Soft Biometric Recognition from Comparative Crowdsourced Annotations

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

Vision Learning and Control,
School of Electronics and Computer Science,
University of Southampton, UK,
{dmc, msn, jnc}@ecs.soton.ac.uk

Keywords: Soft biometrics, crowdsourcing, semantic recognition, surveillance search

Abstract

Soft biometrics provide cues that enable human identification from low quality video surveillance footage. This paper discusses a new crowdsourced dataset, collecting comparative soft biometric annotations from a rich set of human annotators. We now include gender as a comparative trait, and find comparative labels are more objective and obtain more accurate measurements than previous categorical labels. Using our pragmatic dataset, we perform semantic recognition by inferring relative biometric signatures. This demonstrates a practical scenario, reproducing responses from a video surveillance operator searching for an individual. The experiment is guaranteed to return the correct match in the the top 7% of results with 10 comparisons, or top 13% of results using just 5 sets of subject comparisons.

1 Introduction

Biometrics are distinguish and identify human features, providing information with which to perform automatic human recognition [1]. However, using traditional biometrics to perform pedestrian identification from video surveillance footage is still a largely unsolved topic. Retrieving identity cues from a limited number of low quality images proves challenging when applied in unconstrained environments at-a-distance. Due to these constraints, common biometrics like face, fingerprint or gait are often partially hidden or unobservable.

Soft biometrics are a new form of biometric that fill in these information gaps, as they rely only on human perception and description to systematically label subjects. They have been shown to be objective, salient, reliable and robust to changes in distance [2, 3].

The power of soft biometrics lies in their ability to bridge the semantic gap between high-level human description and low-level biometric features generated from images [3]. This opens up a considerable number of opportunities, such as Content Based Image Retrieval (CBIR) and human accessible search queries based only on verbal descriptions. Such methods are capable of addressing the limitations of conventional monitoring systems, where human operators are required to comb vast archives of recorded material for forensic investigations.

This paper moves towards solving this complex problem, by investigating the potential of soft biometrics for human description. Specifically, using comparative, global soft biometric descriptors, annotated via crowdsourcing. From these annotations, precise relative subject signatures are generated to facilitate accurate semantic recognition. Later work will focus on automatically predicting labels using computer vision and machine learning techniques.

Crowdsourcing enables the collection of data from globally diverse annotators. This provides the best opportunity to remove cultural annotation bias, while producing innovative ground-truth information. Crowdsourcing also simulates a working environment, whereby variations in descriptive responses can be analysed to provide better search queries for video surveillance investigations. Our contributions are threefold:

- To provide a comprehensive, public dataset[1] of 59400 unique crowdsourced comparative human annotations detailing 100 subjects through 12 global soft biometric traits.
- To provide insight into crowdsourcing methodologies that utilise genuine human responses, to form high quality annotations.
- To demonstrate semantic recognition in a surveillance scenario, by modelling the search queries of a surveillance operative using only a limited number of comparative judgments.

The paper is organised as follows: Section 2 explores related literature. Section 3 describes the original image dataset. Section 4 answers what attributes should be annotated, how they should be crowdsourced and analyses the results. Section 5 establishes the ranking inference process. Section 6 describes and analyses the relative semantic recognition experiment. Finally, Section 7 reiterates our findings and future plans.

[1] `http://users.ecs.soton.ac.uk/dmc1g14/#icdp-2015`

2 Related work

Many studies focus on biometric fusion techniques to perform identification, by combining *ancillary* soft biometric information with traditional *hard biometrics* like gait [4, 3], face [2, 5] or fingerprints [6, 7]. However, these systems present a serious limitation for practical surveillance systems - subjects must first be enrolled into the system in order to match a known hard biometric signature. Furthermore, hard biometrics are likely to be unobservable or occluded in many CCTV images.

Reliance on soft biometrics means subjects need not be pre-enrolled in a system, as identification is performed through human description alone. This is a compelling premise for our work; to investigate the power of standalone soft biometrics in performing identification, showing they provide more than just subsidiary information.

Earlier work on soft biometrics described subjects through an absolute semantic space, using categorical labels [4, 8, 9]. Later work proposed the use of comparative measurements, able to predict relative attribute strengths of faces and natural scenes [10] and texture [11]. Reid et al. presents a psychologically grounded justification for using comparative soft biometric descriptors and performs accurate retrieval of subjects using the Elo rating system [3]. The study reveals that comparative labels are more objective soft trait measures, over often unreliable conventional absolute labels. By asking for comparative annotations between two subjects, the affects of *individual human bias* are mitigated. Comparative measurements also allow continuous relative measurements to be inferred, improving subject recognition. Adjeroh et al. investigates the correlations and predictability of human metrology, providing extra information for identification at-a-distance using soft biometrics [12].

Another topic area to receive much attention is re-identification; matching individuals across multi-camera networks. Earlier approaches deal with low-level, appearance-based matching methods in the visual space [13, 14, 15]. However, there is a growing trend to solve re-identification using human describable attributes. More recent studies are moving towards using mid-level and high-level semantic attributes [16, 17, 18], some of which discuss zero-shot identification [10, 19]. Therefore, the question is not *if*, but *how* such information can be discerned and utilised.

Furthermore, it has been shown that traits like gender [20, 21, 22], height and colour [23] and demographics like age and race [24] can be automatically estimated successfully from body images. By combining these concepts, it will be possible to automatically generate relative soft biometric labels from images and perform content based image retrieval, therefore enabling automatic human identification. In practice, an automated process will radically cut the time spent manually searching large network of low quality footage.

Traits	**Polar labels** `pole A`	`pole B`
Gender	Feminin	Masculine
Age	Old	Young
Height	Tall	Short
Weight	Heavy	Light
Figure	Fat	Thin
Chest size	Big	Small
Arm thickness	Thick	Thin
Leg thickness	Thick	Thin
Skin colour	Dark	Light
Hair colour	Dark	Light
Hair length	Long	Short
Muscle build	Muscle	Lean

Table 1: Lexicon of traits and their polar labels.

3 Multi-biometric tunnel dataset

To simulate an idealised surveillance environment, our original dataset consists of images from the University of Southampton multi-biometric tunnel dataset [25].

From this we extracted a gender balanced dataset consisting of 100 subject images aligned to a similar position along the tunnel, via a single forward-facing camera and cropped to equal size (Figure 1).

The dataset records many other camera viewpoints at the same time, which will allow future extensions of this work to investigate view-invariant approaches. Categorical annotations were also provided by [4], with which we can compare to our newly annotated relative labels.

4 Crowdsourcing task

In this section we detail the design decisions made when building the crowdsourcing task that led to the large collection of high quality comparative annotations. We used the CrowdFlower[2] platform to build and run the crowdsourced annotation task. The platform provides comprehensive data analysis and quality control tools, allowing customers to accept a range of responses while rejecting non-genuine answers. It also connects to global pools of contributors, therefore unambiguous and decisive questions must be presented.

We would ideally like to improve upon the crowdsourcing work of [24], who spent a significant sum of money collecting a large number of human intelligence tasks (HITs), only to gain few valid responses. Additionally, the goal is to collect geographically unconstrained data to better model average human perception and description of others, compared to more isolated annotation tasks like [3].

4.1 Trait and label derivation

MacLeod et al. set out the first system to record body attributes, founded on psychological observations of perception and memory [26]. It concludes that more research must be done to understand our own value judgments of others, which both [3, 4] go someway to answering. By

[2] `http://www.crowdflower.com/`

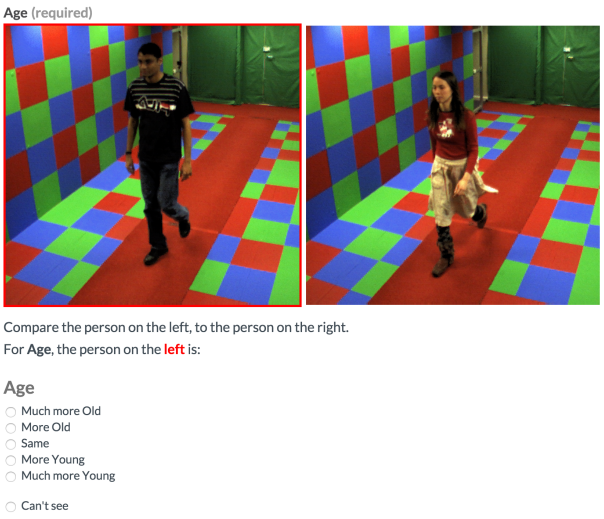
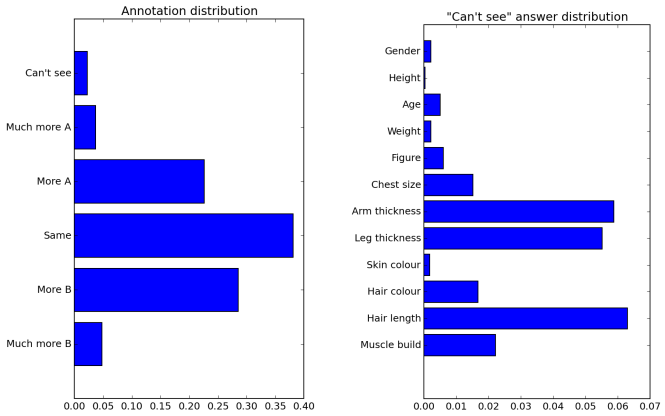


Figure 1: Screenshot of one annotation task question.



(a) (b)

Figure 2: (a) Overall annotation response distribution. (b) Annotation uncertainty distribution per trait.

reviewing the most significant, prevalent and stable features from [3, 4], the final soft biometric trait lexicon of 12 soft traits was deduced, seen in Table 1. The trait and label nomenclature was simplified for a global crowdsourcing audience. Although at times this can appear crude, the intention is always kept clear.

Lucas et al. argue against separating by ethnicity, stating that it is often misinterpreted when describing low quality images [27]. Although distinctive in some cases, there is also no obvious way to represent ethnicity through a single set of binary polar labels.

Finally, it is important to note that gender is collected as a comparative trait. As far as we know, this is the first time gender has been measured in this way on such a scale, being most commonly described in a binary fashion.

4.2 Question and response design

Each annotation question is essentially a psychometric procedure, whereby the respondent is shown two stimuli images and asked to compare the one of the left, to the one on the right, for the 12 traits defined in our new lexicon.

In total $12 \times \binom{100}{2}$ unique annotations were asked, comparing every pair of subjects for each trait. A 5 point answer scale was used for all annotations as in [3, 4, 26], following a consistent format: "Much more `A`", "More `A`", "Same", "More `B`", "Much more `B`".

Reid et al. collected a 'certainty' rating for each annotation [3], but this is too time consuming for crowdsourcing respondents, who are looking to be paid. Instead, an additional "Can't see" option was provided as an acceptable response for hard to distinguish questions. This is very important, as it reduces the chance of collecting feigned and inaccurate responses.

The crowdsourcing platform has the ability to predefine test questions, to measure respondents' accuracy and minimise the number of spurious responses. Respondents were allowed to answer up to 20 pages, each with 10 annotations. The first page consisted totally of test questions, which must be passed in order to proceed as a *trusted respondent* (and be paid), subsequent pages contained 1 test annotation and 9 unique annotations.

Several subsets of questions were trialled, to measure the acceptability of the predefined test questions. To make test questions fair, they were sampled from more obvious comparisons and only the most fundamentally incorrect responses were rejected. However, respondents were required to exceed 80% correct to proceed.

"Can't see" was marked as an acceptable response for all annotations, but capped at a maximum response rate of 20%. 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 [3]. Figure 1 illustrates final question layout and accompanying text.

4.3 Crowdsourced annotation analysis

The annotation task concluded with 59400 unique annotations collected from 892 trusted respondents (124 untrusted respondents were flagged, and 4383 responses rejected). The final task cost, including trail runs, was only $303. Clear instructional text and objective test questions meant our task was much more economic compared to Han et al.'s study, that spent $3000 on 112,519 HITs [24]. Furthermore, 179 respondents rated our task, giving it a favourable overall average of 4.4 out of 5.

Figure 2a details the overall annotation distribution for

the task which was well balanced. Although "Can't see" was always an acceptable response, only 2.4% of answers were marked as such. Figure 2b compares the distribution of "Can't see" responses, forming a measure of uncertainty for each trait. As expected *arm thickness* and *leg thickness* were very uncertain, being the least distinctive traits chosen from previous work [3, 4]. Interestingly, hair length was the most uncertain, due to one subject wearing a head scarf, and many others with long hair obscured by their body, due to the camera angle.

5 Semantic ranking inference

To interpret the annotated pairwise comparisons, we wish to infer the semantic *strength* for each soft biometric trait associated with every subject. Strengths, or *scores*, are measured *relatively*, meaning we can then *rank* subjects by score, forming an ordered list for each trait.

5.1 Ranking function formulation

To infer the rankings, we define a *ranking function* given a set of *pairwise constraints*. For each trait $t \in T$, we say O_t is a set of ordered images $(i, j) \in O_t$, such that image i is described to be more like `pole A` of trait t than image j and S_t is a set of similar image pairs $(i, j) \in S_t$, such that both i and j posses similar qualities for trait t. To reduce the effects of discrepancies between annotation techniques, "Much more" and "More" responses were combined for each polar label.

Our goal is to find T trait target vectors $\mathbf{r}$, such that for ordered images $(i, j) \in O_t$, $r_i > r_j$ and for similar images $(i, j) \in S_t$, $|r_i - r_j| = 0$.

Although this is an NP hard problem, a popular method for approximating the solution is to use Joachims' RankSVM [28], later extended by Parikh to support similarity constraints [10]. As with soft-margin SVMs, we introduce a slack variable ξ_{ij}, which is the ranking error between images i and j. Following the concise formulation of [11], we wish to:

$$\begin{aligned} \underset{\mathbf{r}}{\text{minimize}} \quad & \frac{1}{2}||\mathbf{r}||^2 + C\sum \xi_{ij}^2 \\ \text{subject to} \quad & r_i - r_j \geq 1 - \xi_{ij},\ (i,j) \in O_t, \\ & |r_i - r_j| \leq \xi_{ij},\ (i,j) \in S_t, \\ & \xi_{ij} \geq 0, \end{aligned} \tag{1}$$

where C is the primary RankSVM parameter, trading off between maximising the margin and satisfying the pairwise relative constraints [10]. As this is fundamentally an SVM formulation, it can later be extended to learn rankings from any given feature space, e.g. automatically generated image features.

5.2 Ranking function analysis

Figure 3 contrasts subjects' normalised score distribution against their ordered rankings, illustrating the different properties of each soft biometric trait (semantically ranked using $C = 1$).

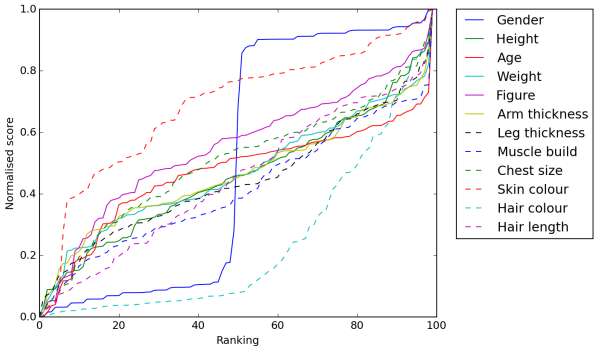


Figure 3: Relative normalised subject scores against subject ranks for each trait.

Gender comparisons produced a highly binary distribution between 'feminine' and 'masculine' polarities. When computing relative gender scores using just 4 subject-to-subject comparisons, the subjects were separated into two subsets, matching the absolute gender labels of Samangooei et al. [4], and being at least as distinguishing. However, the gender response is not a perfect step function, and there are several subjects whose gender is not as pronounced as others. The remaining labels varied between 12% to 26% compared to the equivalent absolute labels of [4], showing more linear correlations between score and rank.

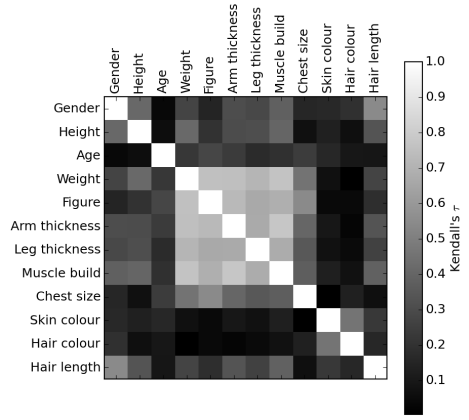


Figure 4: Kendall's tau correlation between ranked traits.

Kendall's τ coefficient is used to measure correlations between traits in Figure 4. Similarly to [12], 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 had high correlations, while *age* varies most independently.

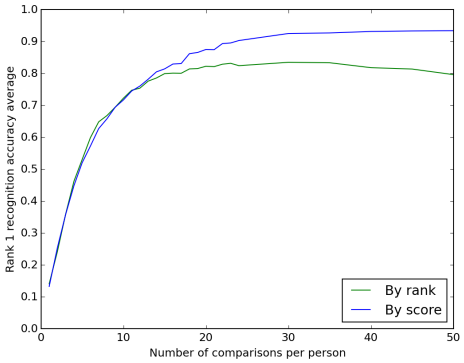


Figure 5: Average rank 1 recognition accuracy, varying the number of comparisons, n, per subject.

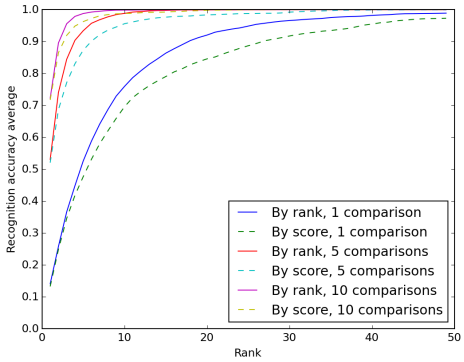


Figure 6: Average recognition accuracy for $n \in (1, 5, 10)$ comparisons per subject, while varying acceptance rank.

6 Performing semantic recognition

This section demonstrates how it is possible to perform recognition using only soft biometrics, from pre-interpreted relative scores. Biometric recognition is the process of identifying an unknown observation (the *probe* or *suspect*), by matching it to a set of known subjects (the *gallery*). This is ideally suited for forensic investigation or performing CBIR to automatically identify an individual in a video surveillance network.

6.1 Recognition 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 can simulate an eye witness testimony that compares the suspect to n known subjects. The recognition methodology is inspired by [3].

The experiment chooses the probe subject from the annotated dataset and removes n sets of randomly sampled comparisons between the probe and n other subjects. The removed comparisons are used to form the a new suspect query, inserted into the dataset. Biometric signatures are generated for each gallery subject and suspect, represented as a vector of T target values for subject i, $\mathbf{x}_i = \{r_i\}$, using the RankSVM technique described in Section 5.

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

6.2 Recognition performance analysis

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

A direct comparison to Reid et al. can be made, who performed recognition with 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 also only collected a subset of 558 annotations from 57 annotators, with the remaining comparisons synthetically inferred [3].

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. For this reason, we find expectantly lower recognition accuracies at lower n values when performing rank 1 recognition, Figure 5. Recognition rates also climb more slowly, suggesting our data includes more inconsistencies. Even so, our ranking process can still attain a maximum recognition rate of 93%, compared to Reid et al.'s 95% [3].

Score based signatures surpassed rank based signatures at higher values, attributed to traits like gender and hair colour, that have regions of similar relative scores (Figure 3). Therefore, relative scores describe the possessed quality of a trait better than ranking positions, which diverge between gallery and probe queries as n increases.

A second experiment assessed recognition accuracy while varying the acceptance rank. This reproduces a surveillance scenario, in which the operator can rapidly eliminate irrelevant subjects, leaving only the most relevant 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% recognition accuracy at rank 7. Using $n = 10$ the recognition rate actually converges faster than [3]. In these cases, rank based signatures outperform score based signatures, as increasing the acceptance rank improves cases where correct matches have small rank differences but proportionally larger score differences.

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.

7 Conclusions

We have discussed how soft biometrics provide a solution to identifying pedestrians from video surveillance footage

and how this could mitigate the limitations of conventional monitoring systems. By applying a RankSVM algorithm to interpret human comparisons, we can build precise, relative soft biometric signatures. With this technique and a small lexicon of soft traits, our experiments perform recognition almost as well as, and in some cases better than [3], using more representative crowdsourced annotations.

The publicly available dataset opens up opportunities to further explore the semantic annotation data, not only to evaluate its intrinsic properties for identification purposes, but to also better understand the variations and contradictions in human responses collected from a highly diverse population.

Future work will build on and combine the ideas presented in Section 2, [10, 11, 19]. The aim is to investigate soft biometric retrieval from a number of surveillance image datasets. By successfully predicting soft biometrics from images, we hope to accomplish automatic soft biometric identification from surveillance footage.

References

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

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

[3] D. A Reid, M. S. Nixon, and S. V. Stevenage. Soft biometrics; human identification using comparative descriptions. *PAMI, IEEE Trans*, 36(6):1216–1228, 2014.

[4] S. Samangooei, B. Guo, and M. S. Nixon. The use of semantic human description as a soft biometric. In *BTAS 2008, 2nd IEEE International Conference*, pages 1–7. IEEE, 2008.

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

[6] A. K. Jain, S. C. Dass, and K. Nandakumar. Can soft biometric traits assist user recognition? In *Defense and Security*, pages 561–572. SPIE, 2004.

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

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

[9] D. Reid, S. Samangooei, C. Chen, M. Nixon, and A. Ross. Soft biometrics for surveillance: an overview. *Machine learning: theory and applications. Elsevier*, pages 327–352, 2013.

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

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

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

[13] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 649–656. IEEE, 2011.

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

[15] B. Prosser, W. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010.

[16] Jianqing Zhu, Shengcai Liao, Zhen Lei, Dong Yi, and Stan Z Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 331–338. IEEE, 2013.

[17] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *BMVC*, volume 2, page 8, 2012.

[18] Y. Deng, P. Luo, C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *Proc. ACM International Conference on Multimedia*, pages 789–792. ACM, 2014.

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

[20] Choon Boon Ng, Yong Haur Tay, and Bok-Min Goi. Recognizing human gender in computer vision: a survey. In *PRICAI 2012: Trends in Artificial Intelligence*, pages 335–346. Springer, 2012.

[21] L. Cao, M. Dikmen, Y. Fu, and T. S. Huang. Gender recognition from body. In *Proc. 16th ACM international conference on Multimedia*, pages 725–728. ACM, 2008.

[22] G. Guo, G. Mu, and Y. Fu. Gender from body: A biologically-inspired approach with manifold learning. In *ACCV 2009*, pages 236–245. Springer, 2010.

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

[24] H. Han, C. Otto, X. Liu, and A. Jain. Demographic estimation from face images: Human vs. machine performance. *PAMI, IEEE Trans.*, PP(99):1–1, 2014.

[25] 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 2008, 2nd IEEE International Conference*, pages 1–6. IEEE, 2008.

[26] 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.

[27] T. Lucas and M. Henneberg. Comparing the face to the body, which is better for identification? *International journal of legal medicine*, pages 1–8, 2015.

[28] T. Joachims. Optimizing search engines using clickthrough data. In *Proc. 8th ACM SIGKDD International Conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.