On Semantic Soft-Biometric Labels

Sina Samangooei and Mark S. Nixon

School of Electronics and Computer Science, University of Southampton, UK

Abstract. A new approach to soft biometrics aims to use human labelling as part of the process. This is consistent with analysis of surveillance video where people might be imaged at too low resolution or quality for conventional biometrics to be deployed. In this manner, people use anatomical descriptions of subjects to achieve recognition, rather than the usual measurements of personal characteristics used in biometrics. As such the labels need careful consideration in their construction, and should demonstrate correlation consistent with known human physiology. We describe our original process for generating these labels and analyse relationships between them. This gives insight into the perspicacity of using a human labelling system for biometric purposes.

Keywords: Soft Biometrics; Human Descriptions; Retrieval; Semantic Labels

1 Introduction

Descriptions of humans based on their physical features has been explored for several purposes including medicine, eyewitness analysis [1] and human identification [2]. Descriptions gathered vary in levels of visual granularity and include features that can be measured visibly and those that are only measurable using specialised tools. To understand the recent use of labels for recognition [3, 4], we must explore the semantic terms people use to describe one another. Once these terms are outlined, the second task becomes the collection of a set of manually ascribed annotations against these terms. In isolation these terms allow the exploration of semantic descriptions as a tool for identification. To explore their capabilities in biometric fusion and automatic retrieval, these annotations must be collected against a set of individuals in an existing biometric dataset.

We developed [3] a set of key semantic terms people use to describe one another at a distance. We start with an overview of human description, from early anthropometry, to modern usage in police evidence forms and in soft biometrics. We then outline a set of key physiological traits observable at a distance and explore a set of semantic terms used for their description. The contents of the semantic annotation datasets are examined and we perform correlation analysis, exploring the underlying structures and other facets of the gathered data. We describe a study of the labels and their properties, concerning in particular information content and utility.

2 Bertillonage

One of the first attempts to systematically describe people for identification based on their physiological traits was the anthropometric system developed by Bertillon [5] in 1879. By 1809 France had abandoned early methods of criminal identification such as branding. However, no systematic method of identification was outlined as an alternative, which meant the verification of repeat offenders or confirmation of criminals' identity of was nearly impossible. Long descriptions, including semantic terms such as "Large" or "Average" to describe height and limbs, proved inadequate due to subjectivity as well as to disproportionate numbers of individuals of "Average" height and "Brown" haired. This, coupled with an uncontrolled lexicon, resulted in many descriptions which added nothing to identification process. By 1840, the photography of criminals was introduced. However, the photographic techniques themselves were not standardised and, though useful for confirmation of identity, a photograph is of little use in discovery of identity when relying on manual search. Bertillon noted the failings of the police identification and cataloguing system and developed his father's anthropological work to a more systematic method of identifying people. His system of anthropometrics, eponymously Bertillonage, outlined the tools and techniques for the careful measurement of:

- physiological features including length/width of head, lengths of certain fingers and the dimensions of the feet, arm, right ear and standing height;
- descriptions of the dimensions of the nose, eye and hair colour; and
- the description and location of notable scars, tattoos and other marks

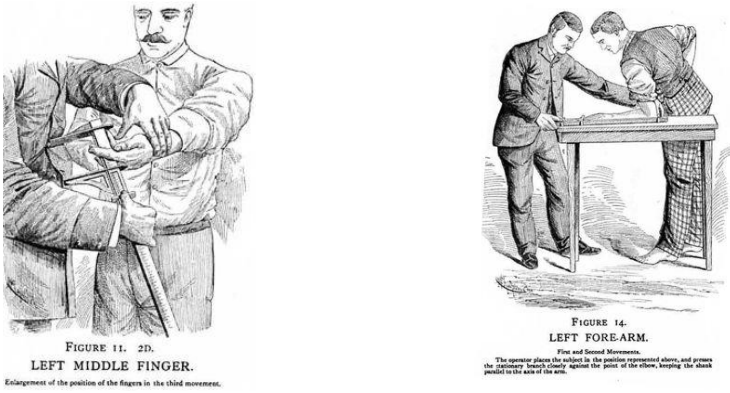

Figure 1 Examples of Bertillon's gathering of measurements [5].

The method for gathering these features was outlined in Bertillon's manual [5] along with a set of diagrams (see Fig. 1). The measurements for a given individual were held on separate slides along with standardised photographs of the individual. The metrics of the system were chosen primarily to be simple so that they could be gathered accurately. As such measurements were taken by a trained individual, though not necessarily a skilled individual. To this end, features were chosen to allow easy identification of points to begin and to end measurement. The success of

Bertillonage came from its ability to geometrically reduce the probability of type 1 errors. Though two individuals may have very similar height, the chance of the same two having similar measurements for the other features is unlikely. Furthermore, Bertillonage inherently allowed for efficient discovery of an individual's existing measurement card and therefore their identity. Cards were stored according to specific range combinations of each metric in a given order. As such that once new measurements of an unidentified individual were taken the identity of the individual could be easily ascertained.

Achieving great success and popularity in France, Bertillonage progressed to the United States as well as Great Britain in the late 19th century. Difficulties in cases such as West vs. West [6] (where Bertillonage could not reconcile differences between identical twins, though this was later disputed) led it being superseded by forms of identification such as fingerprint analysis (since the fingerprints of identical twins differ) and more recently biometric analysis. In spirit, all these systems attempt to reduce the identity of an individual to a representative and measurable set of classification metrics, though not using descriptions of the human body as a whole.

3 Data Acquisition

3.1 Traits

To match the advantages of automatic surveillance media, a primary concern is to choose traits that are discernible by humans at a distance. To do so, it is needed to determine which traits humans are able to consistently and accurately notice in each other and describe at a distance. The traits can be grouped by similar levels of meaning, namely:

- global traits (sex, ethnicity etc.)
- build features that describe the target's perceived somatotype (height, weight etc.); and
- head features, an area of the body humans pay great attention to if it is visible (hair colour, beards etc.).

With regards to global attributes, three independent traits - Age, Race and Sex – are agreed to be of primary significance in cognitive psychology with respect to human description. For gait, humans have been shown to successfully perceive such categories using generated point light experiments and in other adverse viewing conditions involving limited visual cues.

In eyewitness testimony research there is a relatively well formed notion of which features witnesses are most likely to recall when describing individuals. Koppen and Lochun [1] provide an investigation into witness descriptions in archival crime reports. Unsurprisingly, the most accurate and highly mentioned traits were Sex (95% of the respondents mentioned this and achieved 100% accuracy), Height (70% mention 52% accuracy), Race (64% mention 60% accuracy) and Skin Colour (56% mention, accuracy not discussed). Detailed head and face traits such as Eye Shape and Nose Shape are not mentioned as often and when they are mentioned, they appear to

be inaccurate. More prominent head traits such as Hair Colour and Length are mentioned more consistently. Descriptive features which are visually prominent yet less permanent (e.g. clothing) often vary with time and are of less interest than other more permanent physical traits.

Traits regarding build are of particular interest in our investigation having a clear relationship with gait while still being reliably recalled by eyewitnesses at a distance. Few studies thus far have attempted to explore build in any amount of detail beyond passing mention of Height and Weight. MacLeod et al. [7] performed a unique analysis on whole body descriptions using bipolar scales to define traits. There were two phases in their approach towards developing a set of descriptive build traits.

Firstly a broad range of useful descriptive traits was outlined with a series of experiments where a mixture of moving and stationary subjects were presented to a group of annotators who were given unlimited time to describe the individuals. A total of 1238 descriptors were extracted, of which 1041 were descriptions of overall physique and the others were descriptions of motion. These descriptors were grouped together (where synonymous) and a set of 23 traits generated, each formulated as a bipolar five-point scale.

Secondly the reliability and descriptive capability of these traits was gauged. Annotators were asked to watch video footage of subjects walking at a regular pace around a room and rate them using the 23 traits identified. The annotators were then split into two groups randomly from which two mean values were extracted for each subject for each trait. Pearson's product-moment correlation coefficient (Pearson's r) was calculated between the sets of means and was used as an estimate of the reliability for each trait. Principal Components Analysis (PCA) was also used to group traits which represented similar underlying concepts. The 13 most reliable terms, the most representative of the principal components, have been incorporated into the final trait set described later.

3.2 Terms

Having outlined the considerations made in choosing the physical traits which should be collected, the next question is how these traits should be represented. One option for their representation is a free text description for each trait. The analysis of such data would require lexical analysis to correlate words used by different annotators. Following the example of existing soft biometric techniques, a mixture of semantic categorical metrics (e.g. Ethnicity) and value metrics (e.g. Height) could be used to represent the traits. Humans are generally less accurate when making value judgements when compared to category judgements. Therefore a compromise is to formulate all traits with sets of mutually exclusive semantic terms. This approach avoids the inaccuracies of value judgments, being more representative of the categorical nature of human cognition. Simultaneously this approach avoids the complex synonymic analysis that would be required to correlate two descriptions if free text descriptions were gathered. With categorical metrics there is an inherent risk that none of the categories fit, either because the information is unclear or due to the presence of a boundary case where any annotation whatsoever may feel disingenuous.

For this purpose each trait is given the extra term “Unsure”, allowing the user to make the ambiguity known. For reasons covered in Section 4 the “Unsure” annotation is also the default option for any given trait on the annotation user interface. What remains is the selection of semantic terms that best represent the many words that could be used to describe a particular trait. This task can be logically separated by considering those traits which are intuitively describable using discrete metrics and those intuitively requiring value metrics.

4 Semantic Annotation

In this section we describe the process undertaken to gather a novel dataset of semantic annotations of individuals in an existing biometric dataset. We outline the design of the data entry system created to allow the assignment of manual annotations of physical attributes to individuals. Using this system, individuals in the Southampton Large (A) HumanID Database (HIDDB) and the new Southampton Multibiometric Tunnel Database (TunnelDB) datasets [8] were annotated against recordings taken of the individuals in lab conditions. The original purpose of these recordings was the analysis of subject gait biometrics and, in the case of TunnelDB, their face and ear biometrics. We discuss the composition of these datasets in greater detail in Section 5, here we concentrate on the procedure undertaken to assign annotations. Two systems were developed to gather annotations: The PHP based Gait Annotation system (GAnn), and later, the Python/Pylons based Python Gait Annotation system (PyGAnn).

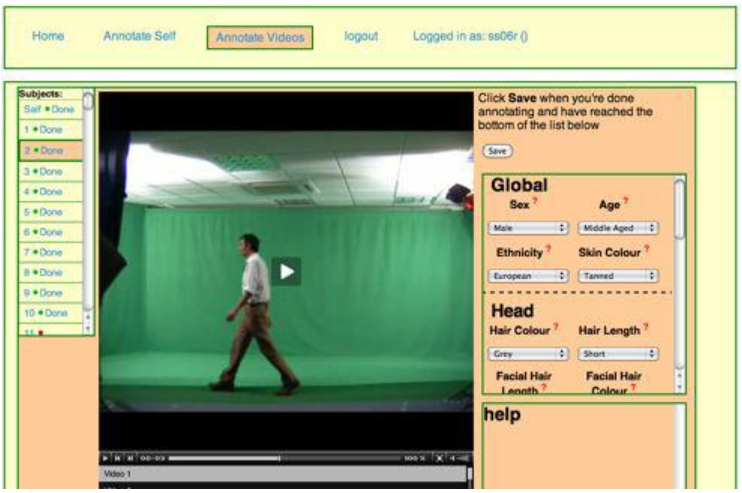

Figure 2 GAnn interface

Both systems were used to collect semantic annotations using the web interface initially designed for the GAnn web application (Fig. 2). This interface allows annotators to view all samples of an arbitrary biometric gathered from a subject as many times as they require. Annotators were asked to describe subjects by selecting

semantic terms for each physical trait. They were instructed to label every trait for every subject and that each trait should be completed with the annotator's own notions of what the trait meant. Guidelines were provided to avoid common confusions, for example that rough overlapping boundaries for different age terms and height of an individual should be assigned absolutely compared to perceived global "Average", while traits such as Arm Length could be annotated in comparison to the subject's overall physique.

To attain an upper limit for the capabilities of semantic data we strive to assure our data is of optimal quality. The annotation gathering process was designed carefully to avoid (and allow the future study of) inherent weaknesses and inaccuracies present in human generated descriptions. The error factors that the system was designed to deal with include:

• **Memory** - Passage of time may affect a witness' recall of a subject's traits.

• **Defaulting** - Features may be left out of descriptions in free recall, often not because a witness failed to remember a feature, but rather that it has a default value.

• **Observer Variables** [9] - A person's own physical features, namely their self perception and mental state, may affect recall of physical variables.

• **Anchoring** - When a person is asked a question and is initially presented with some default value or even seemingly unrelated information, replies given are often weighted around those initial values.

Body			Global	
Trait	Term		Trait	Term
0. Arm Length	(0.1) Very Short		12. Figure	(12.1) Very Thin
	(0.2) Short			(12.2) Thin
	(0.3) Average			(12.3) Average
	(0.4) Long			(12.4) Big
	(0.5) Very Long			(12.5) Very Big
2. Chest	(2.1) Very Slim		13. Age	(13.1) Infant
	(2.2) Slim			(13.2) Pre Adolescence
	(2.3) Average			(13.3) Adolescence
	(2.4) Large			(13.4) Young Adult
	(2.5) Very Large			(13.5) Adult
3. Figure	(3.1) Very Small			(13.6) Middle Aged
	(3.2) Small			(13.7) Senior
	(3.3) Average		18. Facial Hair Length	(18.1) None
	(3.4) Large			(18.2) Stubble
	(3.5) Very Large			(18.3) Moustache

Table 1 Some Semantic Traits and Labels

The semantic data gathering procedure was designed to accommodate these factors. Memory issues were addressed by allowing annotators to view videos of subjects as many times as required, allowing repeat of a particular video if necessary. Defaulting was avoided by explicitly asking individuals for each trait outlined in Table 1, this

means that even values for apparently obvious traits are filled in and captured. This style of interrogative description, where constrained responses are explicitly requested, is more complete than free-form narrative recall but may suffer from inaccuracy, though not to a significant degree. Observer variables can never be completely removed so instead we allowed the study of differing physical traits across various annotators. Users were asked to self-annotate based on self-perception, also certain subjects being annotated themselves provided annotations of other individuals. This allows for some concept of the annotator's own appearance to be taken into consideration when studying their descriptions of other subjects. Anchoring can occur at various points of the data capture process. Anchoring of terms gathered for individual traits was avoided by setting the default term of a trait to a neutral "Unsure" rather than any concept of "Average". Another potential source of anchoring is that attributed by the order subjects are presented to an annotator. A sequence of relatively tall individuals may unfairly weight the perception of an averaged sized individual as short. We aimed to account for this by randomising the order of subjects presented to different annotators. In order to use the annotations in future analysis, they were represented numerically.

5 Dataset Statistics

The Southampton Large (A) HumanID Database (HIDDB) contains between 6 and 20 sample videos of 115 individual subjects each taken from side-on; the later Southampton Multibiometric Tunnel Database (TunnelDB) contains samples of subjects for which 10 gait sample videos from between 8 to 12 viewpoints are taken simultaneously and stored to extract 3D gait information [8]. TunnelDB also contains high resolution frontal videos to extract face information and high resolution still images taken to extract ear biometrics. There are roughly 10 such sets of information gathered for each subject in TunnelDB The GAnn annotation system used to collect data against the HIDDB was designed to allow annotation by anonymous annotators across the internet, though in reality the primary source of annotations came from two separate sessions involving a class of psychology students. In the first session, all the students were asked to annotate the same group of subjects, while in the second session 4 equally sized groups of subjects were allocated between the students.

The PyGAnn annotation system used to collect data against the TunnelDB was designed to gather annotations after recording biometric signatures when annotators were asked to annotate themselves and a group of 15 subjects. Due to time constraints some annotators annotated fewer subjects but all annotators captured provided a self-annotation. We selected 4 groups of 15 subjects to be annotated by progressively few annotators, aiming to maximise the number of annotators describing the same subjects while simultaneously annotating the maximum spread of subjects. Table 2 shows a summary of the data collected. The annotations gathered are discussed in three ways:

- **Self Annotations** - Annotations an individual gave to themselves;
- **Subject Annotations** - Annotations given by an individual to a subject; and

- **Ascribed Annotations** – derived from subjects in TunnelDB who were also annotators.

		HIDDB	TunnelDB	Totals
Terms	Observed	20976	58023	78999
	Self	1659	4957	6616
	Of Annotators	0	31874	31874
Partial Descriptions	Observed	334	956	1290
	Self	10	77	87
	Of Annotators	0	544	544
Complete Descriptions	Observed	625	1685	2310
	Self	63	149	212
	Of Annotators	0	904	904
Individuals Described	Observed	115	71	186
	Self	73	226	299
	Of Annotators	0	43	43

Table 2 Summarising composition of the annotations gathered in 2 biometric datasets

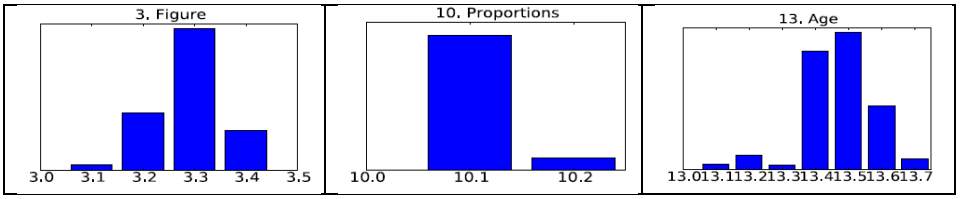

Figure 3 Example distributions of self-annotations of the TunnelDB

6 Dataset Distributions

Trait Distribution Comparison: In the datasets a total of 414 individuals were described. For the normalised distribution of self and subject annotations for all traits in both datasets. An aspect of note is the distribution of measures of physical length including Height, Leg Length and Arm Length. For both datasets ascribed lengths tend towards long and average annotations meaning annotators avoid the use of the term short. This is in contrast to measurements of thickness or bulk such as Figure, Weight, Chest and Arm/Leg Thickness which display a more normal distribution. From these graphs, Fig.3, we can also see different terms for traits such as Proportions were not used. It is possible that such traits were not perceived or the trait itself was not understood by either group of annotators, with most subjects described as having normal Proportions. Alternatively, the subjects collected may indeed portray inherently "Normal" proportions. Leg Direction seemed to enjoy similar term patterns in both datasets, a relatively unexpected result as the HIDDB did not provide the viewpoints one would expect to be necessary to make such judgements. The results for the major global features seem weighted towards Young Adult as Age; White as

Ethnicity and Male as Sex. This distribution is to be expected from the datasets as both contain many subjects from the Engineering departments of the University of Southampton, UK. Overall, we note that self-annotations taken in both systems used semantic terms in ratios comparable to those used in the ascribed annotations, as well as ratios comparable to each other. This is evidence towards the idea that individuals do not wholly believe themselves to be an average; rather individuals can reasonably describe themselves as others might see them, using the full set of semantic terms others might use. Despite this, later use of relative measurements was demonstrated to relieve problems associated with categorical labels, especially height [3].

Trait	**p-value**
Ethnicity	0.62
Hair Colour	0.7
Hair Length	0.84
Facial Hair Length	0.84
Age	0.9
Shoulder Shape	0.91
Sex	0.92
Leg Direcation	0.92

(a) ascribed annotations

Trait	**p-value**
Hair Colour 0.66	0.66
Facial Hair Length	0.66
Skin Colour	0.79
Sex	0.80
Facial Hair Colour	0.86
Ethnicity	0.87
Hair Length	0.92
Figure	0.93

(b) self-annotations

Table 3 Lowest p-values of the difference in annotations between the TunnelDB and HIDDB dataset

Cross-Dataset Distribution Comparison: In Table 3 we explore the differences in the distribution from self-annotations and ascribed annotations of the two datasets. There are small disparities between the self- annotations of HIDDB when compared to those of TunnelDB, though these are mostly insignificant differences with large p-values. The p-values in these tables represent the probability of a shared distribution having created the annotation distributions across the HIDDB and TunnelDB datasets. Two extremely similar distributions will produce p-values close to 1.0 while completely dissimilar distributions will produce p-values close to 0.

The individuals annotated were overall similarly distributed in appearance. More precisely, disparate groups of annotators described the different individuals in the different datasets using similar annotations. Some traits enjoy higher disparity between the datasets and therefore lower p-values; namely Ethnicity and associated attributes of Hair Colour. A special effort was made in the collection of TunnelDB to include individuals of different ethnic backgrounds in order to analyse ethnicity as a covariate of gait; this may explain the apparent higher degree of ethnic disparity reported by annotators of the TunnelDB. Individuals with beards were specifically chosen to be annotated in the TunnelDB due to a lack of such individuals in the HIDDB. This was performed to test the ability of the facial hair related traits to some degree. With regards to self-annotations across the two datasets, both from the graphs and the relatively lower p-values in Table 3, we note a disparity in the ratio of self-annotation of Sex. However, the graphs and p-values show comparatively similar distributions in other traits.

6.1 Internal Correlations

Having outlined the overall content and distributions of the gathered datasets in the previous sections, it is appropriate to explore notable correlations found between the various semantic annotations gathered. The goal of this section is to highlight internal structures inherent in the datasets gathered, some of which are supported by previous studies, therefore confirming the data's validity. In this section the correlation between relevant pairings of self, subject and ascribed annotations (see Section 5) are ex[lored. Though interesting for its own merits, these correlations could also have some useful practical applications. For example, by knowing the correlation between traits, estimated terms for missing traits could be inferred. This would result in more accurate results for a given incomplete semantic query, though such query competition could also be achieved through related techniques. In this section we also explore in greater detail the correlation between especially notable traits, such as Sex and Ethnicity when compared to other physical characteristics.

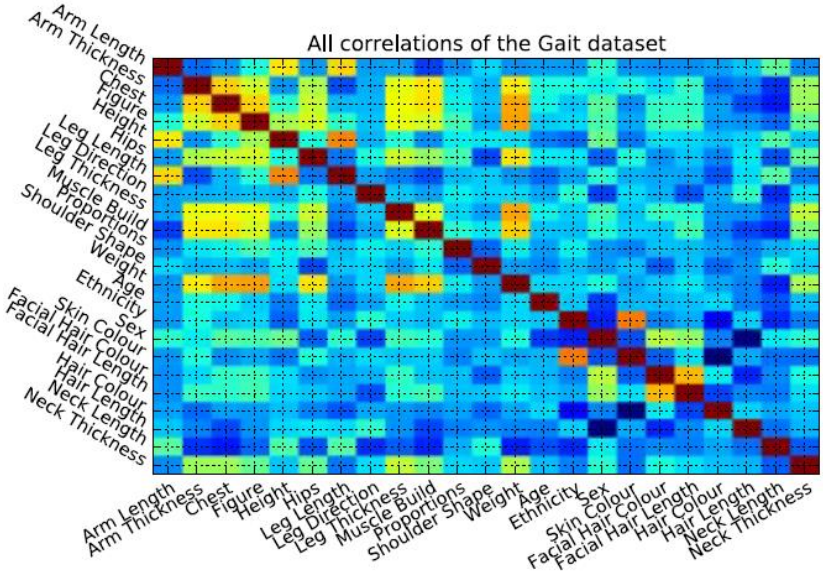

Figure 4 Term Correlations of annotations ascribed by individuals in HIDDB

The correlation matrices containing the Pearson's *r* between each term are represented graphically. Colours closer to red represent correlation coefficients closer to 1.0 and thus a positive correlation, while colours closer to blue represent correlation coefficients closer to -1.0 and thus a negative correlation. Pale green represents positive correlation.

We calculate the correlation coefficient between two terms using individual annotator responses of individual subjects. Pearson's *r* is calculated as:

$$r = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \qquad 2.1$$

where X and Y represent two semantic terms. Each semantic term was set to 1 if the annotation contains the term and 0 if the annotation did not. X_i and Y_i are the value ascribed to an individual in a single annotation, where there exist n annotations. Note that if $(X_i - \overline{X})(Y_i - \overline{Y}) > 0$ then X_i and Y_i lie on the same side of their respective means. In the binary case, where X and Y can only take the values 0 or 1, this denotes simultaneous annotation. Therefore, Pearson's r when applied to these semantic annotations is positive if X_i and Y_i are simultaneously present in an annotation. Furthermore, a higher correlation simultaneously represents how far an appearance of X or Y is from the mean, as well as the frequency of simultaneous appearances of X and Y across all n annotations.

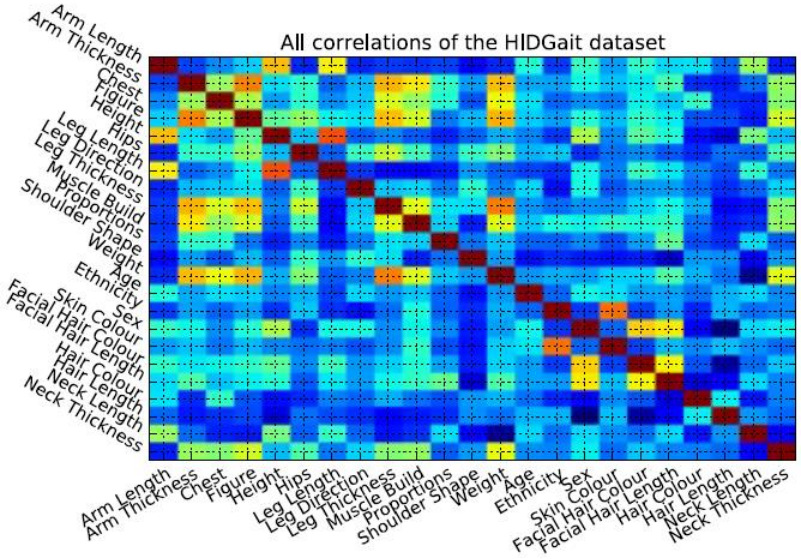

Figure 5 Term Correlations of self annotations in HIDDB

In Fig. 4 we explore the correlations between subject annotation autocorrelation, representing how often individual trait and term pairings were used by annotators. Due to its nature, in the identity of the graph we achieve a perfect correlation. This is a trivial result meaning simply that a term appeared with itself every time it was used in an annotation. More informative correlations can be seen firstly between traits 0 to 12. These are build traits whose terms describe overall thickness and length of the body, as well as extremities. We note that Figure and Weight are highly correlated. In turn they are both correlated with Arm Thickness, Leg Thickness and Chest annotations. Correlation can also be noted between Height and Leg Length, each also portraying correlations with Arm Length. We also notice some inverse correlations.

In Neck Length against Neck Thickness we see signs of thinner necks being correlated with longer necks, bulky necks with shorter necks and so on. This inverse correlation can also be noted in both Neck Length and Neck Thickness compared to other traits of bulk and length respectively, though it should be noted that these inverse relationships are not as significant. There seems to exist two groups of traits

whose terms correlate in ascending order. Namely traits denoting some notion of bulk or girth (represented by Weight, Figure etc.) and those denoting some notion of length (represented by Height and appendage lengths).

In Fig. 5 we see the auto-correlations of self-annotations. The correlations in self annotations are very similar to those found between ascribed annotations and many of the same statements with regards to build and global features can be made as above. This shows that in describing themselves that annotators are as consistent as they are when describing other people. This corresponds well with the similarity in annotations distributions noticed

7 Conclusions and Future Work

A new approach to soft biometrics aims to use human description as part of the recognition/ retrieval process. A semantic labelling system has been described and some of the properties explored. The semantic labels have been chosen with psychology in mind: the labels are those derived from human vision and attention must be paid to minimise bias introduced by the (human labelling) process. As the procedure has been design for use with surveillance video, with necessarily low resolution and quality, it would prove interesting to study the effect of these on the correlations noted here. Equally, a fruitful avenue of research might be to explore the structure revealed here, as this might enable recovery of occluded labels.

References

1. Koppen V, Lochun SK (1997) Portraying perpetrators; the validity of offender descriptions by witnesses, *Law and Human Behavior* **21**(6):662–685
2. Interpol Disaster Victim Identification Form (Ante Mortem, Yellow), 2008 http://www.interpol.int/INTERPOL-expertise/Forensics/DVI-Pages/Forms
3. Samangooei S, Guo B, Nixon MS, The use of semantic human description as a soft biometric, *Proc. IEEE BTAS'08*
4. Reid D, Nixon MS, Stevenage S (2014) Soft biometrics; human identification using comparative descriptions. *IEEE Trans. PAMI* **36**(6): 1216 - 1228
5. Bertillon A (1889) Instructions for taking descriptions for the identification of criminals and others, by means of anthropometric indications. *American Bertillon Prison Bureau*
6. Cole SA (2007) Twins, Twain, Galton, and Gilman: Fingerprinting, Individualization, Brotherhood, and Race in Pudd'nhead Wilson. *Configurations*, **15**(3): 227-265
7. MacLeod MD, Frowley JN, Shepherd JW (1994) Whole body information: its relevance to eyewitnesses, *Adult Eyewitness Testimony* CUP, Chapter **6**:125-142
8. Seely RD, Samangooei S et al (2008) The University of Southampton multi-biometric tunnel and introducing a novel 3D gait dataset. *Proc. IEEE BTAS'08*
9. Flin RH, Shepherd JW (1986) Tall stories: Eyewitnesses' ability to estimate height and weight characteristics. *Human Learning*, **5**(1):29-38
10. Samangooei S, Nixon MS (2010) Performing content-based retrieval of humans using gait biometrics. *Multimedia Tools and Applications* **49**:195-212