot identification						
SRCNN	abs-bin	2.8±2.2	11.4±5.4	19.7±6.9	41.2±9.0	66.5±5.1
SRCNN	rel-bin	1.0±2.3	5.0±3.4	13.5±8.5	39.8±6.7	61.7±4.4
SRCNN	rel-con	4.8±3.1	15.5±5.6	29.0±7.8	**58.5±6.2**	**71.5±3.0**
ET	abs-bin	1.7±3.2	15.2±7.1	**28.2±9.6**	48.4±10.5	66.5±5.7
ET	rel-bin	2.6±2.3	15.8±6.6	26.4±8.9	52.3±9.3	70.8±4.3
ET	rel-con	**6.7±6.7**	**16.4±11.2**	27.9±13.4	53.9±14.1	70.2±8.4

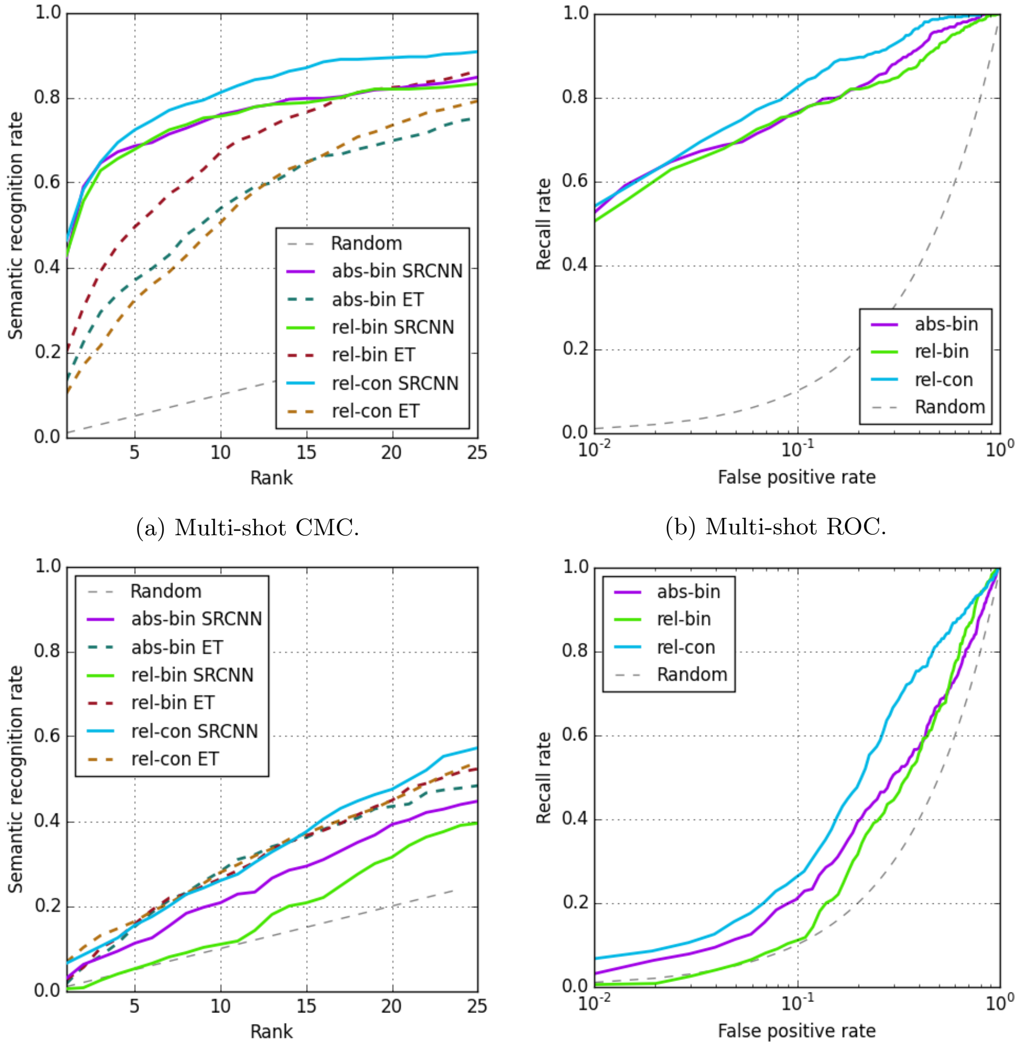
Table 4.6: SoBiR multi-shot and zero-shot with SRCNN and ET, reporting retrieval CMC% and normalised Area Under Curve (nAUC).

We also report the optimised one-shot attribute weightings for ET in Table 4.5. As certain traits perform distinctly better than others, weighting coefficients are divisive across all labelling modalities. Interestingly, alongside *gender*, *height* and *skin colour*, the weighting scheme consistently finds *hair colour* and *hair length* to be most salient, although often occluded in certain camera viewpoints. Furthermore, *leg thickness* and *weight* are the least informative traits for retrieval, closely followed by *muscle build* which are all visually prominent. This indicates that a trait's visibility only partially explains its significance and in fact, traits which exhibited more stable annotation characteristics in Section 3.6.2 are more likely to be essential for identification. The exception is *age*, which is notoriously hard to estimate visually and may overly rely on visual cues from the face.

Overall, attribute weightings improve ET's nAUCs as little as 1.5% for abs-bin, to as much as 7.1% for rel-bin. In contrast, weightings improve SRCNN's nAUC less than 0.5% and weight coefficients vary less than 10^{-3}, indicating that attribute estimations are well optimised and of near equal importance for retrieval. Therefore, we conclude SRCNN is fully exploiting intercorrelations between traits as well as inferring accurate image predictions, having learnt to estimate traits jointly.

4.9.2 Multi-shot Re-identification & Zero-shot Identification

Figure 4.11 reports the remaining two sets of results, again performing semantic recognition with SoBiR, but now in multi-shot and 'open-world' zero-shot scenarios. SRCNN now far outperforms the ET approach for multi-shot re-identification, gaining a top rank-1 increase of 26.3%, Table 4.6a. By providing a larger number of training samples per probe subject, recognition and retrieval performances are radically improved.



(a) Multi-shot CMC.

(b) Multi-shot ROC.

(c) Zero-shot CMC.

(d) Zero-shot ROC.

Figure 4.11: SoBiR multi-shot and zero-shot semantic identification with SRCNN and ET.

In stark contrast, zero-shot identification attains relatively low recognition performance across all labels, with SRCNN rel-bin predictions fairing particularly poorly. In fact, only SRCNN rel-con labelling offers any improvement over the ET approach in this scenario, gaining 1.3% nAUC, Table 4.6b. Furthermore, standard deviations are considerably large for low zero-shot ranks with ET rel-con, yet less than half that for SRCNN rel-con. Intriguingly, SRCNN's abs-bin slightly 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 for zero-shot identification, attribute retrieval rates are drastically reduced, similarly to Layne et al.'s findings [44]. This shows that while SRCNN is able to recognise 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. Compared to single-shot, the multi-shot experiment requires on average 1.6× more training epochs and zero-shot requires 0.4× fewer epochs on SRCNN, indicating the extent and depth of each learning process.

4.10 Conclusions

In this chapter we perform automatic person description and identification using three soft biometric modalities from a novel, publicly accessible dataset, SoBiR. Two retrieval solutions are proposed; a baseline hand-crafted (ET) method and a deep learning (SR-CNN) method. Both approaches learn and predict semantic attributes from stand-alone images and enable identification from eye-witness testimony.

Our results indicate that relative labels are not only more discriminative than binary alternatives, but also enhance semantic identification performance when recognised automatically. Importantly, these findings are reiterated across all three identification scenarios and for both ET and SRCNN approaches, highlighting the aptness of relative attributes and the necessity for higher precision descriptions in recognition and identification.

Furthermore, we demonstrate a significant step in semantic recognition and identification performance with deep learning. SRCNN achieves a top rank-1 increase of 23.2% and 26.3% over the ET method in one-shot and multi-shot re-identification scenarios on the SoBiR dataset. Notably, this is the first time a deep learning approach has been employed to regress subject labels, clearly surpassing binary classification. While extending the potential of soft biometric identification, issues pertaining to zero-shot learning must be tackled to facilitate suspect identification from eye-witness testimony and to enable truly ubiquitous, real-world applications. Encouragingly, our findings provide a clear direction to approach this in future chapters; combining the power of deep learning with high fidelity attribute descriptions.

So far our study extends front-facing annotations to multi-viewpoint images, assuming coherence across viewpoints. We have not yet concerned questions surrounding the applicability of such an approach at scale or with highly unconstrained data. Therefore, in Chapter 5 we increase the scope of our annotation to address questions regarding the large-scale collection of precise soft biometrics descriptions, building upon the knowledge gained so far.

Chapter 5

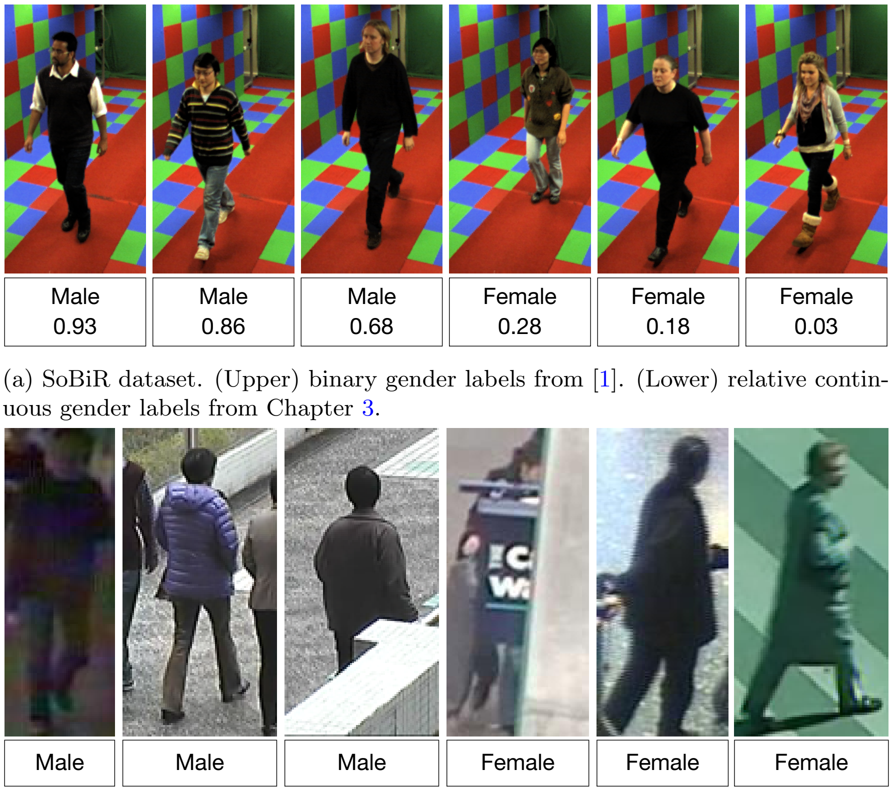
Categorising Gender with Super-Fine Attributes

5.1 Introduction

Chapters 3 and 4 introduce techniques to first collect, then automatically estimate precise human descriptions from images, successfully enhancing subject identification. We now build upon this new found knowledge and tackle the predominant challenges arising from this work, returning back to the human annotation of soft biometrics. In surveillance, expert operatives and non-expert eyewitnesses must generate search queries relying solely on their perception of the surveillance imagery. It is therefore imperative to understand exactly *what* users perceive in such images, in order to generate more illustrative descriptions and eliminate restrictive expertly-defined vocabularies.

This chapter provides an in-depth analysis of gender-from-body, the single most pertinent and commonly annotated human image attribute. This enables succinct experimentation with new labelling techniques, without the added complexity of interactions between multiple traits and modalities. Our aim is to overcome issues pertaining to restrictive description methodologies and very low quality surveillance images, by exploring an innovative, more apposite and more pragmatic annotation method.

Section 5.2 exemplifies the intricacies of perceiving gender as an identifying human characteristic leading into Section 5.3, where we introduce the idea of *super-fine attribute descriptions*, enhancing upon traditional relative and 'fine-grained' attributes. Section 5.4 proposes a novel similarity comparison crowdsourcing methodology and Section 5.5 discusses the formation of super-fine attributes by topologically interpreting pairwise similarity responses. Section 5.6 then compares a one-dimensional ranking approach to previous measures inferred from ordered comparisons and experiments with novel multi-dimensional trait descriptions. Lastly, Section 5.7 concludes our findings.



(a) SoBiR dataset. (Upper) binary gender labels from [1]. (Lower) relative continuous gender labels from Chapter 3.

(b) PETA dataset, original binary gender labels from [57].

Figure 5.1: (a) Categorising gender on a sliding scale. (b) Examples of subject images in which gender is hard-to-see.

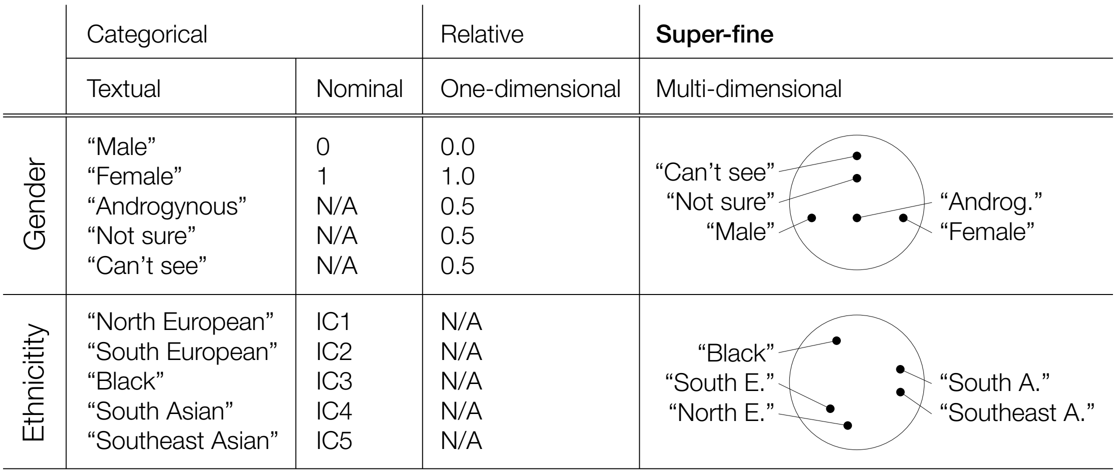
5.2 Gender Identity & Perception

Gender identity is something every human possesses and is the most commonly labelled human image attribute, typically categorised as binary 'male' or 'female' classes [1, 30, 57]. As a contemporary topic, gender has been shown to be perceived on a sliding scale [124] (as in Chapter 3) and more complex representations of gender are becoming widespread, with services like Facebook now offering 71 gender options in the UK[1] and US universities offering non-binary pronouns beyond 'he' and 'she'[2].

Visually discerning gender is highly dependent on the observer, intertwining multiple features and cultural cues e.g. face shape, chest size, body proportions, hair length, clothing appearance, accessories, make-up etc. However, such cues are not always discernible, and unfamiliar combinations can be contradictory. Figure 5.1a highlights varying degrees of masculinity to femininity in visually clear images from the SoBiR dataset, while Figure 5.1b illustrates the difficulty in perceiving gender in real-world surveillance footage. Intriguingly, both image sets are originally manually labelled with, sometimes controversial, binary ground-truths.

[1] https://www.facebook.com/facebookdiversity/posts/774221582674346

[2] http://www.bbc.co.uk/news/magazine-34901704



	Categorical		Relative	**Super-fine**
	Textual	Nominal	One-dimensional	Multi-dimensional
Gender	“Male” “Female” “Androgynous” “Not sure” “Can’t see”	0 1 N/A N/A N/A	0.0 1.0 0.5 0.5 0.5	“Can’t see” “Not sure” “Male” “Androg.” “Female”
Ethnicity	“North European” “South European” “Black” “South Asian” “Southeast Asian”	IC1 IC2 IC3 IC4 IC5	N/A N/A N/A N/A N/A	“Black” “South E.” “North E.” “South A.” “Southeast A.”

Figure 5.2: Exemplifying the difficulty of describing gender and ethnicity traits through traditional textual, categorical, one-dimensional rankings. Super-fine attributes now describe concepts as coordinates in multi-dimensional space, capturing complex inter-concept relations (illustrated example).

In state-of-the-art image attribute studies binary gender-from-body recognition performs significantly worse than recognition from face [4, 5, 31], and even human performance is often far from perfect [30, 37]. This further suggests that binary categorisation may not always be suitable, especially when dealing with highly unconstrained imagery, captured ‘in-the-wild’.

Our experiments with ordered comparisons in Chapters 3 and 4 move closer to encapsulating the truly perceived presence of several soft traits. In other areas of soft biometric identification, relative attributes have also been successfully applied to precisely describe the face [125] and clothing [26]. Alternatively, in fine-grained attribute studies, descriptive precision is increased by using a large number of sub-categories to differentiate visual appearances [126].

However, none of these approaches distinguish between attribute ambiguity (Figure 5.1a) or image obscurity (Figure 5.1b), crucial for ubiquitous identification techniques. Table 5.2 exemplifies the initial difficulty in mapping ambiguity, uncertainty and abstract concepts to traditionally constrained textual, nominal and one-dimensional ordering scales. For *gender*, the concepts of “Androgynous” and “Can’t see” cannot be described in a binary fashion, or are inaccurately conjoined on a relative ranking scale. In the example of *ethnicity*, we find ‘expertly defined’ police IC codes [28] are arbitrarily mapped to nominal values or are entirely inapplicable to one-dimensional ranking.

As such, solutions that require expertly-defined sparse or ordinal representations are unsuited to describing obscure, ambiguous or abstract subject matter captured in unconstrained, real-world image data. Therefore, a new annotation solution must not only facilitate highly discriminative labelling, but also allow for highly adaptive descriptions, in order to semantically discriminate images accurately in any situation.

5.3 Super-fine Attributes

Inspired by fine-grained and relative attributes discussed in Section 2.6, we introduce *super-fine attributes* as the next step towards true semantic image discrimination. Super-fine attributes aim to describe multiple, integral concepts of a single trait, represented as multi-dimensional coordinates in a perceptual space. These novel semantic spaces can describe both clear and subtle discriminations between images and enable vastly more powerful and intricate visual descriptions over coarse-grained categorical vocabularies or restrictive one-dimensional orderings.

In Figure 5.2 we hypothesise a visualisation of super-fine attribute representations, describing concepts as coordinates in a (two-dimensional) semantic space. While previous descriptive methods struggle to interpret abstract concepts, super-fine semantic spaces can capture complex inter-concept relations. For example, a new semantic space describing *gender* may relate the concepts of “Androgynous”, “Not sure” and “Can’t See” equidistant from “Male” and “Female”. However, “Not sure” and “Can’t See” might exhibit a larger distance than “Androgynous” from the clear binary gender classes. Likewise for *ethnicity*, broad “European”, “Black” and “Asian” groups are likely to be roughly equidistant, while sub-concepts would be separated by much smaller distances. By describing the *distance*, or *similarity* between concepts and subjects images, we can elucidate much more complex, higher-dimensional visual descriptions.

5.3.1 Metric vs. Non-metric

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 provide continuous distance measurements applicable to *metric* topological embeddings that seek to preserve inter-point distances e.g. *“how similar is a to b?”*.

A large corpus of work extends Non-metric Multidimensional Scaling (NMDS) [127] with triplet comparisons, including Generalized Non-metric Multidimensional Scaling [128], Crowd Kernel algorithm [129] and Stochastic Triplet Embedding [130]. These works find ordinal embeddings from a subset of pairwise dissimilarities, tackling the issues of collecting all $\mathcal{O}(n^3)$ ordered comparisons. In addition, Amid et al. consider ambiguity and multiple interpretations, proposing Multiview Triplet Embedding to find a number of low-dimensional maps corresponding to hidden attributes, discarding the notion of one correct solution [131].

However, perceptual dissimilarity measures and interpretations remain highly contentious. Agarwal et al. argue against metric embedding, stating that dissimilarity magnitudes are unreliable and difficult to measure [128]. On the other hand, Demiralp et al. comprehensively investigate 5 types of similarity judgement, reporting that while triplet matching exhibits lower variance, pairwise ratings are less costly and also possess greater granularity [132]. An early psychological study also compares alternative methods, suggesting pairwise comparisons are best for reducing respondent fatigue and maintaining a high level of information [133]. Furthermore, Fu et al. propose annotating subjective visual properties with pairwise comparisons and use an outlier detection method to prune inconsistent responses [75]. As super-fine attributes must concern inter-point distances and not merely order (Figure 5.2) we employ *continuous pairwise comparisons* with Metric Multidimensional Scaling (MDS) embeddings [127].

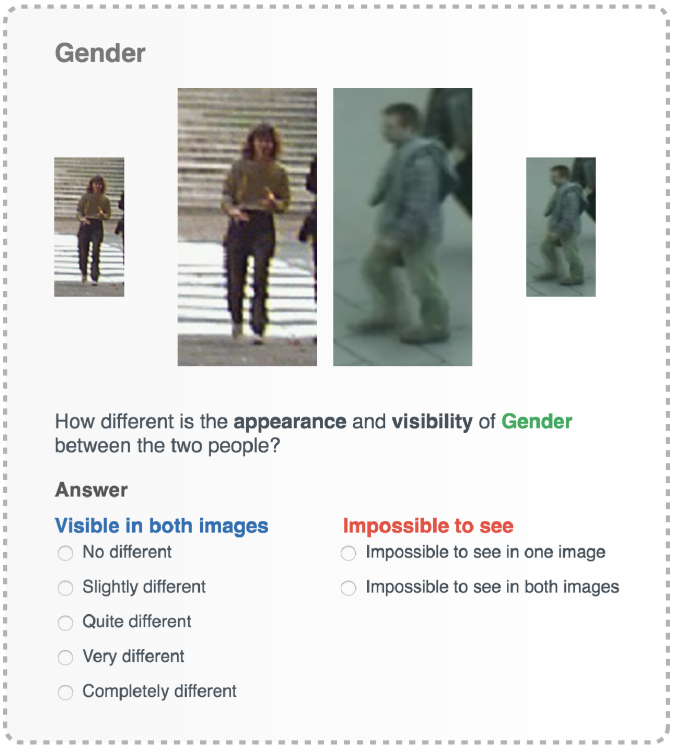
5.3.2 Gender with Similarity

Therefore, we propose to annotate gender-from-body by crowdsourcing *pairwise similarity comparisons* for the first time. Each pair of subject images is annotated by visually comparing the *perceived difference* in gender or its *invisibility*, thereby learning a consensus from the crowd. By collecting more open-ended annotations, our approach addresses the challenges of labelling hard-to-see, confusing and multi-concept attributes, further narrowing the semantic gap. We demonstrate our approach on two datasets; SoBiR and PETA [57], the largest and most diverse pedestrian re-identification dataset to date, annotated as 62.9% ‘male’ alongside 60 additional binary attributes.

Two studies also investigate similarity comparisons via crowdsourcing [79, 80], with several important distinctions. Firstly, both [79, 80] collect overall similarity annotations from a wide range of image subject matter, finding broad, basic-level categories. Instead, our approach discovers super fine-grained visual concepts within a specific trait of pedestrian images. Secondly, we deal with very low-quality and highly subjective images, necessitating the need to discern concepts of ambiguity and uncertainty, not previously dealt with. Finally, rather than grouping images [79] or matching a subset to a query image [80], we explicitly annotate each image pair, ensuring no pairwise comparison can be overlooked.

5.3.3 Similarity in Psychology

Similarity is a hotly debated topic in psychology, being argued as the composition of features [134], represented as a dynamic cognitive process [135] and modelled as geometric distances [127, 136, 137]. According to [124], gender predominantly means *“male-female difference”*, and *“in contemporary psychology is represented as a continuum of psychological difference”*. Though many visual cues are intrinsic to gender identity, it is almost



Gender

How different is the **appearance** and **visibility** of **Gender** 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

Figure 5.3: Example crowdsourcing task question.

always represented as a sole feature. Collectively, this suggests that while gender is hard to decompose, it can be discerned through *(dis)similarity*.

Lately, Edelman & Shahbazi argue 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 [82]. 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 [79, 80], we delve deeper into these trait-level subspaces, exploring their *integral* dimensions for the first time (likened to Amid et al.'s non-metric low-dimensional attribute maps [131]).

In computer vision, distance metrics are learnt to match images based on similarity and attribute simile classifiers have been show to outperform binary classifiers [69]. Image similarity comparisons have also been crowdsourced to discover basic-level categories from a continuous embedded similarity space [79, 80]. However, to our knowledge, no other similarity-based system deals with very low-quality or highly subjective images of homogeneous subject matter.

5.4 Crowdsourcing Similarity Comparisons

We design a crowdsourcing task on Crowdflower, similarly to Chapter 3. The task collects $\binom{N}{2}$ pairwise comparisons from both SoBiR, $N = 100$ and PETA, $N = 95$

datasets. As pairwise labelling is of $\mathcal{O}(n^2)$ space, we collect a subset of just 1% of PETA's 8709 total unique subjects. From this, we generate a representative visual taxonomy, to enable more refined categorical annotations of the remaining dataset. Furthermore, to ensure indistinct subjects are not overlooked, respondents are explicitly asked to judge every possible image pair, rather than grouping subsets [79] or matching to a query image [80].

Although similarity is commonly interpreted geometrically [127, 136, 137], it has been shown to be asymmetric if judging "subject A to subject B" [134]. In order to regularise responses, our questions instead judge "between the two subjects", and the task randomly shuffles the presentation order of images. Questions also judge *difference* over similarity, as it often defines gender [124] and is more succinct in describing subtle variations of an attribute.

The crowdsourcing task asks respondents to judge the difference in both *appearance* and *visibility* of gender, as in Figure 5.3. Images are displayed twice, once at their original resolution and a second scaled to fixed height for more direct side-by-side comparison. Answers are annotated on a 5-point Likert-type answer scale: "No different", "Slightly different", "Quite different", "Very different", "Completely different". Respondents may also answer "Impossible to see in one image / both images" to clearly state there are no visible cues, serving as a measure of uncertainty, and mitigating feigned responses. This enables our approach to differentiate between ambiguity (open to more than interpretation) and uncertainty (having imperfect or unknown information).

Crowdsourcing respondents are vetted by requiring at least 80% test question accuracy throughout the task. We present an initial quiz page of 10 test questions, with remaining pages containing 1 test question and 9 genuine questions. Test questions are carefully crafted to allow a range of acceptable responses, ensuring respondents understand the task, without overzealous priming. 'Gold-standard' responses are crowdsourced for 100 test questions in an initial annotation stage.

5.5 Topological Perceptual Interpretation

The crowdsourcing task provides us with raw pairwise comparison data, which we now need to interpret in order to form a global consensus of the crowd's perception. This involves combining disparate pairwise annotations into a single perceptual distance measure δ_{ij}, later enabling topological representations of the data.

Table 5.1 encodes dissimilarity annotations as pairwise proximities p_{ij} and uncertainties u_{ij} between subject images i and j, where $i, j \in 1, ..., N$. Similarity comparisons may then be mapped to geometric dissimilarity distances using an appropriate monotonic

Annotation	Interpretation	p_{ij}	u_{ij}
No different	Completely similar	0	0
Slightly different	Very similar	0.25	0
Quite different	Quite similar	0.5	0
Very different	Slightly similar	0.75	0
Completely different	Not similar	1	0
Impossible to see in one image	Not similar	1	1
Impossible to see in both images	Completely similar	0	1

Table 5.1: Encoding dissimilarity annotations to proximity p_{ij} and uncertainty u_{ij} measures between subjects i and j.

metric. We opt for exponential decay as a suitable distance measure following [136, 137]:

$$g(p_{ij}) = \exp(\lambda(1 - p_{ij})), \tag{5.1}$$

where λ is decay rate. Setting $\lambda \gg 1$ represents steps between similarities more evenly, approximating a linear function, while $\lambda \ll 1$ emphasises steps between larger differences, spacing distinct concepts further apart, but more compactly in the perceptual space.

Uncertainty measures, $0 \leq u'_i \leq 1$, are calculated per subject, as the fraction of all "Impossible to see.." u_{ij} annotations:

$$u'_i = \frac{\sum_{j \in N \wedge j \neq i} u_{ij}}{N - 1}. \tag{5.2}$$

Next, we define the pairwise uncertainty weighting function between u'_i and u'_j as follows:

$$v_{ij} = (|u'_i + u'_j|/2)^{\psi}, \tag{5.3}$$

where $0 < \psi \leq 1$ is the weighting eccentricity. Setting $\psi \approx 1$ represents ambiguity and uncertainty more equally, while $\psi \ll 1$ accentuates distances between subjects with different uncertainties.

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

$$\delta_{ij} = v_{ij}|u'_i - u'_j|g(0) + (1 - v_{ij})g(p_{ij}). \tag{5.4}$$

forming a symmetric, positive semidefinite distance matrix with zero diagonal $\mathbf{\Delta} = [\delta_{ij}] \in \mathbb{R}^{N \times N}$. We can now apply a number of unsupervised data exploration techniques to our bespoke spatial interpretation.

5.5.1 Multi-Dimensional Scaling Embedding (MDS)

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 $\mathbf{X} = [\boldsymbol{x_i}] \in \mathbb{R}^{N \times M}$, where $M \ll N$. 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 most masculine to most feminine pair, or most masculine to second most feminine 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 [127]. This considers minimising the normed Stress-1 function defined as:

$$\sigma_1(\mathbf{X}) = \frac{||d_{ij}(\mathbf{X}) - \delta_{ij}||_2}{||d_{ij}(\mathbf{X})||_2}. \tag{5.5}$$

where $d_{ij}(\mathbf{X})$ is the Euclidean distance between points $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ in $\mathbf{X}$:

$$d_{ij}(\mathbf{X}) = ||\boldsymbol{x_i} - \boldsymbol{x_j}||_2. \tag{5.6}$$

Each step increments k and computes the Guttman transform, updating $\mathbf{X}$ as follows:

$$\mathbf{X}^k \leftarrow \frac{\mathbf{B}(\mathbf{X}^{k-1})\mathbf{X}^{k-1}}{N}, \tag{5.7}$$

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

$$b_{ij} = \begin{cases} -\delta_{ij}/d_{ij}(\mathbf{X}), & \text{for } j \neq i \wedge d_{ij}(\mathbf{X}) \neq 0 \\ 0, & \text{for } j \neq i \wedge d_{ij}(\mathbf{X}) = 0 \end{cases} \tag{5.8}$$

$$b_{ii} = -\Sigma_{j \in N \wedge j \neq i} b_{ij} \tag{5.9}$$

The algorithm iterates until convergence at:

$$\sigma(\mathbf{X}^{k-1}) - \sigma(\mathbf{X}^k) < \epsilon, \tag{5.10}$$

selecting $\epsilon = 10^{-4}$. The values of $\boldsymbol{x_i}$ are randomly initialised uniformly between $[0, 1]$. As the outcome of MDS is heavily dependent on the starting configuration, we repeat the process 1000 times and select the solution with minimal $\sigma_r(\mathbf{X})$. Principle Component Analysis (PCA) is then applied to the solution to uniformly orient the most salient dimension to the x-axis, ensuring a consistent outcome between runs.

5.5.2 Agglomerative Hierarchical Clustering (AHC)

The second data exploration procedure explored is Agglomerative Hierarchical Clustering (AHC), which forms hierarchical groupings of mutually exclusive data subsets [138]. We use it to establish a visual taxonomy, formed as c sets of images, grouped by their perceived gender similarity. AHC is a 'bottom up' approach to clustering, whereby subject observations are iteratively merged up the hierarchy.

Each observation $\boldsymbol{x_i}$ starts in its own cluster C_i. At each iteration, the cluster pair with minimum linkage criteria D_{ij} are merged, until a maximum of c clusters remain. We apply AHC with the popular Ward's linkage criteria [138] which initialises cluster distances as the squared Euclidean distance between points:

$$D_{ij} = D(C_i, C_j) = ||\boldsymbol{x_i} - \boldsymbol{x_j}||_2{}^2. \tag{5.11}$$

On merging clusters C_i and C_j, the updated cluster distance $d_{ij(k)}$ between the new cluster $C_i \cup C_j$ and C_k is computed recursively as:

$$D_{ij(k)} = D(C_i \cup C_j, C_k) = \alpha_i D_{ik} + \alpha_j D_{jk} + \beta D_{ij}, \tag{5.12}$$

$$\alpha_l = \frac{|C_l| + |C_k|}{|C_i| + |C_j| + |C_k|}, \qquad \beta = \frac{-|C_k|}{|C_i| + |C_j| + |C_k|}, \tag{5.13}$$

where $|C_i|$ is the cardinality of cluster C_i. As such, when merging clusters, the within-cluster variance is minimised in relation to the sum of the between-cluster variance. Therefore, this method is suitable to be applied directly to the high-dimensional distance matrix $\boldsymbol{\Delta}$ as minimizing relative cluster variance accounts for potentially noisy dimensions.

5.6 Super-fine Attribute Characteristics

We perform three experiments to evaluate the characteristics of our gender similarity data and demonstrate its representational flexibility. In Section 5.6.1, we first investigate one-dimensional MDS ranking, replicating a relative attributes-based approach for comparison. In Section 5.6.2, we apply AHC to analyse the agreement between new found clusters and the original binary ground-truths of each dataset. Lastly in Section 5.6.3, we analyse the clusters for their consistency and present an example visual taxonomy of the PETA dataset. In each experiment, distance matrices $\boldsymbol{\Delta}$ are computed with $\lambda = 10$ and $\epsilon = 0.7$, found by a preliminary grid search of the most stable overall parameters across each experiment.

5.6.1 One-dimensional Ranking

We compare a one-dimensional representation of our gender similarity to SoBiR's relative-continuous gender labels and PETA's original binary ground-truth labels in Figure 5.4. One-dimensional embeddings of $\mathbf{X}$ are found by applying MDS with $M = 1$ to the distance matrix $\mathbf{\Delta}$, producing subject similarity *scores* and associated ranks. Dotted red lines indicate where a binary female-male split would normally occur around scores of 0.5.

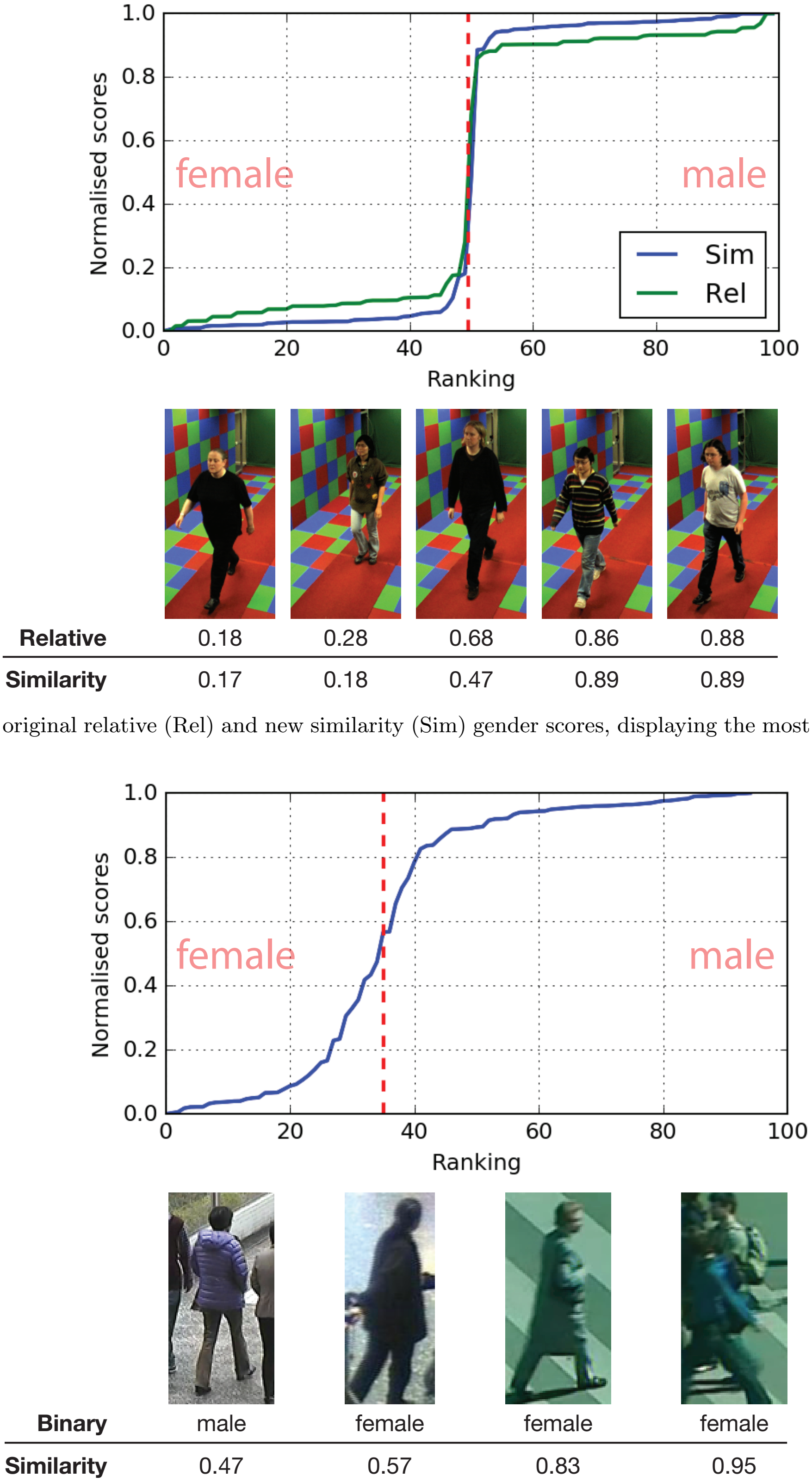
In Figure 5.4a, we observe highly analogous relative and similarity scores across SoBiR's 100 subjects. Although collected through two disparate forms of comparative visual annotation, scores vary on average only by -0.002 ± 0.047, with a Spearman's rank correlation coefficient of $\rho = 0.84$, $p < 10^{-26}$. Furthermore, the dataset's most gender-ambiguous subjects obtain identical ranks and very similar relative scores. This suggests our methodology is at least as informative as the ordered comparison approach in Chapter 3.

Figure 5.4b visualises the crowd's gender perception of a subset of 95 subjects from PETA. Lower quality and more obscure images produce a shallower, less divisive, slope. We find a proportionally similar female-male split to the binary ground truths, with the only four conflicting measures displayed. Evidently, the former two conflicts are highly subjective, attaining scores near 0.5, while the crowd's consensus is more determined for the later two conflicts, with scores > 0.8. These findings are later reiterated in a two-dimensional projection of gender similarities in Figure 6.6 (Section 6.3.5), where it can be seen that the later two subjects are again clustered distinctly as “Male”. Although originally labelled as “Female”, in light of our findings, we suggest that these original PETA classifications are erroneous.

5.6.2 Binary Ground-truth Clustering Agreement

In this experiment we investigate the agreement between the newly perceived gender similarities and the originally annotated binary ground-truths. To directly compare the two sets of data, we cluster the gender similarities into perceived categories via AHC. Agreement between the newly perceived categories and original binary labels is then measured as the Adjusted Mutual Information score (AMI), in the range $[0, 1]$. AMI quantifies the information shared by two partitions of mutually exclusive subsets, adjusted for the effect of chance. A score of 0 indicates purely independent (random) label assignments, while a score of 1 indicates two label assignments are equal.

To observe how AMI scores vary with different populations and numbers of categories, we generate unique distance matrices $\mathbf{\Delta\prime} \in \mathbb{R}^{n \times n}$ from a subset of $50 \leq n \leq N - 1$ subject similarities and apply AHC to form $2 \leq c \leq 7$ perceived categories. Each



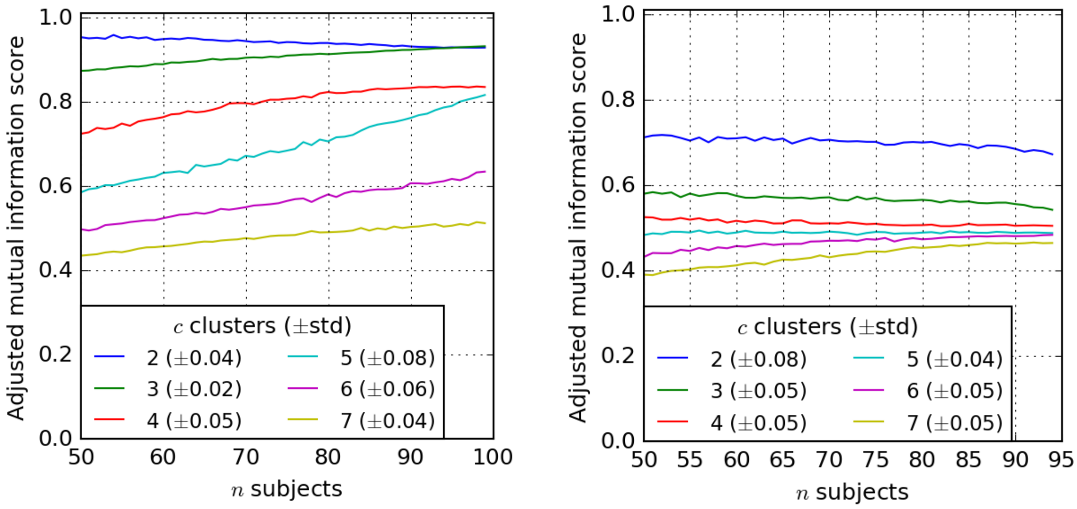
Relative	0.18	0.28	0.68	0.86	0.88
Similarity	0.17	0.18	0.47	0.89	0.89

(a) SoBiR original relative (Rel) and new similarity (Sim) gender scores, displaying the most ambiguous subjects.

Binary	male	female	female	female
Similarity	0.47	0.57	0.83	0.95

(b) PETA original binary and new similarity gender scores, displaying subjects with conflicting measures.

Figure 5.4: One-dimensional similarity ranking of SoBiR and PETA subjects using MDS. Dotted red line indicates female-male split.



(a) SoBiR.

(b) PETA.

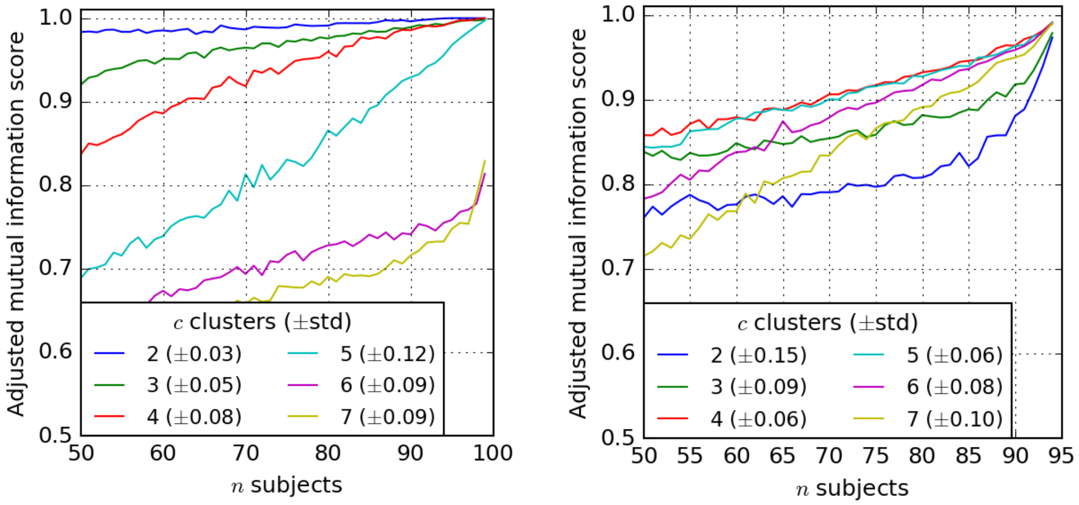
Figure 5.5: Agreement (AMI score) between original binary labels and newly perceived categories of n randomly sampled subjects. Each line records the correspondence between the original binary partitions to c categories generated by clustering a unique distance matrix $\boldsymbol{\Delta}'$ of n perceived subject similarities (averaged over 500 iterations).

random selection of n is repeated across 500 iterations, recording the average AMI score at each n between all c clusters of $\boldsymbol{\Delta}'$ and the original binary subject partitions, in Figure 5.5.

Clustering the similarity data from SoBiR's clear images with $c = 2$ perceived categories closely agrees with the original binary ground-truth labels, resulting in high AMI scores, as seen in Figure 5.5a. When selecting subsets of $n \approx N$ subjects we see AMI scores converging for $c = 2$ and $c = 3$ categories, suggesting an additional cluster may be describing some disparities between the two partitions. As expected, clustered similarity data from PETA agrees much less closely to the original annotations, even with $c = 2$ categories, Figure 5.5b. This indicates that increased image obscurity also increases the disparity between the pre-assigned textual labels and actual perceived similarity.

5.6.3 Clustering Consistency & Visualisation

This experiment aims to discover which number of categories provides the most consistent visual taxonomy from our similarity data. As in the previous experiment, we apply AHC to distance matrices $\boldsymbol{\Delta}'$ generated from n randomly selected subjects similarities. However, we now measure the AMI score clustering correspondence of n subjects from the original $\boldsymbol{\Delta} \in \mathbb{R}^{N \times N}$ and new subset $\boldsymbol{\Delta}' \in \mathbb{R}^{n \times n}$ distance matrices, for each set of $2 \leq c \leq 7$ categories. Figure 5.6 shows the partitioning agreement between the original and new c categories of n uniformly randomly sampled subjects. AMI scores are again averaged over 500 iterations per n.



(a) SoBiR.

(b) PETA.

Figure 5.6: Agreement (AMI score) measuring the agreement of perceived categories between all N subjects and n randomly sampled subject similarities. Clusters are drawn from the original $\boldsymbol{\Delta}$ distance matrix and clustering a new $\boldsymbol{\Delta}\prime$ distance matrix of the n subjects. Each line records the correspondence between the two partitions for c categories (averaged over 500 iterations).

For SoBiR, $c = 2$ categories are very consistent, with almost perfect agreement and low deviation, in Figure 5.6a. Clustering into $c = 3, 4$ categories also converges to AMI scores of 1 at $n \approx N$. Though less consistent, they may be desirable for their increased discrimination. For PETA however, we observe that $c = 2, 3$ categories are inadequate at describing the perceived gender similarities, being the least consistent overall, in Figure 5.6b. Instead, we find that $c = 4, 5$ perceived categories result in the most reproducible and discriminative partitions. Interestingly, AMI scores of $c = 6, 7$ start low, but radically increase with n, intersecting the AMI scores of $c = 2, 3$ which remain fairly constant until $n \approx N$. This indicates that the appropriate number of clusters is somewhat correlated to the number of subjects, and that including additional subject images with large variation is likely to introduce further perceived categories.

The experiment also exhibits a high degree of jitter on the recorded AMI scores across n. This is partially explained by the overall score standard deviations and periodic edge-case clustering conditions, where subjects flip-flop between one cluster and another, occurring due to the exact selection of n and c.

Figure 5.7 displays an example visual taxonomy of gender from PETA, computed by applying MDS to find an $M = 3$ dimensional embedding, subsequently applying AHC with $c = 5$ clusters. Visual groups largely match the original label concepts of ‘male’ and ‘female’ and our labelling of ‘uncertainty’. In this example we attach semantic language descriptions to each visual category. This partitioning only attains an AMI score of 0.45 to the original binary labels, yet on visual inspection, intra-group images are highly similar and clearly correspond to our language descriptions. This indicates that categories may be better defined visually, through related exemplar images. Interestingly, group 3

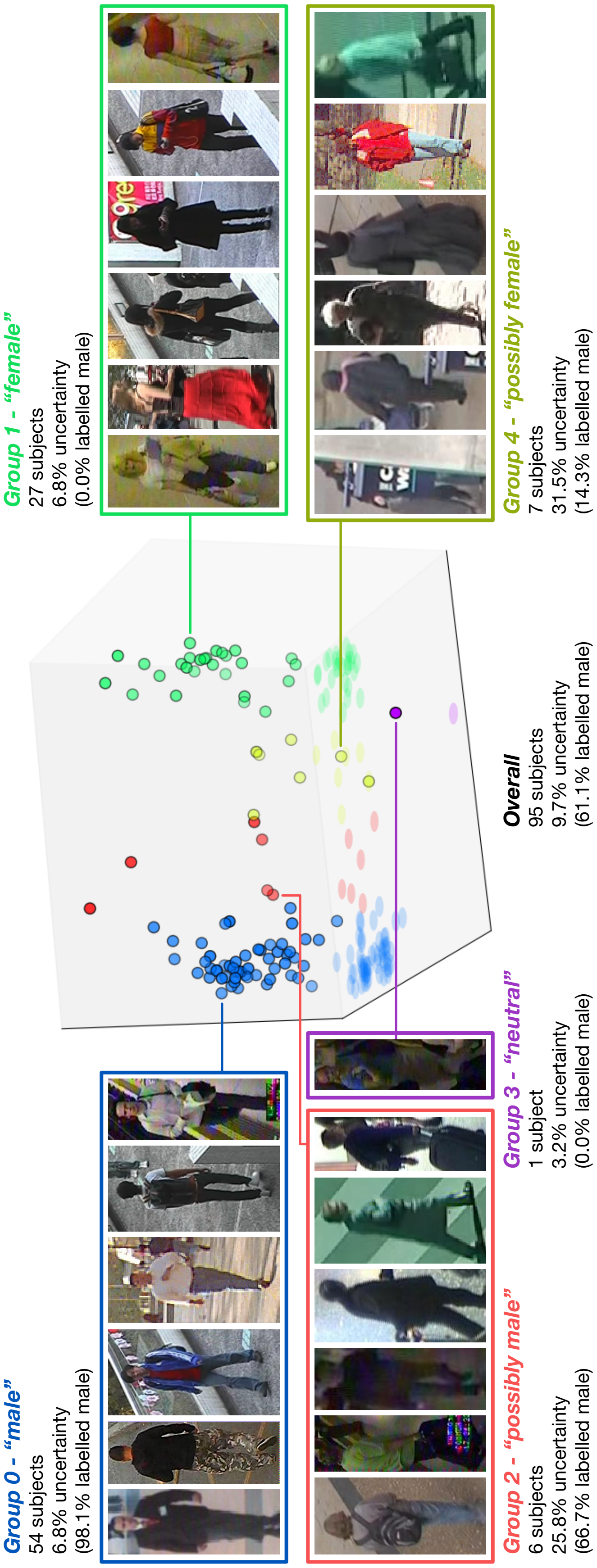


Figure 5.7: Example visual taxonomy of gender similarities on the PETA [57] dataset, formed with $c = 5$ clusters and visualised with 3-dimensional MDS embedding. Including group membership, average uncertainty and original binary ground-truth (brackets).

comprises just one subject, annotated confidently by respondents but contradictory to other subject images, forming its own 'neutral' group.

5.7 Conclusions

We have introduced super-fine attributes as an adaptive form of soft trait representation, enabling precise descriptions of gender-from-body in challenging surveillance imagery. We propose a crowdsourcing task to collect innovative similarity comparisons and discover super fine-grained visual taxonomies from subject images alone. This unique method of perceptual annotation is applicable to ambiguous, uncertain and multi-concept traits and easily extended to other image contexts, transitioning cutting-edge soft biometrics towards real-world datasets.

We experiment with two forms of dimensionality reduction, embedding and clustering, to interpret open-ended similarity annotations and provide a topological reasoning of the crowd's perceptual consensus. We demonstrate that pre-defined binary categorisation is insufficient in describing the subtleties of gender from the challenging PETA dataset. In fact, we advocate the inclusion of 'neutral' and 'uncertain' categories for studies involving human perception and demographic labelling, as in [30].

Lastly, we demonstrate that perceived gender concepts can be defined through related exemplar images, circumventing the need for restrictive, pre-defined textual lexicons. This also clearly distinguishes our work from traditional approaches that attempt to recreate biological ground-truth. In Chapter 6 we build upon this idea and annotate two additional soft traits, addressing issues of very large-scale annotation and investigate the potential of super-fine attributes for automatic identification.

Chapter 6

Super-fine Attributes with Crowd Prototyping

6.1 Introduction

We have seen how precise, relative attributes enhance retrieval and identification in Chapters 3 and 4, and how super-fine attributes provide objective gender descriptions from challenging images in Chapter 5. This chapter pools these ideas together, centring on providing a universal approach to automatic soft biometric identification, applicable ubiquitously, at scale and to highly unconstrained images captured in-the-wild.

We propose *crowd prototyping* to efficiently crowdsource super-fine labels. Crowd prototyping pre-discovers salient perceptual concepts, subsequently enabling rapid non-verbal, categorical annotation via prototype matching. We focus on re-annotating *gender*, *age* and *ethnicity* traits, which are the most commonly reported characteristics in policing [28], criminal record keeping [27] and identity science [4–6, 35] and are proven to be critical in suspect identification [1, 2]. Furthermore, we collect annotations for all 19000 image instances (8699 identities) of the PETA dataset, investigating instance- and subject-level labelling, recognition and retrieval.

Section 6.2 discusses the key approach considerations regarding the PETA dataset, large-scale annotation task and evaluation. Section 6.3 introduces the crowd prototyping methodology to systematically discover salient perceptual concepts, building upon the investigation in Chapter 5. Section 6.4 demonstrates efficient large-scale annotation of super-fine labels, comparing them to conventional alternatives. Section 6.5 details our deep learning recognition approach, employing the ResNet-152 CNN with transfer-learning. Section 6.6 then establishes the inference power of this model over previous approaches on the PETA dataset, and demonstrates the superior performance of super-fine labels in multi-shot and zero-shot identification. Section 6.7 concludes our findings.

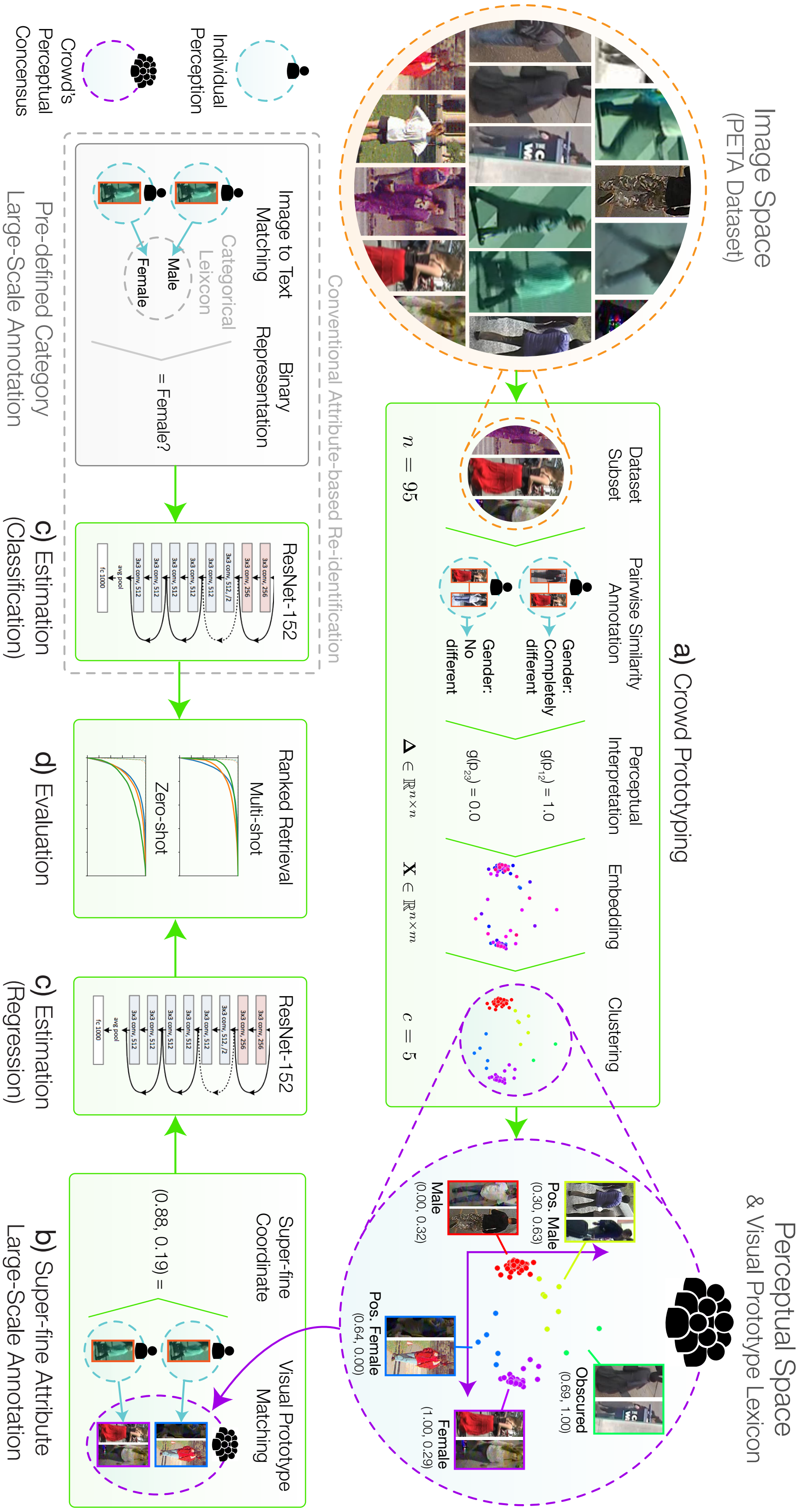


Figure 6.1: Approach overview, **Contributions highted in green**. **a)** Crowd prototyping previews the crowd's perception of an image subset to discover a perceptual space and salient visual prototypes (Section 6.3). **b)** Super-fine annotation efficiently matches unlabelled images to pre-discovered visual prototypes associated with attribute super-fine coordinate labels (Section 6.4). **c)** Image labels are estimated by fine-tuning the ResNet-152 CNN, classifying conventional binary attributes and regressing super-fine attributes (Section 6.5). **d)** Approaches are evaluated comparing ranked retrieval performance in multi-shot and zero-shot scenarios (Section 6.6).

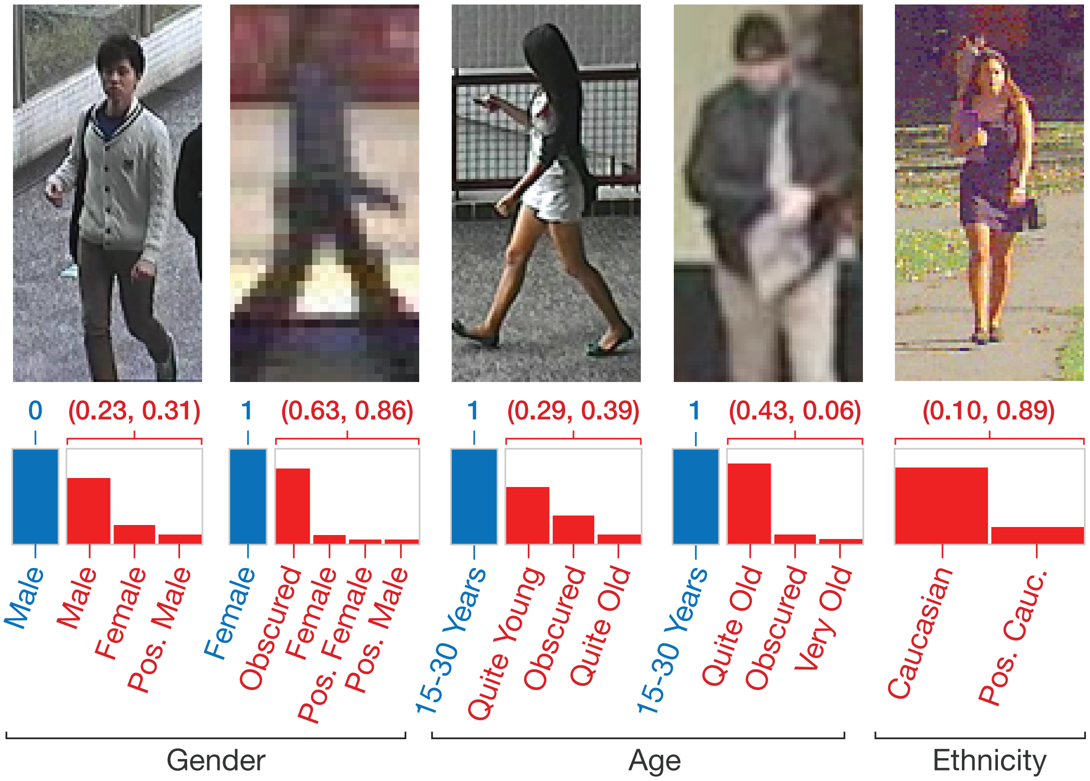


Figure 6.2: **Conventional ground-truth (blue)** vs. **Super-fine (red)** labels on the PETA dataset. Conventional categories are coarse-grained and can be inconsistent and/or irrelevant. In contrast, super-fine annotations improve relevance and objectivity, enabling more precise and accurate descriptions represented as multi-dimensional coordinates.

6.2 Approach Considerations

In this section we outline considerations regarding our approach, image dataset and performance evaluation, illustrated in detail in Figure 6.1 (where green shading indicates our novel contributions and grey shading is the conventional method).

6.2.1 PETA Dataset

The PETA dataset is 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 a number of works experimenting with the dataset [45, 50, 57, 94, 95] provide direct benchmark results. Each subject is annotated with 108 binary attributes, encapsulating global traits, worn accessories, carried objects and a vast number of clothing descriptors. However, the majority of these attributes are extremely imbalanced, occurring in under 10% of the data, and do not include *ethnicity* due to its controversial nature.

Importantly, like most benchmark attribute-based re-identification datasets, PETA assumes binary attribute representations. However, we have shown that such brittle, coarse-grained descriptions are ineffective at recounting subjective and uncertain properties, and inadequate for retrieval and identification in comparison to more adaptive fine-grained approaches. This is further exacerbated by the enormous inter- and intra-class variability of PETA's highly unconstrained images, often resulting in inconsistent or irrelevant binary annotations.

Figure 6.2 contrasts a number of cases where PETA's conventional labels (blue) lack the objectivity and discriminative power in comparison to new super-fine descriptions (red) that are later discovered. For example, the first image is of an effeminate male, originally labelled "Male" but described more precisely as mostly "Male" and somewhat "Female / Possibly Male" with discovered super-fine visual prototypes. The second example of a very low quality image, originally labelled as "Female", is now described more accurately and almost totally by the "Obscured" visual prototype. The third and fourth examples highlight more ambiguous *age* classifications, both originally labelled as "15-30" but distinguished primarily with the "Quite Young" and "Quite Old" visual prototypes respectively. The final example objectively describes *ethnicity* with super-fine attributes, which was previously contentious with expertly defined categories and impossible with a single relative attribute. In each example, the figure represents binary labels as a single class and binary digit, and super-fine attributes as a distribution across named visual prototype classes and as continuous coordinates.

6.2.2 Subject & Instance 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 [139]. 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:

$$\forall i \forall j \big[i = j \rightarrow \forall P (P_i \leftrightarrow P_j) \big] \tag{6.1}$$

This holds true in the case of PETA's original labels, 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.

In this chapter, we therefore also label image instances individually, estimating and evaluating both instance-level and subject-level image retrieval with super-fine attributes. As soft biometrics intertwine the notions of identity and appearance, instance- and subject-level labelling are linked to theories of *perdurance* (appearances may alter with time, therefore identities are temporally distinct) and *endurance* (descriptions of appearance must be consistent, therefore identities endure temporally) 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 both super-fine labelling methods significantly outperform conventional subject-level binary attributes in Section 6.6.

6.2.3 Label Fidelity

Leibniz's second principle, the 'Identity of Indiscernibles' states, if i and j have all the same P properties, then i is identical to j (reversing the implication of Equation 6.1):

$$\forall i \forall j \left[\forall P (P_i \leftrightarrow P_j) \rightarrow i = j \right] \tag{6.2}$$

This predicate does not hold true with PETA's current labels as only 7769 (89.3%) unique attribute configurations exist across all 8699 subjects. On selecting the 35 evaluation attributes, 218 (2.5%) 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 (*gender*, *age*, *hair colour*, *hair length*) only 128 total configurations exist, able to uniquely describe just 1.5% of the population. 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.

6.2.4 Label Representation

Traditional, one-size-fits-all lexicons often result in subjective labelling, suffering from annotator bias and anchoring effects [2]. This is especially apparent when relating textual descriptions to challenging surveillance imagery (Figure 6.2). Problems consistent with conventional labelling are also as a result of assuming the existence of one, 'gold-standard' ground-truth. 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 (Figure 6.1a-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. As such, this non-verbal communication process provides an improved approximation of the truly perceived consensus.

6.3 Crowd Prototyping

In this section we propose *crowd prototyping* as a novel methodology to gauge the crowd's perceptual consensus of each trait, discovering a number of discrete visual prototypes and their relationships. The process distils the perceptual structure of high-level traits and elicits a typical user's perception through crowdsourcing. This later facilitates efficient large-scale annotation of super-fine attributes.

Section 6.3.1 discusses the primary influences for crowd protoyping. Section 6.3.2 extends the pairwise similarity task in Section 5.4, additionally annotating *age* and *ethnicity* traits from the small set of $N = 95$ subject images. Section 6.3.3 then re-examines the perceptual interpretation method from Section 5.5, investigating three alternative interpretation strategies, which are evaluated alongside MDS embedding in Section 6.3.4. The optimal interpretation and embedding strategy pair are then carried forward to form a topological perceptual consensus for each trait. Lastly, Section 6.3.5 clusters each space with AHC, generating the final set of novel visual prototypes and associated perceptual coordinates for each trait.

Later, in Section 6.4, large-scale annotation tasks respondents to match new images to corresponding prototypes, efficiently representing attributes as coordinate labels. By previewing the crowd's perception of each trait from a small set of images in detail, we discover improved descriptive vocabularies accounting for the most visually prominent concepts. This ensures newly annotated labels pertain to the dataset, enhancing relevance, precision and accuracy over expertly-defined alternatives.

6.3.1 Crowdclustering & Prototype Theory

In crowdsourcing, the vast majority of works expertly-define categories, only collecting fixed labels from the crowd. In Gnomes et al.'s *crowdclustering* study, the aim is to discover categories and labels simultaneously through noisy crowd annotations, inferring worker preferences with a Baysian model [79]. Likewise, Wah et al. employ grid-based similarity comparisons alongside a human-in-the-loop classifier to find more powerful fine-grained categorisations [80]. The work especially highlights *"the weaknesses of attribute-based methods for fine-grained visual categorisation, and the need to eliminate expertly-defined vocabularies that constrain user terminology"*.

In cognitive science, Gärdenfors suggests that convex regions in conceptual spaces represent natural categories, which can be represented via the most central *"prototypical"* members [83]. Wittgenstein argues that categories are not connected by one common feature, but are instead resembled by a family of overlapping similarities [140]. Following these ideas, we cluster regions of each perceptual space to form multiple visual *prototypes*. 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 similar prototype. In fact, when scaling the ILSVRC challenge from 100 to 1000 ImageNet classes in 2010, Russakovsky et al. document the need to have prototype image examples representing each class in order to achieve any level of representative annotation at such a challenging scale [60].

Inspired by crowdclustering and prototype theory, we coin the term *crowd prototyping*, as a result of depicting crowdclusters with exemplar image prototypes. Our approach sequentially applies MDS and AHC to the pairwise similarity annotations, first finding salient conceptual dimensions and reducing annotation noise, then subsequently discovering a fixed number of perceptual categories describing each trait. In this way, our novel approach enables subject descriptions to simultaneously exist in a coarse, yet efficient, nominal categorical space (clusters), and in a multi-dimensional continuous coordinate space (embedding), facilitating ubiquitous labelling and improved automatic identification. In a similar manner, Husson et al. propose a methodology called Hierarchical Clustering on Principal Components (HCPC), initially applying PCA as de-noising technique prior to AHC, improving visualisation for exploratory data analysis [141].

6.3.2 Pairwise Similarity Task

Similarity comparisons for *gender*, *age* and *ethnicity* are collected in three separate tasks, performed identically to Section 5.4. Each task collects one response per subject pair for each trait, resulting in 13395 annotations from 614 respondents located across 62 countries. Figure 6.3 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 are found to shy away from extreme judgement for these two traits, indicating more indecision. This also highlights the necessity of a continuous scale in allowing subtle differentiation, in comparison to binary triplets-based approaches [79, 80].

Incidentally, each task consumed under $20 in crowdsourcing costs, yet provides a wealth of information. The next sections focus on discovering a perceptual consensus per trait from this data. Though a single perceptual consensus is intangible, we later show with image recognition that our following efficient approximations far outperform pre-defined alternatives for automatic identification.

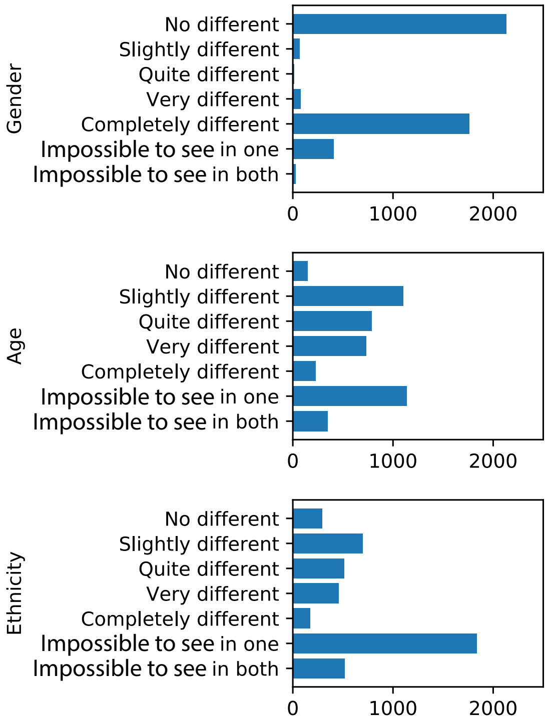


Figure 6.3: Response distributions.

6.3.3 Interpretation Strategies

Table 5.1 (Section 5.4) represents dissimilarity annotations as pairwise proximities p_{ij} and uncertainties u_{ij} between subject images i and j, where $i, j \in 1, ..., N$. In Section 5.5, we interpret similarity comparisons using exponential decay, proposed by both [136, 137] as a suitable geometric distance measure, with a constant λ. Meanwhile, [127] exemplifies a linear interpretation of the Likert Scale and attribute discovery works only deal with binary interpretations [79, 80]. In light of this, we now investigate three forms of proximity mapping, visualised in Figure 6.4 and as follows:

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

$$g^0(p_{ij}) = p_{ij} \tag{6.3}$$

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

$$g^1(p_{ij}) = \begin{cases} 0.00, & \text{if } p_{ij} \leq 0.25 \\ 0.50, & \text{if } p_{ij} = 0.50 \\ 1.00, & \text{if } p_{ij} \geq 0.75 \end{cases} \tag{6.4}$$

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

$$g^2(p_{ij}) = \frac{\exp(\lambda(1 - p_{ij})) - e^{\lambda}}{1 - e^{\lambda}} \tag{6.5}$$

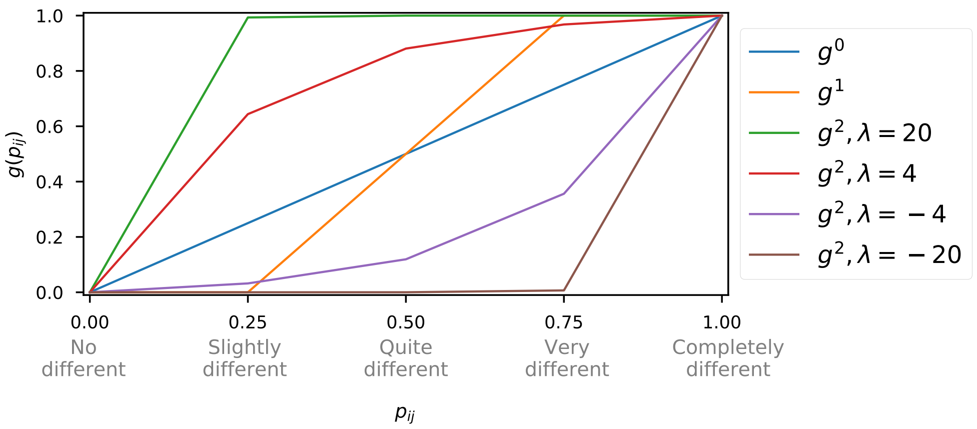


Figure 6.4: Dissimilarity proximity interpretations.

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 $\lambda = -20$ the mapping is equivalent to a binary interpretation, taking only "Completely different" to mean "different", or only "No different" to mean "same" at $\lambda = 20$.

As in Section 5.5, we describe pairwise dissimilarities under one measure δ_{ij}, assimilating proximity and uncertainty annotations. Uncertainty measures per subject u'_i are derived as in Equation 5.2. As we now experiment with many more proximity interpretations, we calculate v_{ij} as the average pairwise uncertainty, removing hyper-parameter ψ from Equation 5.3:

$$v_{ij} = |u'_i + u'_j|/2 \tag{6.6}$$

We also update the definition of δ_{ij} from Equation 5.4, incorporating the normalised $g(p_{ij})$ proximities with absolute pairwise uncertainty difference $|u'_i - u'_j|$, which are sum weighted by average uncertainties v_{ij}:

$$\delta_{ij} = (1 - v_{ij})g(p_{ij}) + v_{ij}|u'_i - u'_j|, \tag{6.7}$$

forming the symmetric, positive semidefinite distance matrix with zero diagonal $\mathbf{\Delta} = [\delta_{ij}] \in \mathbb{R}^{N \times N}$. In the next section, we evaluate the optimal parameter choice for our bespoke distance interpretation. Empirically, we find this weighted combination of uncertainty and dissimilarity produces the most stable and coherent embeddings. This is in contrast to alternative approaches that include interpreting uncertainty as a fixed distance, or ignoring pairwise constraints for uncertain annotations, that result in slightly poorer metrics than the ones reported in the next section.

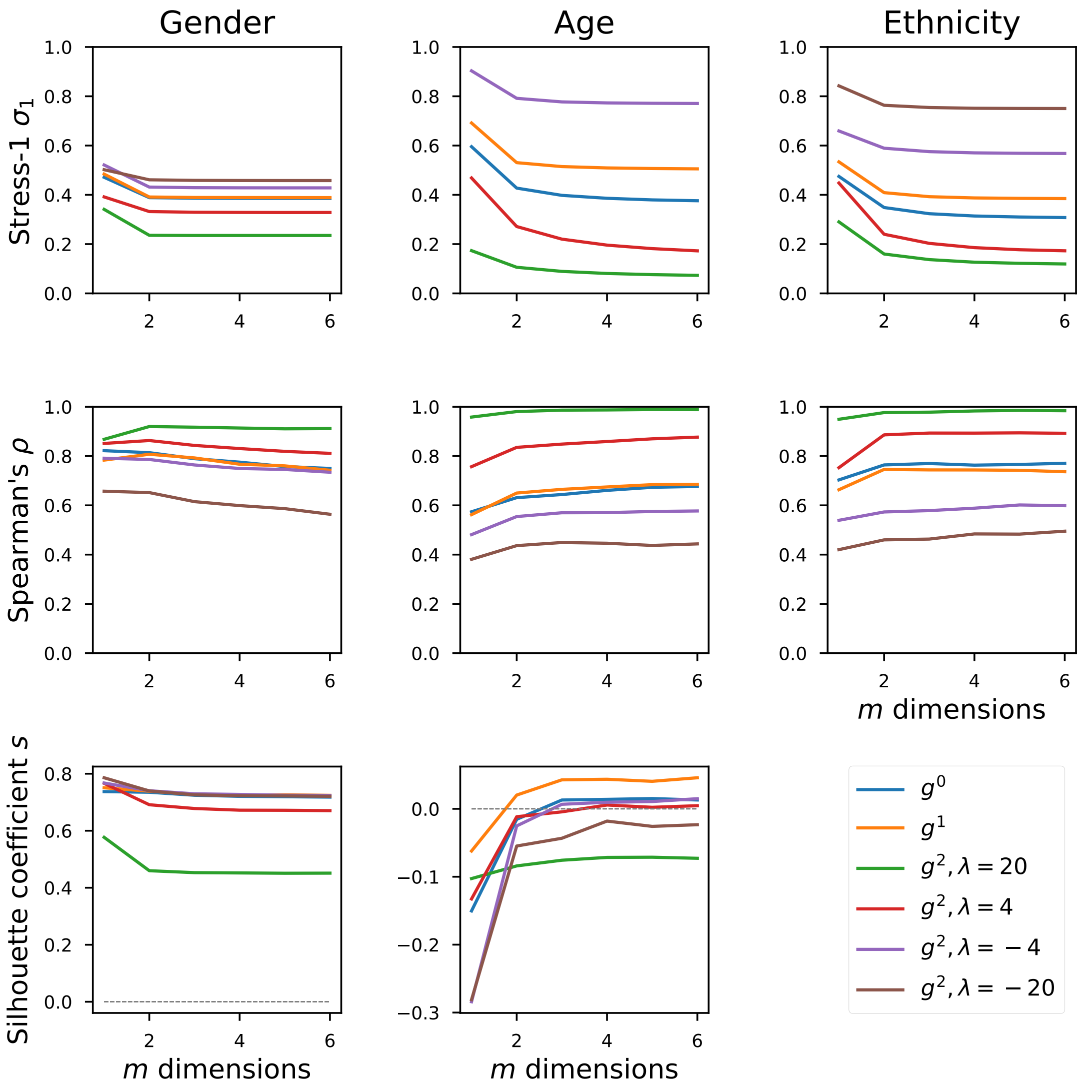


Figure 6.5: MDS embedding strategy evaluation of g^0, g^1 and g^2, measuring Stress-1 σ_1, Spearman's rank correlation coefficient ρ and Silhouette coefficient s as M varies.

6.3.4 Interpretation Strategy & Embedding Evaluation

To find the optimal interpretation strategy, we perform MDS embedding with $M \in 1, ..., 6$ dimensions and generate distance matrices $\boldsymbol{\Delta}$ with all proximity mappings g^0, g^1, g^2 and $\lambda \in \{-20, -4, 4, 20\}$. Alongside evaluating normed Stress-1 σ_1 values, we also compare Spearman's rank correlation coefficient ρ between δ_{ij} and d_{ij} (from Equation 5.6).

In the quest to discover a new ground-truth from wholly subjective data and without a gold-standard to emulate, these measures alone can be misleading. To better realise how informative our new embeddings are, we inspect their cohesion to PETA's original categorical labels. Our aim is not to faithfully reconstruct the original categories, but use them only as an indicator of pertinence for *gender* and *age*. The silhouette score coefficient $-1 \leq s \leq 1$ measures the consistency of each categorical cluster projected into embedded space, where a is the mean intra-cluster ℓ_1 distance and b the mean nearest, non-member cluster ℓ_1 distance over all samples:

$$s = \begin{cases} 1 - a/b, & \text{if } a < b \\ 0, & \text{if } a = b \\ b/a - 1, & \text{if } a > b \end{cases} \tag{6.8}$$

Results are presented in Figure 6.5. As M is incremented, σ_1 and ρ values consistently decrease and increase, improving respectively. We select $M = 2$, as an elbow occurs across all three traits, with minimal stress and correlation improvement thereafter. Proximity mapping $g^2, \lambda = 20$ seems to produce the most consistent embeddings when evaluating σ_1 and ρ, unlike its counterpart $\lambda = -20$. However, its significantly low silhouette coefficients indicate a poorer interpretation of dissimilarity annotations for *gender* and *age*. Interestingly, using $\lambda = 20$ to separate only totally dissimilar subjects, produces much higher silhouette scores. These results show that an increased granularity, especially for more similar subject annotations is preferential.

Embeddings of *age* are much less related to the original categories than those for *gender*, due to increased perceptual difficulty. In fact for *age*, only four proximity mappings ever produce $s > 0$; g^0, g^1 and $g^2, \lambda \in \{-4, 4\}$. Of these, we select g^1 for its best likeness to the original *age* categories and positive s value at $M = 2$.

This analysis highlights the difficulty in creating any ground-truth data and evaluating its objectiveness, without a pre-defined gold-standard target. However, our findings demonstrate that conceptual embeddings benefit from interval scale annotations and question the overall perspicacity of binary alternatives emulated with $\lambda = \{-20, 20\}$.

6.3.5 Perceptual Space Clustering

The final step of the crowd prototyping process is to draw visual prototypes for each perceptual trait space via clustering. In Section 5.5.2 we clustered the high-dimensional distance matrix $\mathbf{\Delta}$ using Ward's linkage criteria to account for noisy measurements. We now cluster the embedded space $\mathbf{X}$ with the average linkage criteria, as it produces more predicable clusters when applied to fewer data points in lower-dimensions, often being employed to generate phylogenetic trees, analogous to our application.

Each perceptual space observation $\boldsymbol{x_i}$ starts in its own cluster C_i and the average linkage criteria distances D_{ij} are initialised with the ℓ_1 distance measure:

$$D_{ij} = D(C_i, C_j) = ||\boldsymbol{x_i} - \boldsymbol{x_j}||_1. \tag{6.9}$$

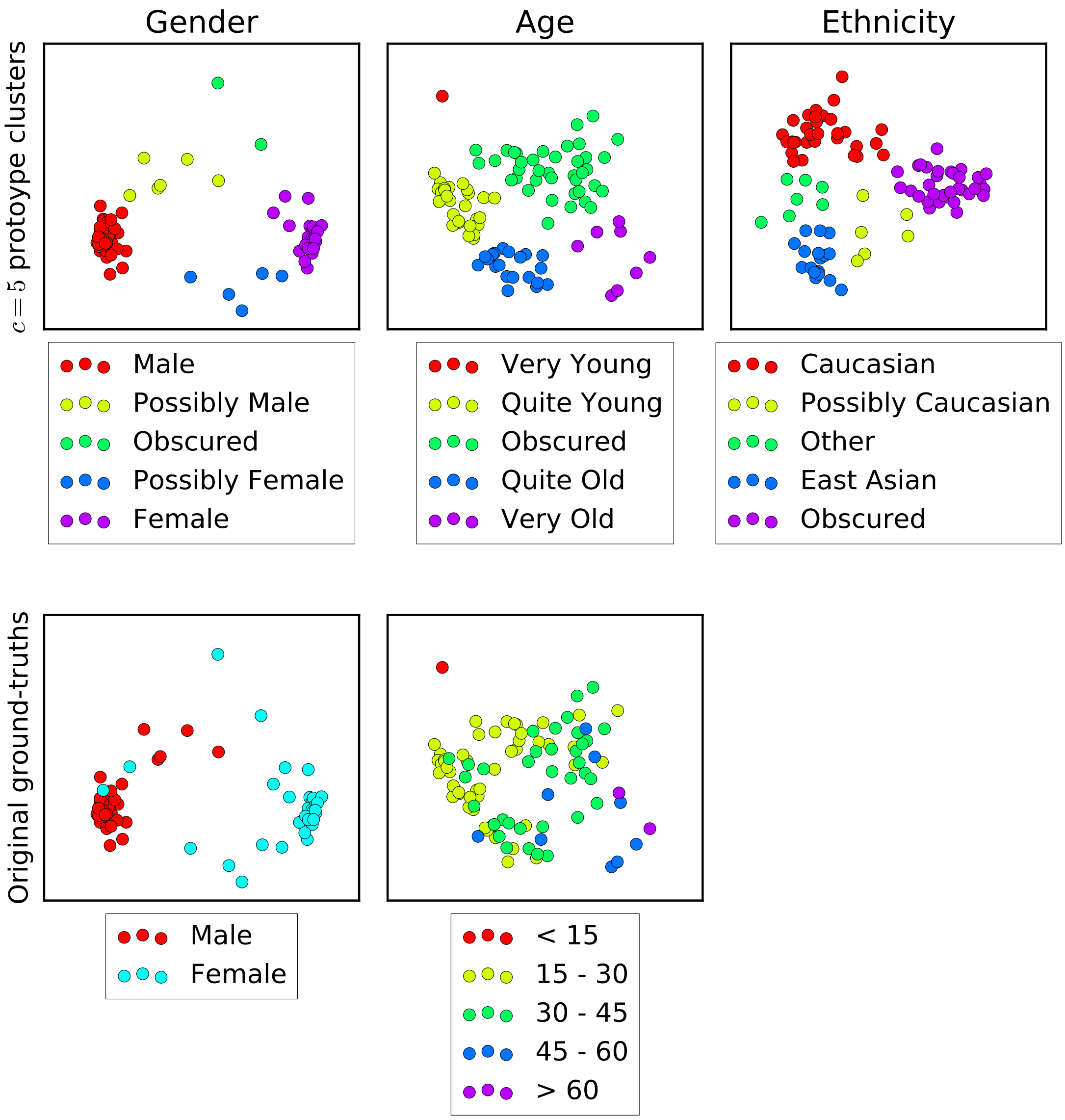


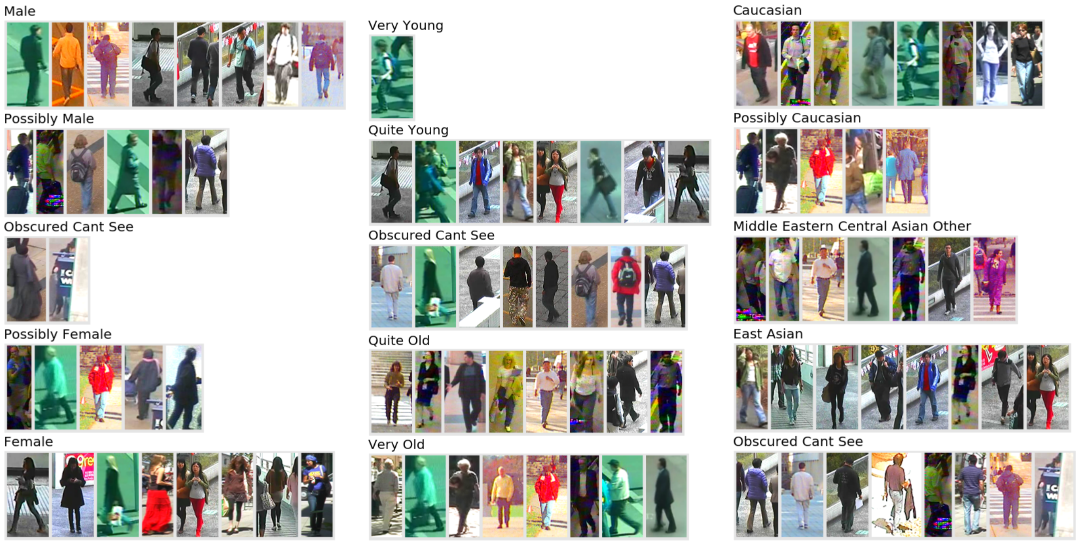
Figure 6.6: Comparison of new prototype clusters against original ground-truth categories projected in embedded conceptual space.

At each clustering iteration, linkage criteria distances are updated between joined clusters C_i and C_j and new cluster C_k, proportionally averaging the D_{ik} and D_{jk} distances:

$$D_{ij(k)} = D(C_i \cup C_j, C_k) = \frac{|C_i|D_{ik} + |C_j|D_{jk}}{|C_i| + |C_j|} \tag{6.10}$$

where $|C_i|$ is the cluster cardinality. 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. New clusters and



(a) Gender.

(b) Age.

(c) Ethnicity.

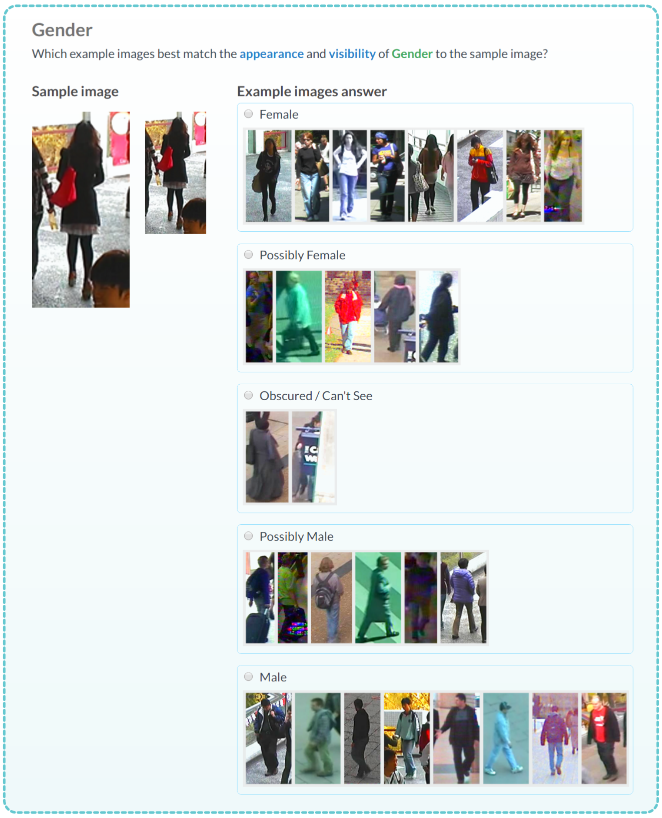
Figure 6.7: Discovered crowd prototypes, depicting distinct trait concepts from the crowd's perceptual consensus. Prototypes are formed by clustering conceptual spaces, selecting up to 8 images from each region's centroid. Semantic text descriptions are added manually for discussion.

original categories are compared in Figure 6.6, projected in embedded perceptual space. The final three sets of prototypes are shown in Figure 6.7.

Crowd prototyping *gender* discovers two clear gender classes, with the remaining three prototypes resembling varying levels of uncertainty (Figure 6.6 gender upper). In fact, clustering almost perfectly divides the original categories (Figure 6.6 gender lower), apart from two images we suggest are previously mislabelled "Female", as seen our earlier one-dimensional ranking in Figure 5.4b (Section 5.6.1). Interestingly the "Possibly Male" cluster is embedded closer to the "Can't See" cluster than the "Possibly Female" cluster. Although this seems incorrect, the embedding strategy analysis does not indicate any improvement when increasing M dimensions, suggesting that less clear females are more distinct from very unclear images than less clear males.

Prototyping *age* finds four clear age characteristics and one central obscured image group (Figure 6.6 age upper and Figure 6.7b). Unlike *gender*, the original categories are much more dispersed in the *age* embedding (Figure 6.6 age lower) reflected in the lower overall silhouette scores (Figure 6.5). However, the "Very young" and "$<$ 15" classes are extracted identically.

Prototyping *ethnicity* finds clearly distinguished "Caucasian" and "East Asian" ethnicities to dominate the dataset (Figure 6.6 ethnicity). Overall, crowd prototyping discovers certain concepts that are more specific than others e.g. "Caucasian" vs. "Middle Eastern / Central Asian / Other" (Figure 6.7c), representing differences in variance and prevalence of conceptual characteristics throughout the dataset.



Gender

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

Figure 6.8: Example large-scale visual prototype matching task question.

The strength of crowd prototyping is its ability to represent traits at varying levels of granularity, from high-detail conceptual maps to coarse-grained conceptual regions. This enables labels to be collected at any level of precision, constrained only by the cost of annotation. As prototypes are mapped to conceptual space coordinates, label conflicts can be treated as a source of additional information. This is in contrast to pre-defined categories or relative comparisons, where the precision is fixed and conflicting annotations are seen as a source of error to be avoided.

6.4 Large-Scale Annotation

This section discusses 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.

6.4.1 Visual Prototype Matching Task

The prototype matching task asks respondents to match new images to the most visually similar prototype. Each question displays one original-sized query image alongside the

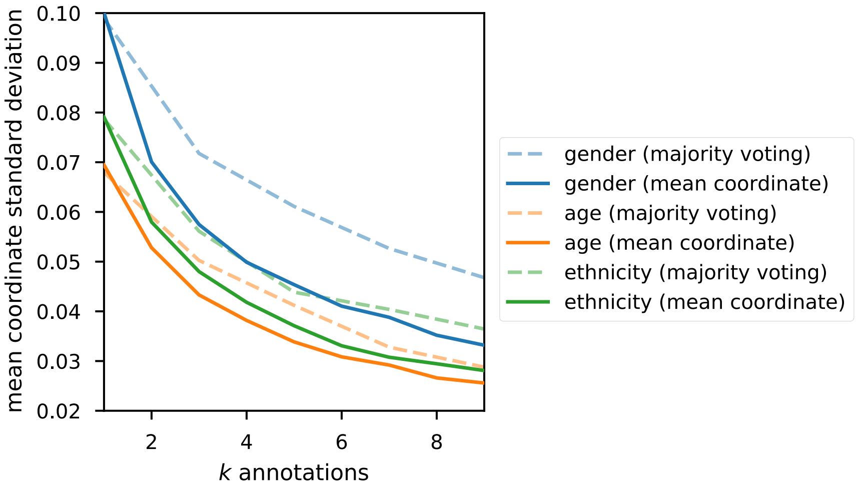


Figure 6.9: Mean coordinate standard deviations across k annotations.

images from all 5 prototypes of one trait, illustrated in Figure 6.8. Visual annotation enables rapid and intuitive categorical image labelling, requiring minimal effort and expense. The task is run similarly to previous pairwise crowdsourcing tasks, where test questions are drawn from 100 initial responses and respondents are constantly monitored for test question accuracy.

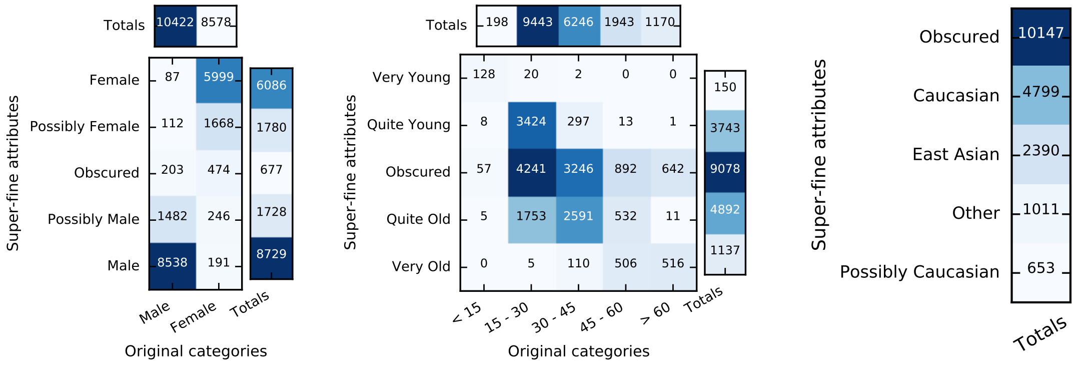
Matching images to visual categories is found to be more objective (and enjoyable) than annotating textual categories. However, it is still expected that repeat image annotations will elicit a number of different responses. This is entirely valid, and is the primary reason for associating visual prototypes 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 three annotations per instance.

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 instance of that subject. As the original subject-level labelling methodologies for VIPeR [44] and PETA [57] are not described, we assume limiting labels to just one annotation per instance to be a fair comparison.

6.4.2 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.



(a) Gender. (b) Age. (c) Ethnicity.

Figure 6.10: Label distributions and confusion between subject-level super-fine attributes and original binary categories.

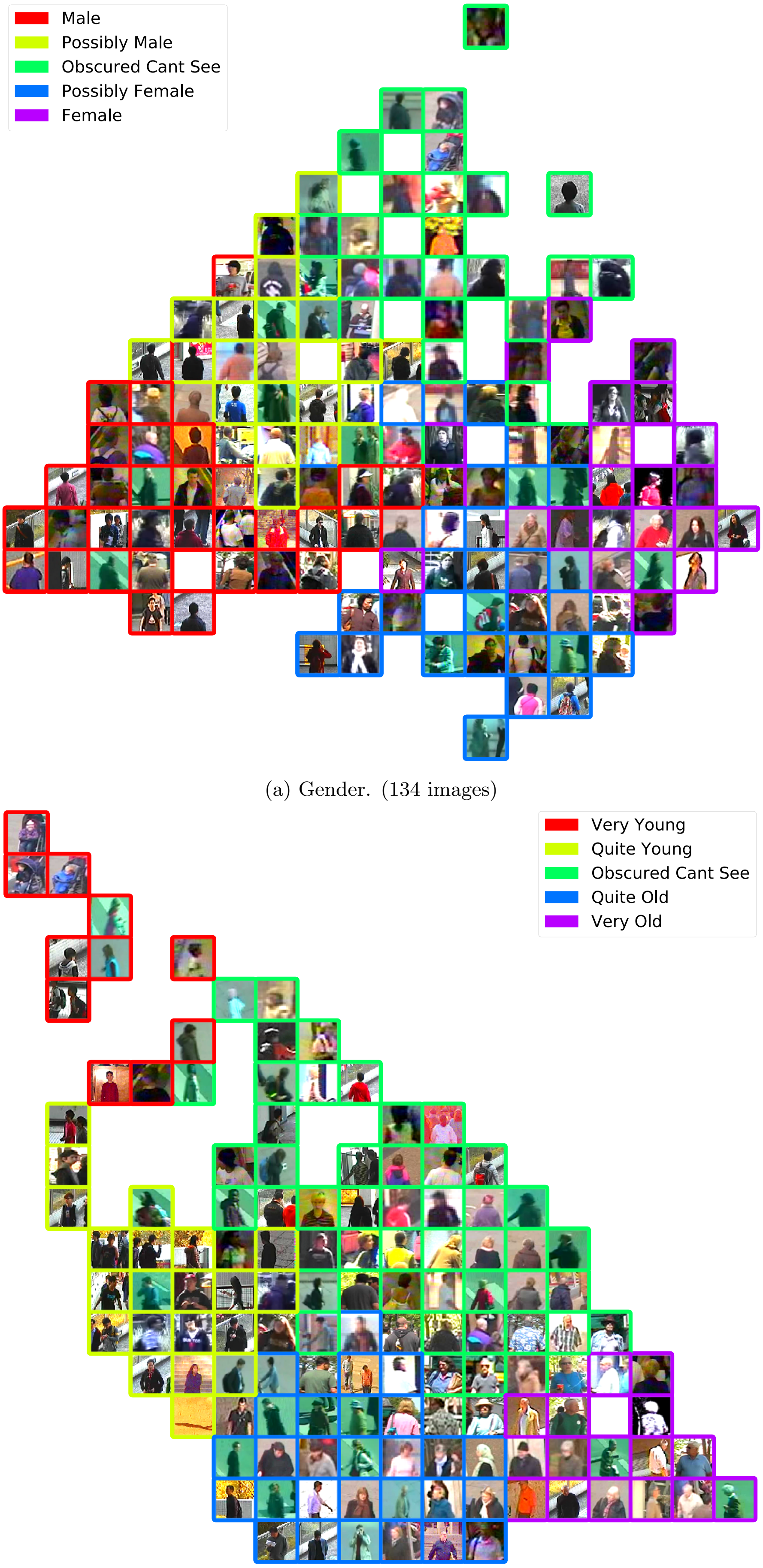
Fig 6.9 reports the mean standard deviation of instance-level simulated coordinates with both majority voting and mean calculation schemes at each k. Mean coordinate calculations are clearly more consistent than the majority voting scheme. We also find subject-level labels to perform similarly, but with 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 practical cost.

6.4.3 Super-fine Annotation Results vs. Binary Labels

Over 1600 respondents contributed to the annotation tasks, annotating just under 190000 total attributes and costing under $250 per trait. Tasks were also highly rated by respondents, averaging 4.0/5.0, for ease, fairness and pay.

Figure 6.10a-c (r.h.s. totals) 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 the face, whereas *gender* has many more visual cues to infer from, as can be seen in Figure 6.11. In fact, a high number of obscure annotations may even benefit recognition by training estimators to better detect image clarity, discussed in Section 6.6.3.

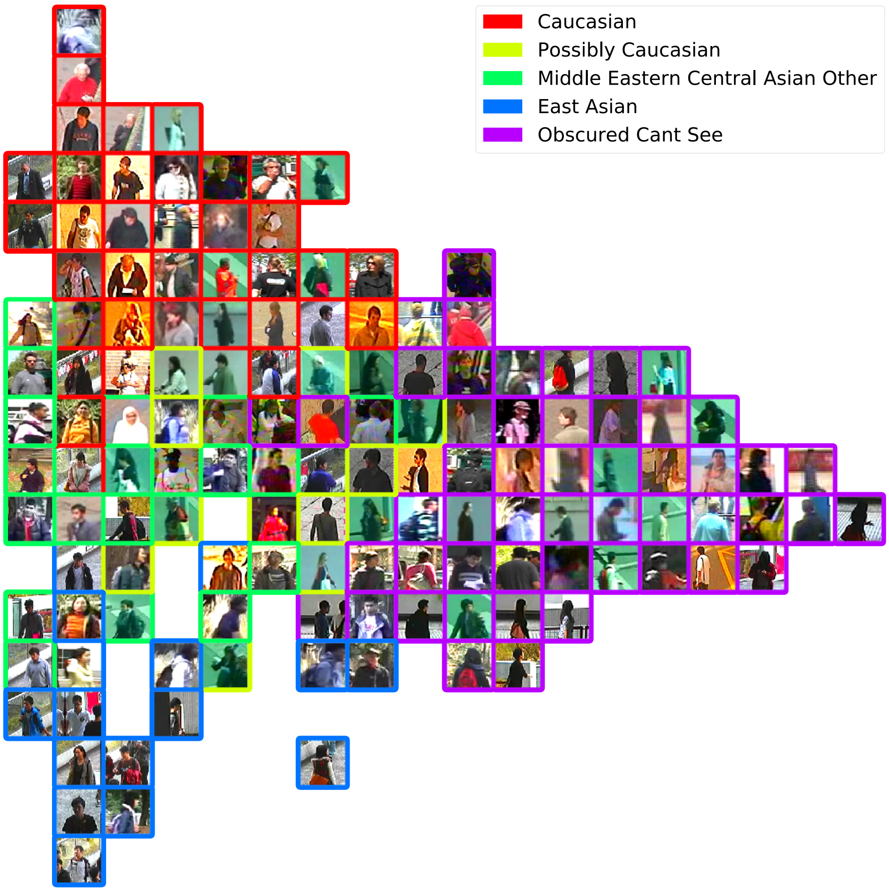
Figure 6.10 also shows the co-occurrence between super-fine and binary labels at subject-level. For *gender*, an exponential decay can be seen for the confusion of male and female labels, with only a small fraction of complete male-female reversals. This affirms the quality of our new labels in replicating previous schemes, while also aiding precision.



(a) Gender. (134 images)

(b) Age. (132 images)

Figure 6.11: (Continues on next page.)



(c) Ethnicity. (140 images)

Figure 6.11: 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.

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 median 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 an enormous step in discriminative power compared to conventional approaches. In the next sections, we investigate just how much more effective these labels are for identification, once automatically estimated.

Figure 6.11 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”, “Quite Old” and finally “Very Old” in Figure 6.11b. Qualitatively, this demonstrates the rational behind averaging multiple coordinate annotations to generate image labels in a continuous space.

6.5 ResNet-152 Attribute Recognition

For automatic label estimation, we select the ResNet-152 CNN model [142], which won the 2015 ImageNet [87] detection, ImageNet localisation, COCO [61] 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 due to the vanishing gradient problem [142].

Our ResNet model is pretrained on the ILSVR2012 challenge, employing ‘feature representation transfer’ as a form of unsupervised transfer learning [55]. This means that already extremely well generalised low-level feature descriptors are used to initialise our network, prior to task specific training. This provides a significant advantage over previous PETA attribute recognition works that do not employ transfer learning techniques (e.g. MAResNet [50]). Though PETA is a large and highly diverse surveillance dataset, the ILSVR2012 dataset is almost 100 times larger, producing more highly generalised feature descriptors that cannot be learnt from PETA alone.

The model accepts input image sizes of 224×224. We therefore directly scale images to fit, regardless of their original dimensions. With approximately 10 times more instances and 90 times more subjects than SoBiR, PETA contains much greater class variance. Therefore images are only augmented with an equal chance of mirroring around the vertical axis at training time. This skips the computationally expensive operations outlined in our augmentation strategy in Section 4.8.1, avoiding prohibitively long training times, while still resulting in state-of-the-art recognition accuracy.

The penultimate layer of ResNet pools 2048 features, which are fully connected to the final Q binary and $Q \cdot M$ super-fine output logits. For fine-tuning the model, we employ the Adam optimiser [143] 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 and label representations are used for binary classification and super-fine regression:

Binary classification. The original high-level PETA traits are encoded into one-hot binary form. Learning multiple traits at once is modelled as a multilabel classification problem, in which each binary class is independent and not mutually exclusive. We employ Sigmoid Cross Entropy (SCE) to determine the loss per subject i across all attributes $q \in 1, ..., Q$:

$$\begin{aligned} L_{SCE}(\boldsymbol{y_i}, \boldsymbol{\hat{y}_i}) &= \frac{1}{Q} \sum_{q \in Q} \Big(-\hat{y}_{iq} \log \big(\frac{1}{1+\exp(-y_{iq})} \big) - (1-\hat{y}_{iq}) \log \big(\frac{\exp(-y_{iq})}{1+\exp(-y_{iq})} \big) \Big) \\ &= \frac{1}{Q} \sum_{q \in Q} \Big(\hat{y}_{iq} - \hat{y}_{iq} y_{iq} + \log \big(1 + \exp(\hat{y}_{iq}) \big) \Big) \end{aligned} \tag{6.11}$$

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 measure the loss per subject i across all super-fine attributes $q \in 1, ..., Q$ and embedded dimensions $m \in 1, ..., M$:

$$L_{MSE}(\boldsymbol{y_i}, \boldsymbol{\hat{y}_i}) = \frac{1}{Q \cdot M} \sum_{q \in Q} \sum_{m \in M} (y_{iqm} - \hat{y}_{iqm})^2. \tag{6.12}$$

Our configuration of Sigmoid Cross Entropy loss and Adam optimisation is now a popular choice for training very deep, multi-label binary classifications networks, and can be found in other pedestrian attribute recognition works [45, 50]. These studies also incorporate an attribute weighting component to adjust for the affect of unbalanced label distributions at training time. We omitted this adjustment, as it was empirically found to result in lower overall accuracy for more balanced, and therefore more discriminative attributes. Furthermore, in Section 4.8.3 we propose a post-training attribute weighting scheme to successfully optimise independently learnt attributes, but which showed little use for joint classification. Therefore, we do not investigate attribute weighting with our three jointly learnt super-attributes, instead focussing on extensive comparative evaluation under different scenarios. The ResNet-152 model is implemented in Python using the Tensorflow library and run on a GPU using CUDA and CuDNN.

6.6 Soft Biometric Recognition & Identification

We present three experiments investigating our estimation model and super-fine label performance. Section 6.6.1 introduces the recognition and retrieval methodology. Section 6.6.2 first benchmarks our model against previous works for binary attribute recognition. Sections 6.6.3 and 6.6.4 then compare binary and super-fine label retrieval performance in multi-shot and zero-shot identification scenarios.

6.6.1 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. In Section 6.6.2, we draw direct comparison with previous works on PETA, evaluating binary estimation accuracy with an identical methodology. In Sections 6.6.3 and 6.6.4 we evaluate super-fine and binary labels with ranked retrieval, emulating suspect search given an eyewitness description. As binary and super-fine approaches are either classified or regressed, comparing accuracy alone is not sufficient. Instead, ranked retrieval measures the true discriminative power of each subject identification approach, incorporating both attribute precision and recognition accuracy.

Set-split criteria. To fully contrast performance and highlight the challenges involved, we experiment with two set-split criteria, original *multi-shot* in Sections 6.6.2 and 6.6.3 and *zero-shot* in Section 6.6.4 (as defined in Section 4.9).

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% training, 10% validation and 40% testing. We call this case *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 [46]. Therefore, we also experiment with splitting the dataset at subject-level as in *zero-shot identification* [37]. 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. With the increased fidelity of super-fine attributes, ground-truth descriptions are assigned either independently (instance-level), or uniformly across the same subject (subject-level) as discussed in Section 6.4.1. The majority of attribute-based re-identification works perform subject-level labelling, assuming invariance to pose, viewpoint and environment. In practice, we find this may be sub-optimal, especially with the precision of super-fine attributes, as significant variation is possible between instances of a subject. We therefore evaluate both instance- and subject-level labelling.

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 Receiver Operating Characteristics (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 ROCs for instance- and subject-level labelling and retrieval synonymously.

Training split criteria		Instance-level (multi-shot)						Subject-level (zero-shot)
		MRFr2 (S)	MRFr2 + DDN (S)	MLCNN (J)	DeepMAR (J)	MAResNET (J)	ResNet-152 (J)	ResNet-152 (J)
Attribute	ratio	[57]	[94]	[95]	[45]	[50]	**Ours**	Ours
Gender Male	0.55	81.70	86.50	84.34	89.90	76.60	**93.06±0.33**	83.96±0.56
Age 15-30	0.50	83.80	86.80	81.05	85.80	78.38	**87.28±0.48**	74.39±1.24
Age 30-45	0.33	78.80	83.10	79.87	81.80	75.55	**82.98±0.47**	60.60±3.10
Age 45-60	0.10	76.40	80.10	**92.84**	86.30	80.87	83.95±0.97	55.10±0.71
Age >60	0.06	89.00	93.80	**97.58**	94.80	86.29	92.74±0.53	56.59±4.08
Acces. Hat	0.10	86.70	90.40	**96.05**	91.80	81.69	91.64±0.26	59.45±3.09
Acces. Muffler	0.08	91.30	93.70	**97.17**	96.10	85.59	94.28±0.70	63.69±1.67
Acces. Nothing	0.75	80.00	82.70	86.11	85.80	74.65	**87.16±0.44**	67.77±1.37
Acces. Sunglasses	0.03	51.70	53.90	-	69.90	**76.18**	60.66±1.16	61.09±1.41
Carry. Backpack	0.20	67.20	70.50	**84.30**	82.60	74.19	83.91±1.13	76.69±0.75
Carry Messenger Bag	0.30	75.50	78.30	79.58	82.00	71.99	**83.06±0.21**	64.80±1.38
Carry. Nothing	0.28	71.50	76.50	80.14	83.10	71.31	**84.45±0.38**	68.34±0.56
Carry. Other	0.20	68.00	73.00	**80.91**	77.30	69.71	75.34±0.45	54.31±0.73
Carry. Plastic Bag	0.08	75.50	81.30	**93.45**	87.00	77.85	82.75±1.19	54.91±1.67
Footwear Leather	0.30	81.70	87.20	85.26	87.30	79.11	**87.48±0.81**	66.91±1.64
Footwear Sandals	0.02	50.30	52.20	-	67.30	**71.06**	63.56±2.51	63.91±2.45
Footwear Shoes	0.36	56.50	78.40	75.78	**80.00**	70.33	79.22±0.49	60.29±0.63
Footwear Sneakers	0.22	69.30	75.00	**81.78**	78.70	72.48	78.70±0.61	69.10±1.77
Hair Long	0.24	72.80	80.10	88.12	88.90	75.94	**89.81±0.22**	78.97±1.45
Lower Casual	0.86	71.30	78.20	**90.54**	84.90	77.39	86.90±0.96	63.94±2.16
Lower Formal	0.14	71.90	79.00	**90.86**	85.20	77.96	84.54±0.96	60.59±1.94
Lower Jeans	0.31	76.00	81.00	83.13	85.70	73.67	**86.05±0.86**	74.62±1.22
Lower Shorts	0.05	56.50	65.20	-	80.40	78.19	**80.57±1.50**	80.05±1.11
Lower Skirt	0.03	64.30	69.60	-	**82.20**	72.37	78.93±0.52	62.28±0.47
Lower Trousers	0.52	76.50	82.20	76.26	84.30	71.51	**85.74±0.70**	50.00±0.83
Upper Casual	0.85	71.30	78.10	**89.25**	84.40	75.20	86.58±0.37	82.98±1.34
Upper Formal	0.13	70.00	78.70	**91.12**	85.10	78.13	**84.58±0.40**	60.28±0.77
Upper Jacket	0.07	67.90	72.20	**92.34**	79.20	74.32	73.33±1.15	55.52±1.41
Upper Logo	0.04	50.70	52.70	-	68.40	66.77	**69.42±0.72**	93.93±1.43
Upper Other	0.46	83.90	**87.30**	81.97	*86.10*	78.72	*86.63±0.32*	75.21±0.48
Upper Plaid	0.03	65.00	65.20	-	81.10	71.87	**92.77±1.86**	76.58±3.44
Upper Short Sleeve	0.14	71.60	75.80	**88.09**	87.50	77.35	86.62±0.69	85.06±0.77
Upper Stripes	0.02	52.30	52.30	-	66.50	65.41	**67.26±1.33**	63.99±3.45
Upper T-shirt	0.08	64.20	71.40	**90.59**	83.00	75.62	77.16±0.73	76.47±0.24
Upper V-Neck	0.01	51.10	53.30	-	69.80	**75.63**	58.79±1.54	51.19±0.22
Average	0.24	70.63	75.59	-	**82.60**	75.43	81.65±0.83	66.45±1.49
All Correct	-	-	-	-	-	-	**40.12**±1.38	3.31±0.32

Table 6.1: Mean recognition accuracy (mA %) comparison of binary labels across previous PETA studies and our work. (S) Separate estimation. (J) Joint estimation. (**Bold**) Highest accuracy. Standard instance-level (multi-shot) and subject-level (zero-shot) training split criteria are evaluated for our approach.

6.6.2 Recognition Model Benchmark Results

Our first experiment compares the pretrained ResNet model to five previous works on PETA which record binary attribute recognition accuracy [45, 50, 57, 94, 95], estimating traditional binary labels in a multi-shot scenario.

The three most recent works [45, 50, 95] employ CNN architectures to jointly classify the 35 binary attributes. Zhu et al. update their original MLCNN model [41] (discussed in Section 4.7) with more convolutional layers and remove attribute-region constrains in favour of fully connected output layers in [95]. Li et al. employ an AlexNet architecture for DeepMAR with attribute-level weighting, achieving promising results [45]. Most notably, Bekele et al. employ an extremely similar Multi-Attribute ResNet (MAResNet) model, published at the time of writing this thesis [50]. However, none of these approaches employ a pre-trained network.

Many of the binary attributes are highly imbalanced (Table 6.1 ratio) as mentioned in Section 6.2.1. To address this, these works employ custom loss functions based on sigmoid cross entropy [45, 50] or softmax [95]. In contrast, when introducing PETA in 2014, Deng et al. learnt attribute predictors separately with a Markov Random Field (MRF) graph and feature descriptors similar to `ELF` (Section 4.4) [57]. Training independent predictors enables dataset augmentation to correct for imbalanced attributes at training time, as in [57, 94]. As we are concerned with retrieval performance in the next experiment, we choose not to prioritise the learning of under-represented classes over well-represented classes which incidentally contain more discriminative power.

Table 6.1 reports the 'mean recognition accuracy' which accounts for imbalances between attributes' positive to negative sample ratio, measured as $mA = (\frac{TP}{P} + \frac{TN}{N})/2$ following the evaluation methodology in [57]. We find our model performs best on the majority of attributes and significantly outperforms previous methods on more balanced and discriminative attributes e.g. Gender, Age 15-30, Lower Trousers. Therefore, in the next experiments, super-fine attributes must compete with already highly accurate binary label estimations. In this scenario ResNet achieves an average of 40.12±1.38% totally correct label estimations.

For imbalanced attributes our approach occasionally attains the top performance e.g. Upper Plaid and other times the least correspondence e.g. Upper V-Neck. Interestingly, our pre-trained model gains consistently higher results than MAResNet [50], which trains a similar ResNet architecture without pre-trained initialisation.

We repeat the experiment with zero-shot subject-level set-split criteria to highlight the challenges of training and testing on disjoint subject sets. In this scenario the average mA is 15.2% lower, with only 3.31±0.32% totally correct image estimations, most affected by traits with ratios in the range $[0.10, 0.35]$, where inferring the most common occurrence is detrimental. This reiterates the findings of [46], where gender accuracy rates dropped over 16% after removing quasi-identical images from the PETA dataset. Zero-shot evaluation evidences that much of the measured estimator accuracy is due to training on highly similar images examples in a multi-shot scenario, and less likely from learning truly generalised descriptors.

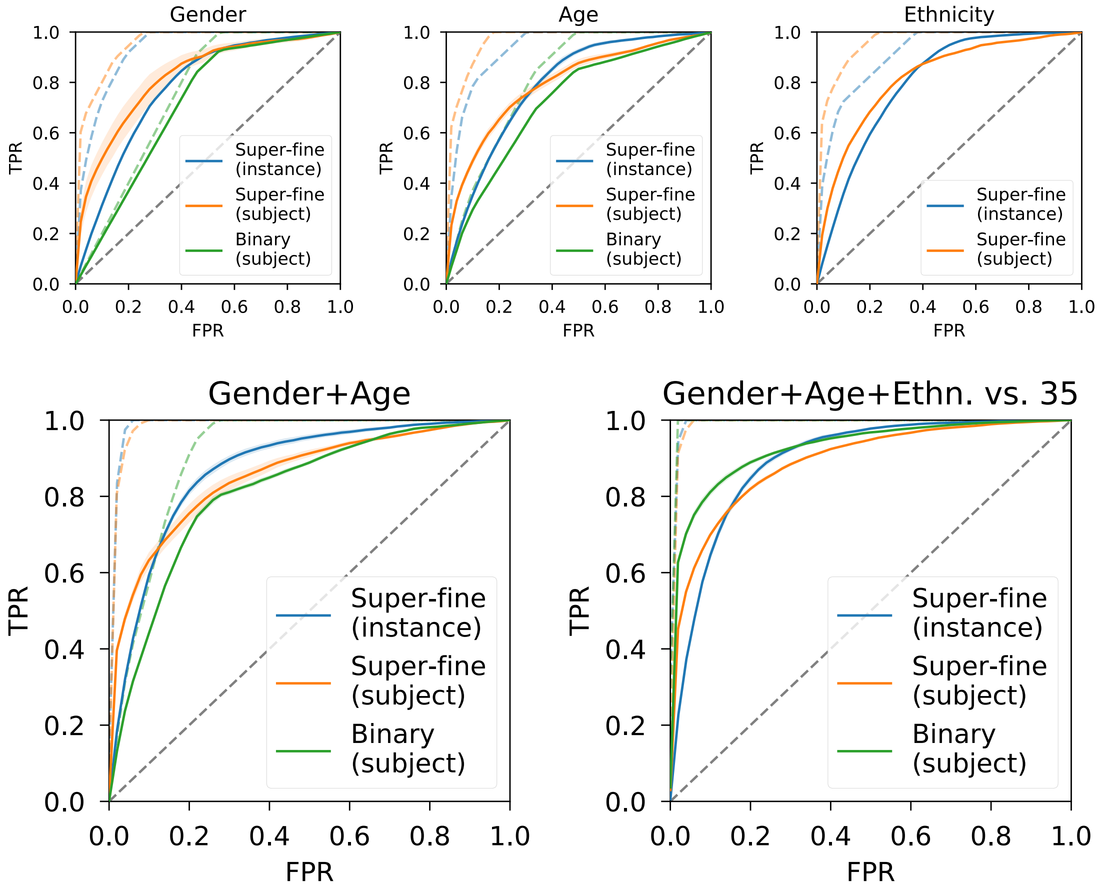


Figure 6.12: Multi-shot scenario Ranked Retrieval ROC curves (instance-level set-split criteria). Dotted lines indicate maximum performance assuming perfect label estimation. Shaded areas represent standard deviation. (TPR) True Positive Rate. (FPR) False Positive Rate.

6.6.3 Multi-shot Re-identification Results

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

Figure 6.12 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 by up to 11.2% mAP and 7.2% mAP for *gender* and *age* respectively (Table 6.2).

Dotted lines indicate the maximum possible performance limit given perfect label estimations. Although represented with 5 classes, binary *age* only slightly enhances its theoretical 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 much (if any) greater annotation effort. Remarkably, estimated super-fine labels

	Attribute modality	**Super-fine** (mAP)		**Binary** (mAP)
Retrieval	Labelling	Instance-level	Subject-level	Subject-level
Multi-shot	Gender	0.776±0.002	**0.821±0.031**	0.709±0.003
	Age	0.788±0.007	**0.795±0.012**	0.723±0.002
	Ethnicity	0.801±0.001	**0.821±0.001**	-
	Gender+Age	**0.871±0.006**	0.854±0.010	0.813±0.004
	Gender+Age+Ethnicity	0.893±0.004	0.890±0.002	-
	All 35	-	-	**0.927±0.003**
Zero-shot	Gender	0.731±0.012	**0.733±0.002**	0.665±0.007
	Age	**0.762±0.007**	0.749±0.007	0.614±0.015
	Ethnicity	**0.789±0.003**	0.777±0.008	-
	Gender+Age	**0.826±0.006**	0.804±0.004	0.692±0.015
	Gender+Age+Ethnicity	**0.872±0.005**	0.845±0.007	-
	All 35	-	-	0.807±0.006

Table 6.2: Ranked retrieval results - mean Average Precision (mAP). (**Bold**) Highest mAP.

	Attribute modality	**Super-fine** (R^2)		**Binary** (Acc %)
Retrieval	Labelling	Instance-level	Subject-level	Subject-level
Multi-shot	Gender	0.680±0.013	0.725±0.073	92.20±0.54
	Age	0.622±0.024	0.549±0.025	79.85±0.46
	Ethnicity	0.773±0.001	0.696±0.006	-
	Gender+Age	0.663±0.015	0.631±0.007	75.43±0.75
	Gender+Age+Ethnicity	0.688±0.011	0.666±0.008	-
	All 35	-	-	40.23±1.36
Zero-shot	Gender	0.521±0.039	0.514±0.010	83.31±1.62
	Age	0.517±0.029	0.428±0.020	51.27±3.11
	Ethnicity	0.713±0.041	0.418±0.259	-
	Gender+Age	0.521±0.031	0.485±0.009	44.28±2.91
	Gender+Age+Ethnicity	0.610±0.031	0.456±0.117	-
	All 35	-	-	3.31±0.32

Table 6.3: Ranked retrieval results - coefficient of determination (R^2), accuracy percentage (Acc %).

also outperform the theoretical 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 attributes.

Furthermore, *ethnicity* appears to outperform both *gender* and *age* (Table 6.2), indicating how imperative it is 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 information, improving automatic label inference. Intuitively, by enhancing attribute relevance we jointly enhance the accuracy and precision of both ground-truth and estimated labels.

Table 6.3 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,

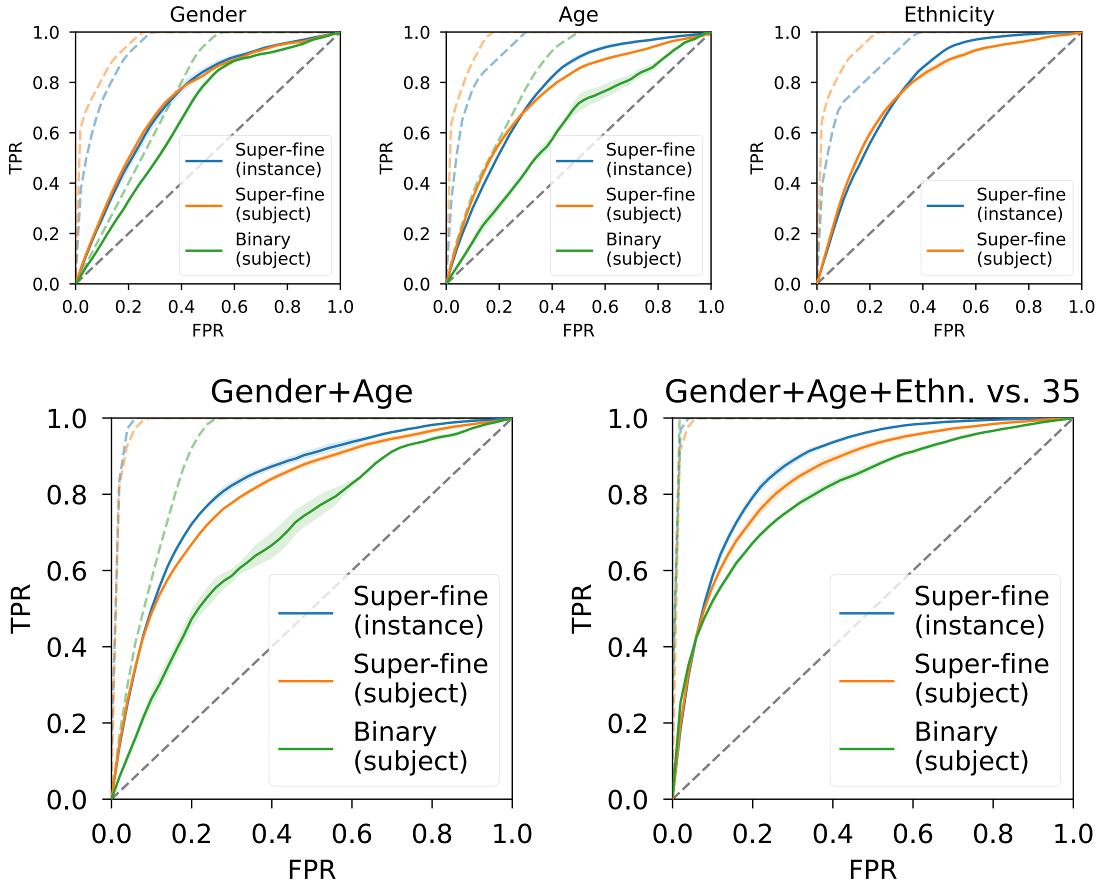


Figure 6.13: Zero-shot scenario Ranked Retrieval ROC curves (subject-level set-split criteria). Dotted lines indicate maximum performance assuming perfect label estimation. Shaded areas represent standard deviation. (TPR) True Positive Rate. (FPR) False Positive Rate.

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 the *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.

6.6.4 Zero-shot Identification Results

This experiment emulates a real-world surveillance scenario, where estimators are evaluated without prior exposure to similar instances. 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. Subject-level binary labels fair particularly poorly in zero-shot, finding super-fine *gender* and *age* to improve mAPs up to 6.8% and 14.8% respectively.

Retrieval performance limits are not affected by alternative set-split criteria. However, in this experiment, instance- and subject-level super-fine labels achieve more comparable performance (Figure 6.13). Relative to multi-shot, instance-level labels are now more accurately estimated over subject-level labels, as they capture intra-subject variation. Consequently, estimators can better generalise feature descriptors to totally dissimilar instances, reiterated by higher R^2 scores in Table 6.3. In particular, *ethnicity* appears to be the most robust trait, suffering the least performance drop 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 *all 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.

6.7 Conclusions

Building upon super-fine attributes from Chapter 5, we propose crowd prototyping to pre-discover salient perceptual concepts that facilitate the efficient large-scale annotation of super-fine attributes. This enables our approach to describe perceived visual variation, ambiguity and uncertainty in challenging images and at scale.

The versatility of crowd prototyping is demonstrated by discovering and annotating *gender*, *age* and *ethnicity* traits on the PETA dataset. Despite traditionally fraught binary, ordinal and categorical measurement scales, our process is able to methodically discover objective, visual discernments for each trait. We provide extensive analysis of the crowd prototyping and annotation processes, comparing our new super-fine labels to conventional ground-truths.

By substantially improving label relevance and fidelity over conventional approaches, we establish the superiority of super-fine over binary labels, for discrimination, ranked retrieval and generalisation in zero-shot identification scenarios. Our incorporation of unsupervised learning techniques with state-of-the-art supervised image recognition not only outperforms conventional approaches, but exceed the maximum possible retrieval performance of binary *gender* and *age* descriptions with estimated super-fine labels in a multi-shot identification scenario. We also note that the coarse-grained and highly varied categories produced by crowd prototyping are more pertinent than expertly-defined lexicons and visual annotation is more objective than conventional text-based approaches. Our findings also highlight the critical importance of *ethnicity* in identification, as the most consistently discriminative trait.

The effectiveness of our novel super-fine attributes is investigated, jointly regressing labels by fine-tuning the ResNet-152 CNN model. Remarkably, this model also outperforms previous approaches on PETA for more balanced binary attribute classification. Better still, regressed super-fine labels outperform automatically classified binary labels across all ranked retrieval scenarios, demonstrating that accurately modelling subjectivity and uncertainty is key to learning robust estimators.

Significantly, instance-level labels supersede subject-level labels in the zero-shot scenario, suggesting they would benefit real-world applications. Finally, in a zero-shot scenario we note that our 3 super-fine attributes outperform 35 conventional attributes for ranked subject retrieval, as a result of our radically new approach.

Chapter 7

Conclusions & Future Work

7.1 Conclusions

Our work presents exciting possibilities for ubiquitous real-world identification and development towards comprehensive machine intelligence, uniting soft biometrics with novel unconstrained annotation and contemporary machine learning techniques.

In Chapter 2 we recapitulate the current state-of-the-art in human identification and image attribute recognition, documenting the paradigm shift in estimating hard, physical and categorical ground-truths from the face to the discovery of finer-grained, human perceived descriptions from the body. Chapter 3 demonstrates the tremendous potential of crowdsourcing comparative annotations that generate precise, relative attributes aiding in subject retrieval. Chapter 4 then advances stand-alone soft biometric identification, exploring two image attribute recognition methodologies, highlighting the power of regressing relative attributes with a deep learning CNN. Chapter 5 breaks away from convention by exploring a novel and pragmatic technique for eliciting objective, yet unconstrained super-fine attribute descriptions from highly subjective images, transitioning our research towards real-world data and providing an in-depth analysis of the fundamental *gender* trait. Finally, Chapter 6 builds upon this new solution, facilitating the large-scale annotation of challenging images with crowd prototyping and radically increased label precision with super-fine attributes. The chapter concludes by recognising super-fine images attributes with a pre-trained CNN, demonstrating state-of-the-art soft biometric identification performance on real-world surveillance imagery.

This thesis tackles a number of pertinent challenges, often overlooked in current identity science literature. As a result, we recommend four points for subsequent human identification approaches:

1. **Increased precision**. Chapters 3, 4, 5 and 6 all confirm that much more precise, relative and super-fine attributes are essential for effective human-machine communication of subject descriptions, leading to consistently superior identification performance. Coarse-grained, categorical and binary descriptions are no longer suitable for the needs of highly discriminative descriptive systems. Therefore, techniques for enhancing descriptive precision, relevance and objectivity should be of primary attention in future studies. Furthermore, a focus should be on refining existing and well studied traits, harnessing their previously untapped discriminative power, rather than supplementing ever more attributes to the annotation process. This is evidenced in Chapter 6, where 3 traits labelled with super-fine attributes outperformed 35 binary traits in zero-shot identification.

2. **Clarity, ambiguity and uncertainty**. In Chapter 5, we advocate the inclusion of 'neutral' and 'uncertain' categories for studies involving human perception and demographic labelling, as endorsed by Golomb et al. in 1990 [30]. The study of less distinguishing perceptual concepts is central to the success of super-fine attributes in Chapter 6, enabling more faithful semantic image representations. This indicates that concepts of uncertainty and ambiguity are imperative in enabling machine learning algorithms to fully generalise to human-centric tasks.

3. **Non-verbal communication**. The advantage of similarity comparisons is their ability to naturally abstract away from linguistic constraints. Enabling identity descriptions to remain within the image domain is key to finding even more relevant, precise and accurate subject descriptions. Such non-verbal communication may also be less susceptible to anchoring problems and the cross-race affect than alternative methods, making it ideal for unconstrained surveillance.

4. **Zero-shot and instance-level labels**. Annotation and retrieval at instance-level is shown to outperform the traditional subject-level approach in Chapter 6, as more specific super-fine attributes generalise better in the challenging zero-shot scenario. With surveillance growing ever more pervasive, identifying subject instances using visual descriptions is becoming a viable solution. Therefore, exploring this avenue will facilitate the next generation of surveillance applications.

7.2 Future Work

7.2.1 Active Learning

Our work evidently improves upon existing identification methods but also raises new questions when dealing with emerging problems in the field. An especially challenging topic is the construction of new ground-truth information from human perception without pre-defined targets. In contrast to other attribute discovery works that aim to reproduce prior 'gold-standard' ground-truths, our discovery process is instead concerned with finding the most *discriminative* attributes. This presents a significant difficulty, as early stage hyper-parameters e.g. M dimensions and c clusters for crowd prototype generation cannot be fully evaluated for their discriminant power until labels are subsequently collected, learnt and retrieved. Alleviating this bottleneck with an all-in-one solution that further optimises efficiency and performance would be of primary attention in future studies.

Facilitating an end-to-end annotation and identification solution would remove the need to find intermediate representations through unsupervised techniques. An interactive, human-in-the-loop approach to gamify labelling e.g. BubbleGame [67], could vastly minimise the number of annotations required and discover optimal internal representations while classifying images iteratively. Alternatively, a Siamese CNN that accepts two images as input and regresses a similarity score as output could be trained from similarity comparisons alone, as in [144]. This would circumvent the need for prototype discovery and its associated hyper-parameters.

A further avenue of investigation could include evaluating various triplet and grid-based comparison approaches against pairwise comparisons of similarity as in [79, 80]. While these approaches already target selecting discriminative image features to improve recognition accuracy, the challenge would be in extending them to simultaneously learn optimal instance- and subject-level ground-truth discriminations from the crowd's description alone. This would entail iteratively performing model training and ranked retrieval testing as part of an online algorithm's interactive loop.

7.2.2 Behavioural Categorisation

Akhtar et al. note that behavioural classification and identification are becoming integral to surveillance [16]. Therefore, the next frontier of soft biometrics is in dealing with temporal information. In fact, advances in Recurrent Neural Networks (RCNNs) with Long Short-Term Memory (LSTM) blocks are already enabling the next generation of solutions in other areas of video classification. A novel future development would be to extend a form of similarity comparison that utilises the advantages of higher-dimensional descriptions to capture temporarily dependent information such as gait, enabling ever more comprehensive identification.

7.2.3 EFIT-V

A practical application of our work would be its incorporation into existing suspect identification systems like EFIT-V [145], which generate suspect faces using grid-based selection to iteratively refine the visualisation, Figure 7.1. Currently, such systems

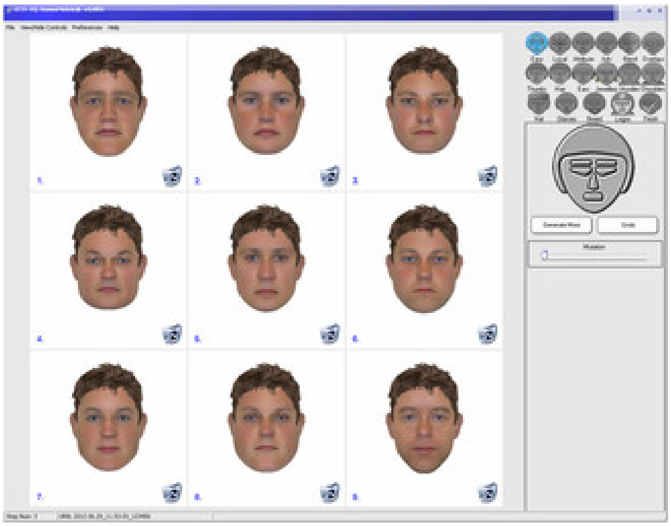


Figure 7.1: EFIT-V presents the witness with groups of computer-generated faces and the basic task of the witness is to make simple decisions of selection and rejection in relation to sequences of faces presented to them. In this way, EFIT-V gradually learns and refines the facial appearance the witness has in mind, progressing to a better and better likeness [145].

randomly select initial variables from a large parameter space and hone in on desirable permutations with user feedback. An online, similarity comparison approach could be utilised to build finer and finer-grained prototype clusters of similar images, limited not only to faces, but also encapsulating body images. With prior knowledge of perceived similarity between a large corpus of images, such systems could be extended to deal with a greater range of realistic imagery and present even more relevant identity examples.

7.2.4 ImageNetV2 with Hierarchical Similarity

The final topic we propose is tangential to human identification, but involves image classification at large. As the most substantial ontology of images, ImageNet [87] has played an important role in the evolution of computer vision in the 21st century. It contains 50 million image instances, labelled across 100K 'synsets' organized according to the WordNet hierarchy. Figure 7.2 overviews the construction of the dataset, which classifies instances into varying granularities of semantic concept, from broad level classes e.g. 'animal' and 'instrument' to very fine-grained classes e.g. 'spotted lynx' and 'acoustic guitar'. The WordNet lexicon is built on hierarchical, ontological relationships between words, where leaf nodes represent the closest corresponding synonyms. This fantastic source of information also comes with a number of caveats, resulting in sometimes peculiar image classifications and only accounts for unambiguous images which contain one clear concept.

A compelling avenue of future investigation would be to create an ImageNetV2 ontology based solely on perceived image similarities at several levels of granularity, forming a similarity hierarchy. Providing a 'semantic distance' between every image pair would

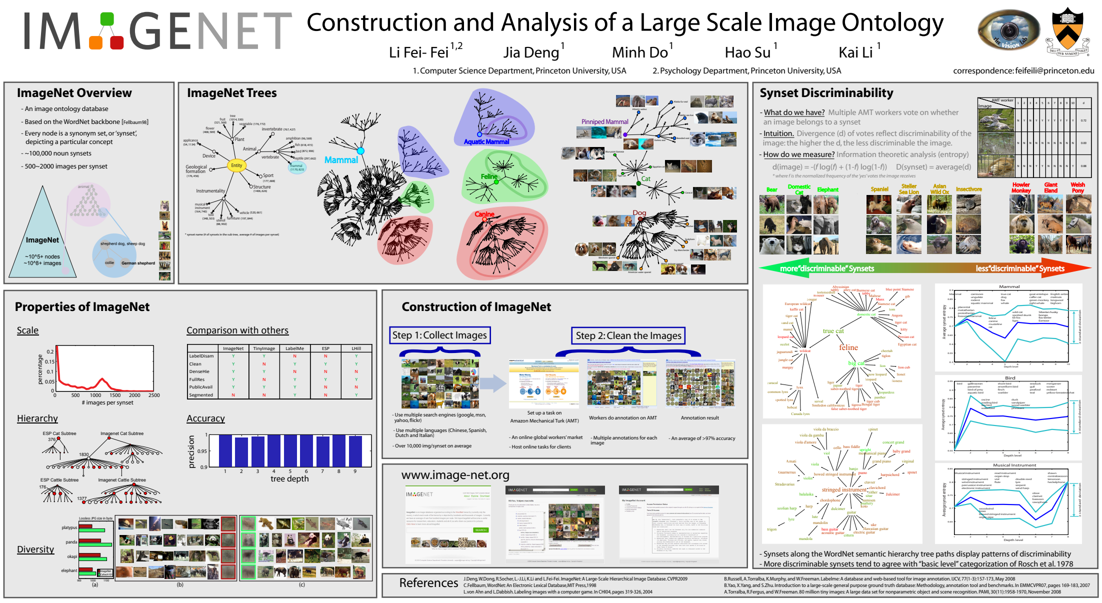

IMAGENET
Construction and Analysis of a Large Scale Image Ontology
Li Fei- Fei 1,2
Jia Deng 1
Minh Do 1
Hao Su 1
Kai Li 1
1. Computer Science Department, Princeton University, USA
2. Psychology Department, Princeton University, USA
correspondence: feifeili@princeton.edu
ImageNet Overview
- An image ontology database
- Based on the WordNet backbone
- Every node is a synonym set, or 'synset', depicting a particular concept
- ~100,000 noun synsets
- 500~2000 images per synset
ImageNet Trees
Mammal
Aquatic Mammal
Feline
Canine
Pinniped Mammal
Cat
Dog
Synset Discriminability
more"discriminable"Synsets
less"discriminable"Synsets
Properties of ImageNet
Scale
Comparison with others
Hierarchy
Accuracy
Diversity
# images per synset
tree depth
precision
Construction of ImageNet
Step 1: Collect Images
Step 2: Clean the Images
www.image-net.org
References

Figure 7.2: Overview of ImageNet an image ontology database based on the WordNet. Every node is a synonym set, or synset, depicting a particular concept with around 100K noun synsets, each with 500-2000 images [87].

better model human image perceptions and reveal a plethora of new enquiries such as the inclusion of difficult to classify images and the annotation of multiple concepts per image each with multiple dimensions. Pursing these topics will establish the next generation of computer vision and machine learning algorithms to rival human intelligence.

7.3 Final Remarks

Looking towards the future, we believe knowledge representation is *the* grand challenge in computer vision and machine learning. Our first steps towards ubiquitous identification will undoubtedly be refined e.g. super-fine attributes, crowd prototyping and image attribute recognition, but the underlying premise will remain; to find higher-dimensional shared embeddings with which to communicate descriptions more freely and 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] S. Samangooei, B. Guo, and M. S. Nixon. The use of semantic human description as a soft biometric. In *BTAS*, pages 1–7. IEEE, 2008.

[2] D. Reid, M.S. Nixon, and S. Stevenage. Soft biometrics; human identification using comparative descriptions. *IEEE TPAMI*, 36(6):1216–1228, 2013.

[3] R.D. Seely, S. Samangooei, M. Lee, J.N. Carter, and M.S. Nixon. The university of southampton multi-biometric tunnel and introducing a novel 3d gait dataset. In *BTAS*, pages 1–6. IEEE, 2008.

[4] M.S. Nixon, P.L. Correia, and K. Nasrollahi et al. On soft biometrics. *Pattern Recognition Letters*, 68:218 – 230, 2015.

[5] A. Dantcheva, P. Elia, and A. Ross. What else does your biometric data reveal? a survey on soft biometrics. *IEEE TIFS*, 11(3):441–467, 2016.

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

[7] M.Q. Hill, S. Streuber, and A.J. O'Toole. Creating body shapes from verbal descriptions by linking similarity spaces. *Psychological Science*, 27(11):1486–1497, 2016.

[8] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. In *ICMR*, pages 153–160. ACM, 2014.

[9] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE TPAMI*, 22(12):1349–1380, 2000.

[10] J.S. Hare, P.H. Lewis, P. Enser, and C.J. Sandom. Mind the gap: another look at the problem of the semantic gap in image retrieval. In *Electronic Imaging*, page 9. SPIE, 2006.

[11] A. Bertillon. *Signaletic instructions including the theory and practice of anthropometrical identification*. Werner Company, 1896.

[12] T. Lucas and M. Henneberg. Comparing the face to the body, which is better for identification? *International Journal of Legal Medicine*, 130(2):533–540, 2016.

[13] D. Adjeroh, D. Cao, M. Piccirilli, and A. Ross. Predictability and correlation in human metrology. In *WIFS*, pages 1–6. IEEE, 2010.

[14] I.A. Kakadiaris, N. Sarafianos, and C. Nikou. Show me your body: Gender classification from still images. In *ICIP*, pages 3156–3160. IEEE, 2016.

[15] A.K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE TCSVT*, 14(1):4–20, 2004.

[16] Z. Akhtar, A. Hadid, M. Nixon, M. Tistarelli, J. Dugelay, and S. Marcel. Biometrics: in search of identity and security: Q & a. *IEEE MultiMedia*, April 2017.

[17] A.K. Jain, S.C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In *ICB*, pages 731–738. Springer, 2004.

[18] A.K. Jain, S.C. Dass, and K. Nandakumar. Can soft biometric traits assist user recognition? In *Biometric Technology for Human Identification*, pages 561–572. SPIE, 2004.

[19] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M.S. Nixon. Soft biometrics and their application in person recognition at a distance. *IEEE TIFS*, 9(3):464–475, 2014.

[20] S. Denman, C. Fookes, A. Bialkowski, and S. Sridharan. Soft-biometrics: unconstrained authentication in a surveillance environment. In *DICTA*, pages 196–203. IEEE, 2009.

[21] E.S. Jaha and M.S. Nixon. Soft biometrics for subject identification using clothing attributes. In *IJCB*, pages 1–6. IEEE, 2014.

[22] A. Dantcheva, C. Velardo, A. D'angelo et al., and J.-L. Dugelay. Bag of soft biometrics for person identification. *Multimedia Tools and Applications*, 51(2):739–777, 2011.

[23] U. Park and A.K. Jain. Face matching and retrieval using soft biometrics. *IEEE TIFS*, 5(3):406–415, 2010.

[24] A.K. Jain, K. Nandakumar, X. Lu, and U. Park. Integrating faces, fingerprints, and soft biometric traits for user recognition. In *ICB*, pages 259–269. Springer, 2004.

[25] N.Y. Almudhahka, M.S. Nixon, and J.S. Hare. Unconstrained human identification using comparative facial soft biometrics. In *BTAS*, pages 1–6. IEEE, 2016.

[26] E.S. Jaha and M.S. Nixon. From clothing to identity: Manual and automatic soft biometrics. *IEEE TIFS*, 11(10):2377–2390, 2016.

[27] The pnc user manual. Technical report, National Policing Improvement Agency, 2012. http:/www.levesoninquiry.org.uk/wp-content/uploads/2012/04/Exhibit-KW-NIPA3.pdf.

[28] Codes of practice code d identification of persons by police officers. Technical report, Home Office, 2011. https://www.gov.uk/government/publications/pace-code-d-2011.

[29] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708. IEEE, 2014.

[30] B.A. Golomb, D.T. Lawrence, and T.J. Sejnowski. Sexnet: a neural network identifies sex from human faces. In *NIPS*, pages 572–572. MIT Press, 1990.

[31] C.B. Ng, Y.H. Tay, and B. Goi. Recognizing human gender in computer vision: a survey. In *PRICAI*, pages 335–346. Springer, 2012.

[32] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738. IEEE, 2015.

[33] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *CVPR*, pages 34–42. IEEE, 2015.

[34] X. Wang, R. Guo, and C. Kambhamettu. Deeply-learned feature for age estimation. In *WACV*, pages 534–541. IEEE, 2015.

[35] H. Han, C. Otto, X. Liu, and A.K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE TPAMI*, 37(6):1148–1161, 2015.

[36] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275. Springer, 2008.

[37] R. Layne, Timothy M. Hospedales, and S. Gong. Attributes-based re-identification. In *Person Re-Identification*, pages 93–117. Springer, 2014.

[38] 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*, pages 5315–5324. IEEE, 2015.

[39] X. Li, A. Wu, and M. Cao. Towards more reliable matching for person re-identification. In *ISBA*, pages 1–6. IEEE, 2015.

[40] L. An, X. Chen, M. Kafai, S. Yang, and B. Bhanu. Improving person re-identification by soft biometrics based reranking. In *ICDSC*, pages 1–6. IEEE, 2013.

[41] J. Zhu, S. Liao, D. Yi, Z. Lei, and S.Z. Li. Multi-label cnn based pedestrian attribute learning for soft biometrics. In *ICB*, pages 535–540. IEEE, 2015.

[42] S. Gong, M. Cristani, C.C. Loy, and T.M. Hospedales. The re-identification challenge. In *Person re-identification*, pages 1–20. Springer, 2014.

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

[44] R. Layne, T.M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *BMVC*, page 8, 2012.

[45] D. Li, X. Chen, and K. Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *ACPR*, pages 111–115. IEEE, 2015.

[46] G. Antipov, S. Berrani, N. Ruchaud, and J. Dugelay. Learned vs. hand-crafted features for pedestrian gender recognition. In *ACMMM*, pages 1263–1266. ACM, 2015.

[47] 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.

[48] P. Sudowe, H. Spitzer, and B. Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *ICCV*, pages 87–95. IEEE, 2015.

[49] Y. Li, C. Huang, C.C. Loy, and X. Tang. Human attribute recognition by deep hierarchical contexts. In *ECCV*, pages 684–700. Springer, 2016.

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

[51] O.A. Arigbabu, S.M. Syed Ahmad, W.A. Adnan, and S. Yussof. Integration of multiple soft biometrics for human identification. *Pattern Recognition Letters*, 68:278–287, 2015.

[52] E.S. Jaha and M.S. Nixon. Viewpoint invariant subject retrieval via soft clothing biometrics. In *ICB*, pages 73–78. IEEE, 2015.

[53] S. Denman, M. Halstead, C. Fookes, and S. Sridharan. Searching for people using semantic soft biometric descriptions. *Pattern Recognition Letters*, 68:306–315, 2015.

[54] Z. Zhou, Y. Wang, and E. K. Teoh. A framework for semantic people description in multi-camera surveillance systems. *Image and Vision Computing*, 46(C):29–46, 2016.

[55] S.J. Pan and Q. Yang. A survey on transfer learning. *IEEE TKDE*, 22(10):1345–1359, 2010.

[56] Z. Shi, T.M. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *CVPR*, pages 4184–4193. IEEE, 2015.

[57] Y. Deng, P. Luo, C.C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *ACMMM*, pages 789–792. ACM, 2014.

[58] D. Hall and P. Perona. Fine-grained classification of pedestrians in video: Benchmark and state of the art. In *CVPR*, pages 5482–5491. IEEE, 2015.

[59] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124. IEEE, 2015.

[60] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[61] G. Patterson and J. Hays. Coco attributes: Attributes for people, animals, and objects. In *ECCV*, pages 85–100. Springer, 2016.

[62] A. Kovashka, O. Russakovsky, L. Fei-Fei, and K. Grauman. Crowdsourcing in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 10(3):177–243, 2016.

[63] A.J. O'Toole and P.J. Phillips. Five principles for crowd-source experiments in face recognition. In *FG*, pages 735–741. IEEE, 2017.

[64] A. Rice, P.J. Phillips, V. Natu, X. An, and A.J. O'Toole. Unaware person recognition from the body when face identification fails. *Psychological science*, 24(11):2235–2243, 2013.

[65] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, pages 503–510. IEEE, 2011.

[66] A. Yu and K. Grauman. Just noticeable differences in visual attributes. In *ICCV*, pages 2416–2424, 2015.

[67] J. Deng, J. Krause, M. Stark, and L. Fei-Fei. Leveraging the wisdom of the crowd for fine-grained recognition. *IEEE TPAMI*, 38(4):666–676, 2016.

[68] B. Qian, X. Wang, N. Cao, Y. Jiang, and I. Davidson. Learning multiple relative attributes with humans in the loop. *IEEE Image Processing*, 23(12):5573–5585, 2014.

[69] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar. Attribute and simile classifiers for face verification. In *CVPR*, pages 365–372. IEEE, 2009.

[70] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, pages 2973–2980. IEEE, 2012.

[71] C.A. Meissner, S.L. Sporer, and J.W. Schooler. Person descriptions as eyewitness evidence. *Handbook of eyewitness psychology: Memory for people*, pages 1–34, 2013.

[72] R.H. Flin and J.W. Shepherd. Tall stories: Eyewitnesses' ability to estimate height and weight characteristics. *Human Learning: Journal of Practical Research & Applications*, 5(1):29–38, 1986.

[73] J.C. Yuille and J.L. Cutshall. A case study of eyewitness memory of a crime. *Journal of Applied Psychology*, 71(2):291, 1986.

[74] M.H. Kiapour, K. Yamaguchi, A.C. Berg, and T.L. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, pages 472–488. Springer, 2014.

[75] Y. Fu, T.M. Hospedales, T. Xiang, J. Xiong, S. Gong, Y. Wang, and Y. Yao. Robust subjective visual property prediction from crowdsourced pairwise labels. *IEEE TPAMI*, 38(3):563–577, 2016.

[76] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785. IEEE, 2009.

[77] C.H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36(3):453–465, 2014.

[78] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, pages 1681–1688. IEEE, 2011.

[79] R.G. Gomes, P. Welinder, A. Krause, and P. Perona. Crowdclustering. In *NIPS*, pages 558–566. MIT Press, 2011.

[80] C. Wah, G. Horn, S. Branson, S. Maji, P. Perona, and S. Belongie. Similarity comparisons for interactive fine-grained categorization. In *CVPR*, pages 859–866, 2014.

[81] L.L. Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

[82] S. Edelman and R. Shahbazi. Renewing the respect for similarity. *Frontiers in computational neuroscience*, 6:45–45, 2012.

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

[84] S. Maji. Discovering a lexicon of parts and attributes. In *ECCV*, pages 21–30. Springer, 2012.

[85] D. Ellis and B. Whitman. The quest for ground truth in musical artist similarity. In *ISMIR*, pages 170–177, 2002.

[86] V.S. Sheng, F. Provost, and P.G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *SIGKDD*, pages 614–622. ACM, 2008.

[87] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[88] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In *CVPR*, pages 1777–1784. IEEE, 2011.

[89] W.J. Scheirer, S.E. Anthony, K. Nakayama, and D.D. Cox. Perceptual annotation: Measuring human vision to improve computer vision. *IEEE TPAMI*, 36(8):1679–1686, 2014.

[90] M.D. MacLeod, J.N. Frowley, and J.W. Shepherd. Whole body information: Its relevance to eyewitnesses. *Adult eyewitness testimony: Current trends and developments*, pages 125–143, 1994.

[91] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD*, pages 133–142. ACM, 2002.

[92] T. Matthews, M.S. Nixon, and M. Niranjan. Enriching texture analysis with semantic data. In *CVPR*, pages 1248–1255. IEEE, 2013.

[93] H.A. Perlin and H.S. Lopes. Extracting human attributes using a convolutional neural network approach. *Pattern Recognition Letters*, 68:250–259, 2015.

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

[95] J. Zhu, S. Liao, Z. Lei, and S.Z. Li. Multi-label convolutional neural network based pedestrian attribute classification. *Image and Vision Computing*, 58:224–229, 2016.

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

[97] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159. IEEE, 2014.

[98] E. Ahmed, M. Jones, and T.K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, pages 3908–3916. IEEE, 2015.

[99] C. Liu, S. Gong, C.C. Loy, and X. Lin. Person re-identification: What features are important? In *ECCV*, pages 391–401. Springer, 2012.

[100] S.M. Pizer, E.P. Amburn, J.D. Austin, Ro. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J.B. Zimmerman, and K. Zuiderveld. Adaptive histogram equalization and its variations. *CVGIP*, 39(3):355–368, 1987.

[101] S. Samangooei and M.S. Nixon. On semantic soft-biometric labels. In *BIOMET*, pages 3–15. Springer, 2014.

[102] Wikipedia. Color space — wikipedia, the free encyclopedia, 2015. [Online; accessed 11-June-2015].

[103] C.H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958. IEEE, 2009.

[104] J.S. Hare, S. Samangooei, P.H. Lewis, and M.S. Nixon. Semantic spaces revisited: investigating the performance of auto-annotation and semantic retrieval using semantic spaces. In *CBIR*, pages 359–368. ACM, 2008.

[105] D. Vaquero, R.S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *WACV*, pages 1–8. IEEE, 2009.

[106] M. Demirkus, K. Garg, and S. Guler. Automated person categorization for video surveillance using soft biometrics. In *Biometric Technology for Human Identification*, volume 12, page 54. SPIE, 2010.

[107] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367. IEEE, 2010.

[108] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, pages 1543–1550. IEEE, 2011.

[109] S. Cai, J. Wang, and L. Quan. How fashion talks: Clothing-region-based gender recognition. In *CIARP*, pages 515–523. Springer, 2014.

[110] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893. IEEE, 2005.

[111] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[112] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[113] L. Cao, M. Dikmen, Y. Fu, and T.S. Huang. Gender recognition from body. In *ACMMM*, pages 725–728. ACM, 2008.

[114] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE TPAMI*, 24(7):971–987, 2002.

[115] Wikipedia. Gabor filter — wikipedia, the free encyclopedia, 2015. [Online; accessed 11-June-2015].

[116] D. He and L. Wang. Texture unit, texture spectrum, and texture analysis. *IEEE TGRSL*, 28(4):509–512, 1990.

[117] H. Lian and B. Lu. Multi-view gender classification using local binary patterns and support vector machines. In *ISNN*, pages 202–209. Springer, 2006.

[118] I. Fogel and D. Sagi. Gabor filters as texture discriminator. *Biological cybernetics*, 61(2):103–113, 1989.

[119] B. Prosser, W. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, pages 21.1–11, 2010.

[120] W. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, pages 649–656. IEEE, 2011.

[121] M.D. Zeiler. Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[122] 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.

[123] M.D. Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in optimization and numerical analysis*, pages 51–67. Springer, 1994.

[124] R.T. Hare-Mustin and J. Marecek. The meaning of difference: Gender theory, postmodernism, and psychology. *American Psychologist*, 43(6):455–464, 1988.

[125] N. Almudhahka, M. Nixon, and J. Hare. Human face identification via comparative soft biometrics. In *ISBA*, pages 1–6. IEEE, 2016.

[126] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, pages 580–587. IEEE, 2013.

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

[128] S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D.J. Kriegman, and S. Belongie. Generalized non-metric multidimensional scaling. In *AISTATS*, pages 11–18, 2007.

[129] O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A.T. Kalai. Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*, 2011.

[130] L. Van Der Maaten and K. Weinberger. Stochastic triplet embedding. In *MLSP*, pages 1–6. IEEE, 2012.

[131] E. Amid and A. Ukkonen. Multiview triplet embedding: Learning attributes in multiple maps. In *ICML*, pages 1472–1480, 2015.

[132] Ç. Demiralp, M.S. Bernstein, and J. Heer. Learning perceptual kernels for visualization design. *TVCG*, 20(12):1933–1942, 2014.

[133] T.A. Bijmolt and M. Wedel. The effects of alternative methods of collecting similarity data for multidimensional scaling. *International Journal of Research in Marketing*, 12(4):363–371, 1995.

[134] A. Tversky. Features of similarity. *Psychological review*, 84(4):327–352, 1977.

[135] D.L. Medin, R.L. Goldstone, and D. Gentner. Respects for similarity. *Psychological review*, 100(2):254–278, 1993.

[136] C.L. Krumhansl. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, 85(5):445–463, 1978.

[137] R.N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987.

[138] J.H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[139] G.W. Leibniz. Discourse on metaphysics. In *Philosophical papers and letters*, pages 303–330. Springer, 1989.

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

[141] F. Husson, J. Josse, and J. Pages. Principal component methods-hierarchical clustering-partitional clustering: why would we need to choose for visualizing data. *Applied Mathematics Department, AgroCampus*, 2010.

[142] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE, 2016.

[143] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[144] A. Dubey, N. Naik, D. Parikh, R. Raskar, and C.A. Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *ECCV*, pages 196–212. Springer, 2016.

[145] How efit-v works. Technical report, VisionMetric Ltd, 2012. http://www.visionmetric.com/products/about-efit-v/how-efit-v-works/.