On Categorising Gender in Surveillance Imagery

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

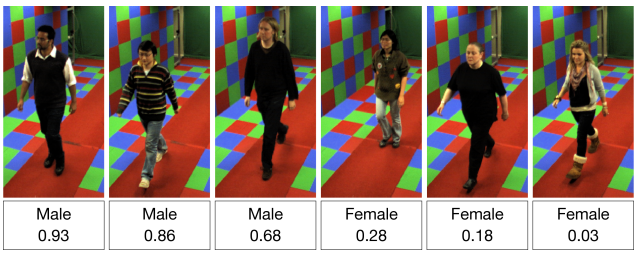
Abstract

Categorising gender for soft biometric recognition is especially challenging from low quality surveillance footage. Our novel approach discovers super fine-grained visual taxonomies of gender from pairwise similarity comparisons, annotated via crowdsourcing. This paper presents our techniques for collection, interpretation and clustering of perceived visual similarities, and discusses the transition from pre-defined categorisation to similarity comparisons between subjects. We compare and evaluate our proposal on two diverse datasets, demonstrating the ability to describe multiple concepts, including ambiguity and uncertainty, that go beyond binary male-female designators. Our method is applicable to a wide range of soft biometric traits and image attributes, and can aid in efficiently annotating large-scale datasets, by generating more discriminative, reproducible and flexible categorical labels.

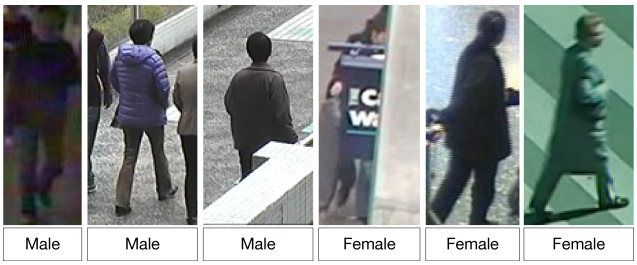
1. Introduction

Describing pedestrian images remains a significant challenge in video surveillance. Society desperately requires a means to search for pedestrians in video footage, matching human descriptions attained from eye-witness testimonies. Soft biometrics are human describable attributes, designed to consistently and precisely describe and identify people in surveillance footage [18, 3]. They can be perceived at-a-distance, in partially obscured, occluded and very low quality surveillance footage and when hard biometrics e.g. fingerprint, iris or gait, are inapplicable.

This paper concerns the annotation of gender-from-body as a soft biometric from surveillance images of pedestrians, where hard biometrics are unavailable. Gender identity is a universal trait and the most commonly predicted human attribute, typically categorised as 'male' or 'female' [4, 21, 6]. As a contemporary topic, gender has been shown to be perceived on a sliding scale [14, 8] and more complex representations of gender are becoming widespread. For example, services like Facebook now offer 71 gender options in the UK[1], and a number of universities in the US now accept non-binary pronouns beyond 'he' and 'she'[2].



(a) SoBiR binary (top) and relative (bottom) labels [15] (cropped).



(b) PETA binary labels [4].

Figure 1: (a) Labelling ambiguous genders continuously. (b) Images in which gender is hard-to-see or uncertain.

Problem. State-of-the-art recognition of gender-from-body performs significantly worse than recognition from face [17, 3, 18] and even human performance is often far from perfect [6, 12]. This suggests binary categorisation may not always be suitable, especially when dealing with challenging surveillance imagery.

Visually discerning gender (as opposed to biological sex) is dependent on the observer interpreting 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 visible, and unfamiliar combinations can be contradictory. Figure 1a high-

[1] https://www.facebook.com/facebookdiversity/posts/774221582674346
[2] http://www.bbc.co.uk/news/magazine-34901704

lights varying degrees of masculinity to femininity in clear images, while Figure 1b illustrates the difficulty in perceiving gender in real-world surveillance footage, although both image sets are originally labelled with binary ground-truths.

Comparative annotation has been successfully applied to soft biometrics [20, 9] and most recently gender [14, 15, 1]. However, previous approaches cannot distinguish between attribute ambiguity (Figure 1a) and image obscurity (Figure 1b), while pairwise comparisons for relative labels are difficult to scale. To meet the demand of ever larger datasets, highly discriminative labelling solutions must also consider efficiency, reproducibility and flexibility.

Proposal. We propose to annotate gender-from-body for pedestrian images using *pairwise similarity comparisons*, collected via crowdsourcing. 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. A *super fine-grained* visual taxonomy is then discovered by clustering subject similarities. 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 and provide comparisons to their original labels. The first dataset, SoBiR [15], is a gender-balanced dataset, including four forms of categorical and relative soft biometric labels, describing subject images captured in a controlled environment. The second dataset, PETA [4], is the largest and most diverse pedestrian re-identification dataset, annotated as 62.9% 'male' alongside 60 additional binary attributes.

We are primarily motivated by comparative soft biometrics [20, 14, 1, 9], but our work is also analogous to [24, 7], which investigate similarity comparisons via crowdsourcing, with several important distinctions. Firstly, both [24, 7] collect overall similarity annotations from a wide range of image subject matter, finding broad, basic-level categories. Instead, this study discovers super fine-grained visual concepts within a specific attribute 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 [7] or matching a subset to a query image [24], we explicitly annotate each image pair, such that no pairwise comparison can be overlooked.

Contributions. (1) The introduction of comparative similarity annotations to describe gender, applicable to ambiguous, uncertain and multi-concept attributes. (2) A discussion of our crowdsourcing methodology and novel spatial interpretation of pairwise distances and uncertainties. (3) A comprehensive evaluation on two diverse datasets, demonstrating categorisation with more than two classes of gender.

2. Related Work

Gender recognition. Two recent soft biometric surveys [18, 3] and a computer vision survey [17] analyse the demographic estimation of gender. These reveal significantly lower performance when classifying gender-from-body in uncontrolled environments, over typically more constrained facial images. An early study in gender recognition from faces [6] demonstrated an 8.1% error rate, compared to 11.6% averaged by humans. The approach labels gender as binary, but also includes a binary 'certainty' label and suggests a 'special category' to permit outliers to be correctly classified. Furthermore, the first attribute-based pedestrian re-identification study [12], found 4.5% annotator disagreement for the 'male' attribute on the PRID dataset. Almost all contemporary re-identification studies classify gender as binary, achieving accuracies up to 89.9% [4, 13, 26, 5].

Relative ordered comparisons. Relative attributes were introduced in [19] and first applied to soft biometrics in [20], using relative *ordered* annotations of psychologically grounded global and body traits. Comparative soft biometrics have since been employed to describe clothing [9], face [1] and for the first time, gender [14], finding ambiguity in even clear images. Therefore, gender uncertainty and ambiguity concepts are supported in both early [6] and recent [14] works, and are evident visually in Figures 1, 3 and 5.

Ordered comparisons seek more *objective* annotations by comparing two subjects on a bi-polar scale, e.g. "more feminine / more masculine". These approaches produce high quality, relative labels, that outperform traditional categorical labels for biometric recognition and automatic retrieval from body [9, 15]. However, predefined scales and one-dimensional ranking interpretations are not able to concisely describe multiple concepts or visual uncertainty, and remain difficult to scale for very large dataset annotation.

Similarity comparisons. Similarity is a hotly debated topic in psychology, being argued as the composition of features [23], represented as a dynamic cognitive process [16] and modelled as geometric distances [10, 22, 2]. In computer vision, distance metrics are learnt to match images based on similarity and attribute simile classifiers have been show to outperform binary classifiers [11]. Overall image similarity comparisons have also been crowdsourced to discover basic-level categories [24, 7]. These approaches discover a continuous embedded similarity space, that is richer and more versatile than a fixed size vocabulary (predefined categories) or one-dimensional ranking (relative attributes).

According to [8], 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

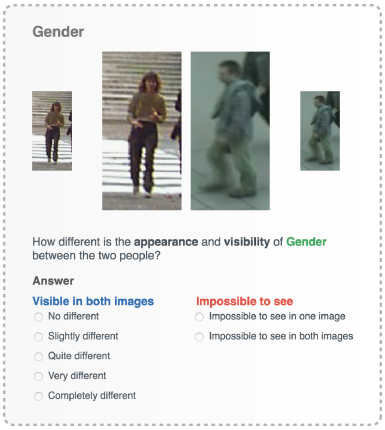


Figure 2: Example crowdsourcing task question.

is almost always represented as a sole feature. Collectively, this suggests that while gender is hard to decompose, it can be discerned through *(dis)similarity*. To our knowledge, no other similarity-based system discovers concepts within an attribute or deals with very low-quality and highly subjective images of homogeneous subject matter.

3. Crowdsourcing Data Collection

We designed a crowdsourcing task on Crowdflower[3] to collect $\binom{N}{2}$ pairwise comparisons from both SoBiR, $N = 100$ and PETA, $N = 95$ datasets. As pairwise labelling is of $\mathcal{O}(n^2)$ space, we annotate a representative subset of just 1% of PETA's 8709 total unique subjects. From this subset, a suitable visual taxonomy can be generated, enabling 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 [7] or matching to a query image [24].

Although similarity is commonly interpreted geometrically [22, 10, 2], it has been shown to be asymmetric when judging "subject A to subject B" [23]. 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 [8] and is more succinct in describing subtle variations of an attribute.

The crowdsourcing task is designed similarly to [14], asking respondents to judge the difference in both *appearance* and *visibility* of gender, as in Figure 2. 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).

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

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

Table 1: Encoding difference annotations to similarity s_{ij}, and uncertainty u_{ij}, between subjects i and j.

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.

4. Spatial Interpretation

We construct a Euclidean *perceptual similarity space*, comprised of $K = \binom{N}{2}$ comparisons of similarity s_{ij} and uncertainty u_{ij}, between all subjects $i, j \in 1, ..., N$. Table 1 details the annotation encoding of similarity as a linear interval scale and uncertainty as a binary value. Uncertainty measures, $0 \leq u'_i \leq 1$, are calculated per subject, as the fraction of all "Impossible to see.." annotations:

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

Similarity comparisons may then be mapped to geometric dissimilarity distances using an appropriate monotonic metric. We opt for exponential decay following [22, 10]:

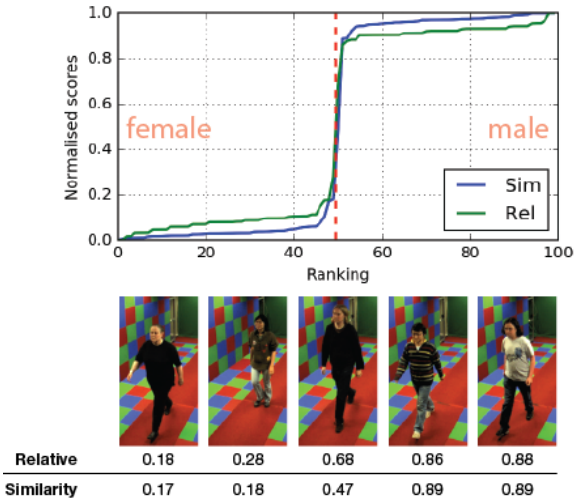
$$d(s_{ij}) = e^{-\lambda s_{ij}},$$

where λ is decay rate. Setting $\lambda \gg 1$ represents changes in difference more evenly, approximating a linear function, while $\lambda \ll 1$ emphasises spatial separations between concepts in the perceptual space. Next, we define an uncertainty weighting function between u'_i and u'_j as follows:

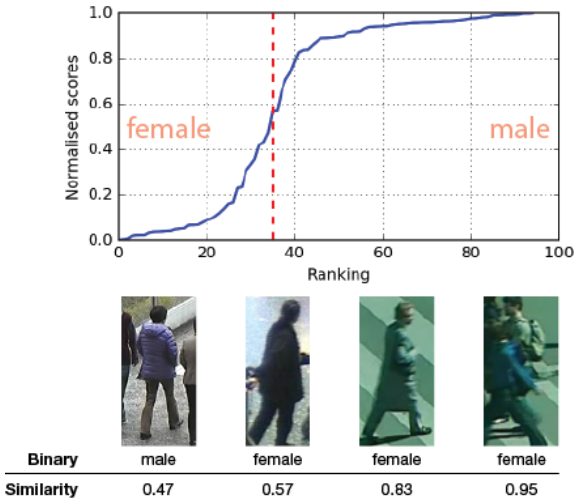
$$w(u'_i, u'_j) = (|u'_i + u'_j|/2)^{\epsilon},$$

where $0 < \epsilon \leq 1$ is eccentricity. Setting $\epsilon \approx 1$ represents ambiguity and uncertainty more equally, while $\epsilon \ll 1$ accentuates distances between subjects with different uncertainties. A distance matrix $\Delta \in N \times N$, is constructed as the dissimilarity distance between subjects i and j:

$$\Delta_{ij} = |u'_i - u'_j| d(0) w(u'_i, u'_j) + d(s_{ij})(1 - w(u'_i, u'_j)).$$



(a) SoBiR original relative (Rel) [14] and new similarity (Sim) gender scores, displaying the most ambiguous subjects.



(b) PETA original binary [4] and new similarity gender scores, displaying subjects with conflicting measures.

Figure 3: One-dimensional similarity ranking of SoBiR and PETA subjects using MDS. Dotted red lines indicate female-male split.

We can now apply a number of data exploration techniques to our spatial interpretation. We use two forms of dimensionality reduction, MDS to visualise the perceptual space, and AHC to discover visual taxonomies.

Multi-dimensional scaling (MDS). MDS is a method for representing dissimilarity measurements as distances between points, primarily applied to find structures in psychology [2]. Given a distance matrix, Δ, MDS attempts to find a lower-dimensional embedding of N vectors with D dimensions, $\mathbf{v}_1, ..., \mathbf{v}_N \in \mathbb{R}^D$, such that, $||\mathbf{v}_i - \mathbf{v}_j|| \approx \Delta_{ij}$. Since we supply a Euclidean distance matrix, we use an efficient, classical MDS approach that derives a coordinate matrix via eigenvalue decomposition.

Agglomerative hierarchical clustering (AHC). AHC is a procedure that forms hierarchical groupings of mutually exclusive data subsets [25]. 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 each subject starts in its own cluster and pairs of clusters are iteratively merged up the hierarchy. We use AHC with the popular Ward's linkage criteria to merge clusters [25], minimising within-cluster variance by reducing squared distances from each cluster centre.

5. Experiments

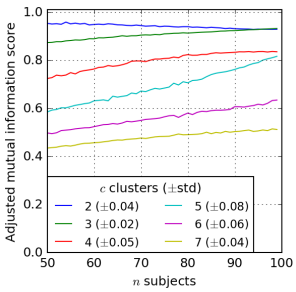
We perform three experiments, to evaluate the characteristics of our similarity data, demonstrate its representational flexibility and compare it to previous approaches. First, we investigate one-dimensional ranking, replicating a relative attributes-based approach. Next, we apply clustering to discover visual taxonomies from randomly sampled subsets of each dataset. Clusters are analysed against binary ground-truths and for their consistency across sampled subsets. Finally, we present an example visual taxonomy of the PETA dataset. Distance matrices are computed with $\lambda = 10$ and $\epsilon = 0.7$, found by a parameter grid search to maximise the overall AMI scores in Sections 5.2 and 5.3.

5.1. One-dimensional Ranking

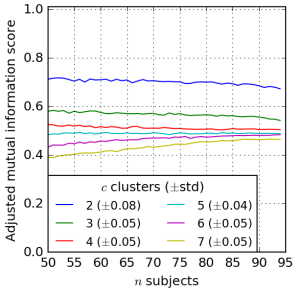
We compare a one-dimensional representation of our gender similarity to SoBiR's *relative continuous* gender labels and PETA's original binary ground-truth labels in Figure 3. The dotted red lines indicate where a binary female-male split would normally occur around scores of 0.5. One-dimensional embeddings are found by applying MDS with $D = 1$ to the distance matrix Δ, producing subject similarity scores and associated ranks.

In Figure 3a, we observe highly analogous relative and similarity scores of 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 $\rho = 0.84$, $p < e^{-26}$. 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 an ordered comparison approach [14].

Figure 3b visualises the crowd's gender perception from a subset of 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 highlighted.

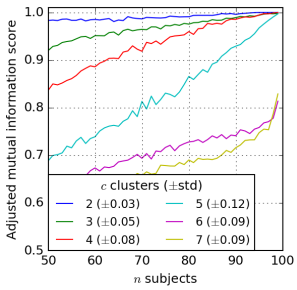


(a) SoBiR.

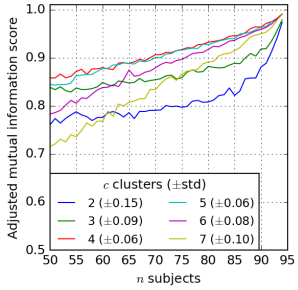


(b) PETA.

Figure 4: Agreement (AMI score) between original binary partitions and c clusters of n randomly sampled subjects.



(a) SoBiR.



(b) PETA.

Figure 5: Agreement (AMI score) between all N subjects and n randomly sampled subjects with c clusters.

5.2. Clustering Agreement to Binary Ground-truths

We investigate the agreement between clustering the perceived gender similarities and previously annotated binary ground-truths. Agreement is measured as an 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. We uniformly randomly sample a subset of n subjects and apply AHC to form c clusters. AMI scores are averaged over 500 iterations per n. As expected, distinct behaviours are observed between SoBiR and PETA, due to disparities in image clarity and demographic distributions.

Clustering the similarity data from SoBiR's clear images with $c = 2$ closely agrees with the original binary ground-truth labels, Figure 4a. At $n \approx N$ we see $c = 2$ and $c = 3$ AMI scores converging, suggesting a third cluster may describe small inconsistencies between the two original partitions. As expected, clustered similarity data from PETA agrees much less closely to the original annotations, even at $c = 2$, Figure 4b. This indicates that an increased image obscurity also increases the disparity between perceived gender similarities and original binary labels.

5.3. Clustering Consistency and Visualisation

To discover consistent visual taxonomies, we apply AHC to distance matrices computed from data subsets. Figure 5 shows the partitioning agreement between all N subjects and n uniformly randomly sampled subjects, clustered into c sets. AMI scores are averaged over 500 iterations per n.

For SoBiR, $c = 2$ is very consistent, with almost perfect agreement and low deviation, in Figure 5a. Clusterings of $c = 3, 4$ also converge to AMI scores of 1 at $n \approx N$. Though less consistent, $c > 2$ may be desirable for increased discrimination. For PETA however, we observe that $c = 2, 3$ are inadequate at describing the perceived gender similarities, being the least consistent, in Figure 5b. Instead, we find that $c = 4, 5$ groups provide more consistently reproducible and discriminative partitions.

Figure 6 displays an example visual taxonomy of PETA, with $c = 5$ clusters. Visual groups largely match the original label concepts of 'male' and 'female' and our labelling of 'uncertainty'. For discussion, we attach semantic language descriptions to each visual category. However, for large-scale annotation, categories would be better defined as related exemplar images. Although partitioning with $c = 5$ only attains an AMI score of 0.45 to the original binary labels, on visual inspection, intra-group images are highly similar and clearly correspond to our language descriptions. Interestingly, group 3 comprises just one subject, annotated confidently by respondents but contradictory to other subject images, forming its own 'neutral' group. Alternatively, setting $c = 4$ merges the 'neutral' and 'female' groups, as both possess low uncertainties.

6. Conclusions

We have introduced an approach for discovering super find-grained taxonomies of gender-from-body in challenging surveillance imagery. Our methodology crowdsources and interprets open-ended pairwise similarity comparisons, demonstrating that fixed binary categories are insufficient when labelling gender from very low quality pedestrian images. As in [6], we advocate the inclusion of 'neutral' and 'uncertain' categories for studies in human perception and demographic estimation.

Generating a visual taxonomy enables efficient and refined annotation of large datasets, and is highly applicable to other soft biometric traits and image attributes. Our spatial interpretation is also more flexible than one-dimensional ranking, permitting alternative label representations e.g. through fuzzy clustering and multi-dimensional ranking. Future work could demonstrate very large-scale annotation of super fine-grained attributes and investigate optimal representations for automatic computer vision estimation.

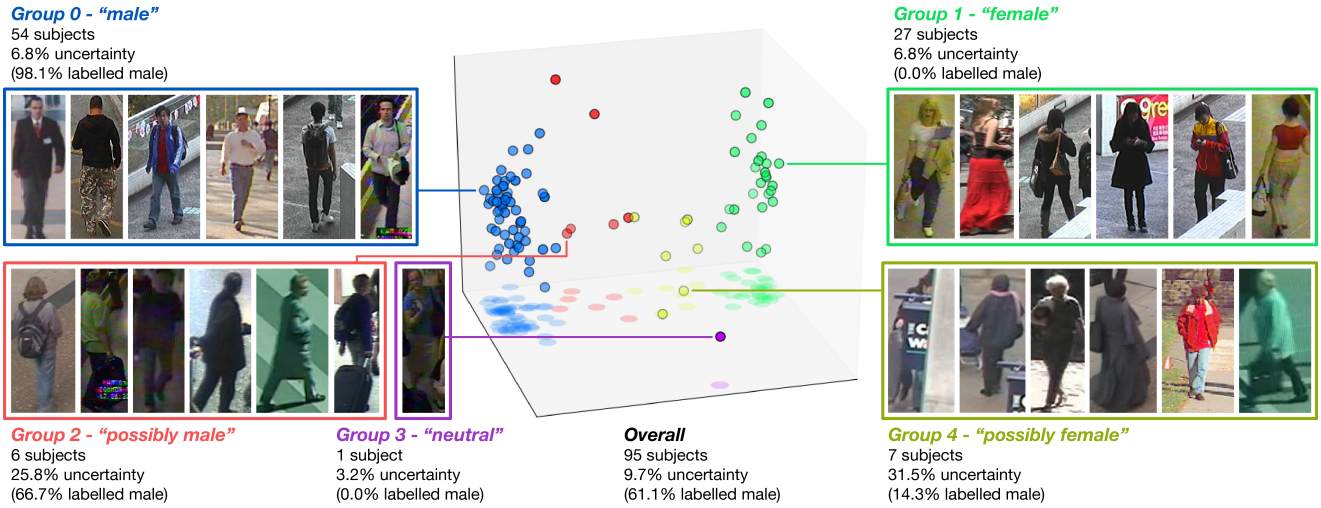


Figure 6: Example visual taxonomy of gender similarities on the PETA [4] dataset, formed with $c = 5$ clusters and visualised with 3-dimensional MDS embedding. Including group membership, average uncertainty and original binary ground-truths (in brackets).

References

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

[2] I. Borg and P. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[3] A. Dantcheva, P. Elia, and R. Arun. What else does your biometric data reveal? A survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, 11(3):441 – 467, Sept. 2015.

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

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

[6] B. Golomb, D. Lawrence, and T. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *NIPS*, 1990.

[7] R. Gomes, P. Welinder, A. Krause, and P. Perona. Crowdclustering. In *Advances in neural information processing systems*, pages 558–566, 2011.

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

[9] E. Jaha and M. Nixon. Viewpoint invariant subject retrieval via soft clothing biometrics. In *ICB*. IEEE, 2015.

[10] C. Krumhansl. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, 1978.

[11] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *CVPR*. IEEE, 2009.

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

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

[14] D. Martinho-Corbishley, M. Nixon, and J. Carter. Soft biometric recognition from comparative crowdsourced annotations. In *ICDP*. IEEE, 2015.

[15] D. Martinho-Corbishley, M. Nixon, and J. Carter. Soft biometric retrieval to describe and identify surveillance images. In *ISBA*. IEEE, 2016.

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

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

[18] M. Nixon, P. Correia, K. Nasrollahi, T. Moeslund, A. Hadid, and M. Tistarelli. On soft biometrics. *Pattern Recognition Letters*, 2015.

[19] D. Parikh and K. Grauman. Relative attributes. In *ICCV*. IEEE, 2011.

[20] D. Reid, M. Nixon, and S. Stevenage. Soft biometrics; human identification using comparative descriptions. *TPAMI*, 36(6):1216–1228, 2014.

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

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

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

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

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

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