Performing Content-Based Retrieval of Humans Using Gait Biometrics

Sina Samangooei and Mark S. Nixon

School of Electronics and Computer Science, Southampton University, Southampton, SO17 1BJ, United Kingdom
{ss06r,msn}@ecs.soton.ac.uk

Abstract. In order to analyse surveillance video, we need to efficiently explore large datasets containing videos of walking humans. At survei llance-image resolution, the human walk (their gait) can be determined automatically, and more readily than other features such as the face. Effective analysis of such data relies on retrieval of video data which has been enriched using semantic annotations. A manual annotation process is time-consuming and prone to error due to subject bias. We explore the content-based retrieval of videos containing walking subjects, using semantic queries. We evaluate current biometric research using gait, unique in its effectiveness at recognising people at a distance. We introduce a set of semantic traits discernible *by humans* at a distance, outlining their psychological validity. Working under the premise that similarity of the chosen gait signature implies similarity of certain semantic traits we perform a set of semantic retrieval experiments using popular latent semantic analysis techniques from the information retrieval community.

1 Introduction

In 2006 it was reported that around 4 million CCTV cameras were installed in the UK[4]. This results in 1Mb of video data per second per camera, using relatively conservative estimates[1] . Analysis of this huge volume of data has motivated the development of a host of interesting automated techniques, as summarised in[7][16], whose aim is to facilitate effective use of these large quantities of surveillance data. Most techniques primarily concentrate on the description of human behaviour and activities. Some approaches concentrate on low level action features, such as trajectory and direction, whilst others include detection of more complex concepts such as actor goals and scenario detection. Efforts have also been developed which analyse non human elements including automatic detection of exits and entrances, vehicle monitoring, etc.

Efficient use of large collections of images and videos by humans, such as CCTV footage, can be achieved more readily if media items are meaningfully *semantically transcoded* or *annotated*. Semantic and natural language description has been discussed [16] [41] as an open area of interest in surveillance. This

[1] 25 frames per second using 352×288 CIF images compressed using MPEG4 (`http://www.info4security.com/story.asp?storyCode=3093501`)

D. Duke et al. (Eds.): SAMT 2008, LNCS 5392, pp. 105–120, 2008.
© Springer-Verlag Berlin Heidelberg 2008

includes a mapping between behaviours and the semantic concepts which encapsulate them. In essence, automated techniques suffer from issues presented by the multimedia semantic gap[44], between semantic queries which users readily express and which systems cannot answer.

Although some efforts have attempted to bridge this gap for behavioural descriptions, an area which has received little attention is semantic appearance descriptions, especially in surveillance. Semantic whole body descriptions (Height, Figure etc.) and global descriptions (Sex, Ethnicity, Age, etc.) are a natural way to describe individuals. Their use is abundant in character description in narrative, helping readers put characters in a richer context with a few key words such as *slender* or *stout*. In a more practical capacity, stable physical descriptions are of key importance in eyewitness crime reports, a scenario where human descriptions are paramount as high detail images of assailants are not always available. Many important semantic features are readily discernible from surveillance videos by humans, and yet are challenging to extract and analyse automatically. Unfortunately, the manual annotation of videos is a laborious[7][16] process, too slow for effective use in real time CCTV footage and vulnerable to various sources of human error (subject variables, anchoring etc.). Automatic analysis of the way people walk[29] (their gait) is an efficient and effective approach to describing human features at a distance. Yet automatic gait analysis techniques do not necessarily generate signatures which are immediately comprehensible by humans. We show that Latent Semantic Analysis techniques, as used successfully by the image retrieval community, can be used to associate semantic physical descriptions with automatically extracted gait features. In doing so, we contend that retrieval tasks involving semantic physical descriptions could be readily facilitated.

The rest of this paper is organised in the following way. In Section 2 we describe Latent Semantic Analysis, the technique chosen to bridge the gap between semantic physical descriptions and gait signatures. In Section 3 we introduce the semantic physical *traits* and their associated *terms*; justifying their psychological validity. In Section 4 we briefly summarise modern gait analysis techniques and the gait signature chosen for our experiments. In Section 5 we outline the source of our experiment's description data, using it in Section 6 where we outline the testing methodology and show that our novel approach allows for content-based video retrieval based on gait. Finally in Section 7 we discuss the final results and future work.

2 Latent Semantic Analysis

2.1 The Singular Value Decomposition

In text retrieval, Cross Language Latent Semantic indexing (CL-LSI) [20], itself an extension of LSI [9], is a technique which statistically relates contextual-usage of terms in large corpuses of text documents. In our approach, LSI is used to construct a Linear-Algebraic Semantic Space from multimedia sources[14][37]

within which documents and terms sharing similar *meaning* also have similar *spacial location*.

We start by constructing an occurrence matrix $\mathbf{O}$ whose values represent the *presence* of terms in documents (columns represent documents and rows represent terms). In our scenario documents are videos. Semantic features and automatic features are considered terms. The "occurrence" of an automatic feature signifies the magnitude of that portion of the automatic feature vector while the "occurrence" of a semantic term signifies its semantic relevance to the subject in the video. Our goal is the production of a rank reduced factorisation of the observation matrix consisting of a term matrix $\mathbf{T}$ and document matrix $\mathbf{D}$, such that:

$$\mathbf{O} \approx \mathbf{TD}. \tag{1}$$

Where the vectors in $\mathbf{T}$ and $\mathbf{D}$ represent the *location* of individual terms and documents respectively within some shared space.

$\mathbf{T}$ and $\mathbf{D}$ can be efficiently calculated using the singular value decomposition (SVD) which is defined as:

$$\mathbf{O} = \mathbf{U\Sigma V}^T \tag{2}$$

Such that $\mathbf{T} = \mathbf{U}$ and $\mathbf{D} = \mathbf{\Sigma V}^T$, and the rows of $\mathbf{U}$ represent positions of terms and the columns of $\mathbf{\Sigma V^T}$ represent the position of documents. The diagonal entries of $\mathbf{\Sigma}$ are equal to the singular values of $\mathbf{O}$. The columns of $\mathbf{U}$ and $\mathbf{V}$ are, respectively, left- and right-singular vectors for the corresponding singular values in $\mathbf{\Sigma}$. The singular values of any $m \times n$ matrix $\mathbf{O}$ are defined as values $\{\sigma_1, .., \sigma_r\}$ such that :

$$\mathbf{Ov_i} = \sigma_i \mathbf{u_i}, \tag{3}$$

and

$$\mathbf{O}^T\mathbf{u_i} = \sigma_i \mathbf{v_i} \tag{4}$$

Where $\mathbf{v}_i$ and $\mathbf{u}_i$ are defined as the right and left singular vectors respectively.

In can be shown that $\mathbf{v}_i$ and $\mathbf{u}_i$ are in fact the *eigenvectors* with corresponding *eigenvalues* $\{\lambda_1 = \sigma_1^2, .., \lambda_r = \sigma_r^2\}$ of the square symmetric matrices $\mathbf{O}^T\mathbf{O}$ and $\mathbf{OO}^T$ respectively, referred to as the *co-occurrence* matrices. The matrix $\mathbf{U}$ contains all the eigenvectors of $\mathbf{OO}^T$ as its rows while $\mathbf{V}$ contains all the eigenvectors of $\mathbf{O}^T\mathbf{O}$ its rows and $\mathbf{\Sigma}$ contains all the eigenvalues along its diagonal. Subsequently:

$$\mathbf{O}^T\mathbf{O} = \mathbf{V\Sigma}^T\mathbf{U}^T\mathbf{U\Sigma V}^T = \mathbf{V\Sigma}^T\mathbf{\Sigma V}^T, \tag{5}$$

$$\mathbf{OO}^T = \mathbf{U\Sigma V}^T\mathbf{V\Sigma}^T\mathbf{U}^T = \mathbf{U\Sigma\Sigma}^T\mathbf{U}^T. \tag{6}$$

To appreciate the importance of SVD and the eigenvector matrices $\mathbf{V}$ and $\mathbf{U}$ for information retrieval purposes, consider the *meaning* of the respective co-occurrence matrices.

$$\mathbf{T_{co}} = \mathbf{OO}^T, \tag{7}$$

$$\mathbf{D_{co}} = \mathbf{O}^T\mathbf{O}. \tag{8}$$

The magnitude of the values in $\mathbf{T_{co}}$ relate to how often a particular term appears with every other term throughout all documents, therefore some concept of the "relatedness" of terms. The values in $\mathbf{D_{co}}$ relate to how many terms every document shares with every other document, therefore the "relatedness" of documents.

By definition the matrix of eigenvectors $\mathbf{U}$ and $\mathbf{V}$ of the two matrices $\mathbf{T_{co}}$ and $\mathbf{D_{co}}$ respectively form two basis for the co-occurrence spaces, i.e. the combination of terms (or documents) which the entire space of term co-occurrence can be projected into without information loss.

In a similar strategy to Principal Components Analysis (PCA), LSA works on the premise that the eigenvectors represent underlying latent concepts encoded by the co-occurrence matrix and by extension the original data. It is helpful to think of these latent concepts as mixtures (or weightings) of terms or documents. Making such an assumption allows for some interesting mathematical conclusions. Firstly, the eigenvectors with the largest corresponding eigenvalues can be thought of the *most* representative latent concepts of the space. This means by using only the most relevant components of $\mathbf{T}$ and $\mathbf{D}$ (as ordered by the singular values), less meaningful underlying concepts can be ignored and higher accuracy achieved. Also as both the document and term co-occurrence matrices represent the same data, their latent concepts must be identical and subsequently comparable[2]. Therefore the position of every term or document projected into the latent space are similar if the terms and documents in fact share similar meaning.

2.2 Using SVD

With this insight, our tasks becomes the choice of semantic and visual terms to be observed from each subject for the generation of an observation matrix. Once this matrix is generated, content-based retrieval by semantic query of unannotated documents can be achieved by exploiting the projection of partially observed vectors into the eigenspace represented by either $\mathbf{T}$ or $\mathbf{D}$.

Assume we have two subject-video collections, a fully annotated training collection and a test collection, lacking semantic annotations. A matrix $\mathbf{O}_{train}$ is constructed such that training documents are held in its columns. Both visual and semantic terms are fully observed for each training document, i.e. a term is set to a non-zero value encoding its existence or relevance to a particular video. Using the process described in Section 2.1 we can obtain $\mathbf{T}_{train}$ and $\mathbf{D}_{train}$ for the training matrix $\mathbf{O}_{train}$.

Content-Based Retrieval. To retrieve the set of unannotated subjects based on their visual gait components alone, a new partially observed document matrix $\mathbf{O}_{test}$ is constructed such that visual gait terms are prescribed and semantic terms are set to zero. For retrieval by semantic terms, a query document matrix is constructed where all visual and non-relevant semantic terms are set to zero

[2] It can also be shown that the two sets of eigenvectors are in fact in the same vector space[37] and are subsequently directly comparable.

while relevant semantic terms are given a non-zero value (usually 1.0), this query matrix is $\mathbf{O}_{query}$. The query and test matrix are projected in the latent space in following manner:

$$\mathbf{D}_{test} = \mathbf{T}_{train}^{T}\mathbf{O}_{test}, \tag{9}$$

$$\mathbf{D}_{query} = \mathbf{T}_{train}^{T}\mathbf{O}_{query}. \tag{10}$$

Projected test documents held in $\mathbf{D}_{test}$ are simply ordered according to their cosine distance from query documents in $\mathbf{D}_{query}$ for retrieval. This process readily allows for automatic annotation, though exploration in this area is beyond the scope of this report. We postulate that annotation could be achieved by finding the distance of $\mathbf{D}_{test}$ to each term in $\mathbf{T}_{train}$. A document is annotated with a term if that term is the closest compared to others belonging to the same physical *trait* (discussed in more detail in Section 3).

We show results for retrieval experiments in Section 6.

3 Human Physical Descriptions

The description of humans based on their physical features has been explored for several purposes including medicine[34], eyewitness analysis and human identification [3]. Descriptions chosen differ in levels of granularity and include features both visibly measurable but also those only measurable through use of specialised tools. One of the first attempts to systematically describe people for identification based on their physical traits was the anthropometric system developed by Bertillon [5] in 1896. His system used eleven precisely measured traits of the human body including height, length of right ear and width of cheeks. This system was quickly surpassed by other forms of forensic analysis such as fingerprints. More recently, physical descriptions have also been used in biometric techniques as an ancillary data source where they are referred to as *soft biometrics*[28], as opposed to primary biometric sources such as iris, face or gait. In behaviour analysis, several model based techniques[1] attempt the automatic extraction of individual body components as a source of behavioural information. Though the information about the individual components is not used directly, these techniques provide some insight into the level of granularity at which body features are still discernible at a distance.

When choosing the features that should be considered for semantic retrieval of surveillance media, two major questions must be answered. Firstly, which human traits should be described and secondly, how should these traits be represented. The following sections outline and justify the traits chosen and outline the semantic terms chosen for each physical trait.

Physical Traits

To match the advantages of automatic surveillance media, one of our primary concerns was to choose traits that are discernible by humans at a distance. To

[3] Interpol. Disaster Victim Identification Form (Yellow). booklet, 2008.

do so we must firstly ask which traits individuals can *consistently* and *accurately* notice in each other at a distance. Three independent traits - Age, Race and Sex, are agreed to be of primary significance in cognitive psychology. For gait, humans have been shown to successfully perceive such categories using generated point light experiments [39] with limited visual cues. Other factors such as the target's perceived somatotype [26] (build or physique attributes) are also prominent in cognition.

In the eyewitness testimony research community there is a relatively mature idea of which concepts witnesses are most likely to recall when describing individuals [42]. Koppen and Lochun [19] provide an investigation into witness descriptions in archival crime reports. Not surprisingly, the most accurate and highly mentioned traits were Sex (95% mention 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, a result also noted by Yarmey and Yarmey [43]. 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, 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 the brief mention of Height and Weight. MacLeod et al. [25] performed a unique analysis on whole body descriptions using bipolar scales to define traits. Initially, whole body traits often described by people in freeform annotations experiments were gauged using a set of moving and stationary subjects. From an initial list of 1238 descriptors, 23 were identified as unique and formulated as five-point bipolar scales. The reliability and descriptive capability of these features were gauged in a separate experiment involving subjects walking at a regular pace around a room. Annotations made using these 23 features were assessed using product moment correlation and their underlying similarity was assessed using a principal components analysis. The 13 most reliable terms and most representative of the principle components have been incorporated into our final set of traits.

Jain et al. [17] outline a set of key characteristics which determine a physical trait's suitability for use in biometric identification, a comparable task to multimedia retrieval. These include: Universality, Distinctiveness, Permanence and Collectability.

The choice of our physiological traits keeps these tenets in mind. Our semantic descriptions are universal in that we have chosen factors which everyone has. We have selected a set of subjects who appeared to be semantically distinct in order to confirm that these semantic attributes can be used. The descriptions are relatively permanent: overall Skin Colour naturally changes with tanning, but our description of Skin Colour has racial overtones and these are perceived to be more constant. Our attributes are easily collectible and have been specifically

selected for being easily discernible at a distance by humans. However much care has been taken over procedure and definition to ensure consistency of acquisition (see Section 5).

Using a combination of the studies in cognitive science, witness descriptions and the work by MacLeod et al. [25] we generated a list of visual semantic traits which is given in Table 1.

Semantic Terms

Having outlined which physical traits should allowed for, the next question is how these traits should be represented. Soft biometric techniques use a mixture of categorical metrics (e.g. Ethnicity) and value metrics (e.g. Height) to represent their traits. Humans are generally less consistent when making value judgements in comparison to category judgements. Subsequently, in our approach we formulate all traits with sets of mutually exclusive semantic terms rather than using value metrics. This approach is more representative of the categorical nature of human cognition [38] [26] [39]. This is naturally achieved for certain traits, primarily when no applicable underlying value order exists (Sex, Hair Colour etc.). For other traits representable with intuitive value metrics (Age, Lengths, Sizes etc.) bipolar scales representing concepts from *Small* to *Large* are used as semantic terms. This approach closely matches human categorical perception. Annotations obtained from such approaches have been shown to correlate with measured numerical values [8]. Perhaps the most difficult trait for which to find a limited set of terms was Ethnicity. There is a large corpus of work [12] [33] [2] exploring ethnic classification, each outlining different ethnic terms; ranging from the use of 3 to 200, with non necessarily convergent. Our ethnic terms encompass the three categories mentioned most often and an extra two categories (Indian and Middle Eastern) matching the UK census[4].

4 Automatic Gait Descriptions

In the medical, psychological and biometric community, automatic gait recognition has enjoyed considerable attention in recent years. Psychological significance in human identification has been demonstrated by various experiments [39] [18]; it is clear that the way a person walks and their overall structure hold a significant amount of information used by humans when identifying each other. Inherently, gait recognition has several attractive advantages as a biometric. It is unobtrusive, meaning people are more likely to accept gait analysis over other, more accurate, yet more invasive biometrics such as finger print recognition or iris scans. Also gait is one of the few biometrics which has been shown to identify individuals effectively at large distances and low resolutions. However this flexibility also gives rise to various challenges in the use of gait as a biometric. Gait is (in part) a behavioural biometric and as such is affected by a large variety of

[4] `http://www.statistics.gov.uk/about/Classifications/ns_ethnic_classification.asp` Ethnic classification

Table 1. Physical traits and associated semantic terms

Body Shape	
1. Arm Length	[Very Short, Short, Average, Long, Very Long]
2. Arm Thickness	[Very Thin, Thin, Average, Thick, Very Thick]
3. Chest	[Very Slim, Slim, Average, Large, Very Large]
4. Figure	[Very Small, Small, Average, Large, Very Large]
5. Height	[Very Short, Short, Average, Tall, Very Tall]
6. Hips	[Very Narrow, Narrow, Average, Broad, Very Broad]
7. Leg Length	[Very Short, Short, Average, Long, Very Long]
8. Leg Shape	[Very Straight, Straight, Average, Bow, Very Bowed]
9. Leg Thickness	[Very Thin, Thin, Average, Thick, Very Thick]
10. Muscle Build	[Very Lean, Lean, Average, Muscly, Very Muscly]
11. Proportions	[Average, Unusual]
12. Shoulder Shape	[Very Square, Square, Average, Rounded, Very Rounded]
13. Weight	[Very Thin, Thin, Average, Fat, Very Fat]

Global	
14. Age	[Infant, Pre Adolescence, Adolescence, Young Adult, Adult, Middle Aged, Senior]
15. Ethnicity	[Other, European, Middle Eastern, Far Eastern, Black, Mixed]
16. Sex	[Female, Male]
17. Skin Colour	[White, Tanned, Oriental, Black]
Head	
18. Facial Hair Colour	[None, Black, Brown, Blond, Red, Grey]
19. Facial Hair Length	[None, Stubble, Moustache, Goatee, Full Beard]
20. Hair Colour	[Black, Brown, Blond, Grey, Red, Dyed]
21. Hair Length	[None, Shaven, Short, Medium, Long]
22. Neck Length	[Very Short, Short, Average, Long, Very Long]
23. Neck Thickness	[Very Thin,Thin,Average,Thick,Very Thick]

co-variates including mood, fatigue, clothing etc. all of which can result in large within-subject (intra-class) variance.

Over the past 20 years there has been a considerable amount of work dedicated to effective automatic analysis of gait with the use of marker-less machine vision techniques attempting to match the capabilities of human gait perception[30]. Broadly speaking, these techniques can be separated into model based techniques and holistic statistical techniques.

The latter approaches tend to analyse the human silhouette and its temporal variation without making any assumptions as to how humans tend to move. An early example of such an approach was performed by Little and Boyd [23] who extract optic flow "blobs" between frames of a gait video which they use to fit an ellipsoids to describe predominant axis of motion. Murase and Sakai [27] analyse gait videos by projecting each frame's silhouettes into the eigenspace separately and using the trajectory formed by all of an individual's separate frames in the eigenspace as their signature. Combining each frame silhouette and averaging by number of frames, or simply average silhouette [13] [24] [40], is the most popular

holistic approach. It provides relatively promising results and is comparatively simple to implement and as such is often used as a baseline algorithm.

Model based techniques start with some assumption of how humans move or a model for human body structure, usually restricted to one view point, though some tackle the problem in 3D. Values for model parameters are estimated which most faithfully represent the sensed video data. An elegant early approach by [31] stacked individual silhouettes in an x-y-time (XYT) space, fitting a helix to the distinctive pattern caused by human legs at individual XT slices. The helix perimeters are used to define the parameters for a five-part stick model. Another, more recent approach by BenAbdelkader et al. [3] uses a structural model and attempts to gather evidence for subject height and cadence.

Model based techniques make several assumptions and explicitly extract certain information from subject videos. Though this would be useful for specific structural semantic terms (Height, Arm/Leg dimensions etc.), the model could feasibly ignore global semantic terms (Sex, Ethnicity etc.) evidence for which could exist in the holistic information[21]. Subsequently we choose the simple yet powerful average silhouette operation for our automatic gait signature both for purposes of simplicity and to increase the likelihood of correlation with global semantic terms.

5 Semantic and Automatic Data Source

In this section we describe the procedures undertaken to extract automatic and manual data sources describing our gait videos. Our videos are of 115 individual subjects each with a minimum of 6 video samples from the Southampton University Gait Database [36] [36]. In our experiments, the videos used are from camera set-up "a" during which subjects walk at a natural pace side on to the plane of the camera view and walking either towards the left or right. Each subject has been annotated by at least two separate annotators, though 10 have been annotated with 40 annotators as part of a previous, more rigourous, though smaller scale experiment [35].

Semantic Features

Semantic annotations were collected using the GaitAnnotate system; a web based application designed to show arbitrary biometric data sources to users for annotation, as shown in Fig. 1. This interface allows annotators to view all video samples of 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 e.g. that Height of an individual should be assigned absolutely in compared to a perceived global "Average" where traits such as Arm Length could be annotated in comparison to the subject's overall physique. This annotation data was also gathered from some subjects present in the video set, as well as from subjects not present (e.g. a class of Psychology students, the main author etc.).

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

- **Memory[10] -** Passage of time may affect a witness' recall of a subject's traits. Memory is affected by variety of factors e.g. the construction and utterance of featural descriptions rather than more accurate (but indescribable) holistic descriptions. Such attempts often alter memory to match the featural descriptions.
- **Defaulting[22] -** Features may be left out of descriptions in free recall. This is often not because the witness failed to remember the feature, but rather that the feature has some default value. Race may be omitted if the crime occurs in a racially homogenous area, Sex may be omitted if suspects are traditionally Male.
- **Observer Variables[11][32] -** A person's own physical features, namely their self perception and mental state, may affect recall of physical variables. For example, tall people have a skewed ability to recognise other tall people but will have less ability when it comes to the description shorter individuals, not knowing whether they are average or very short.
- **Anchoring[6] -** When a person is asked a question and is initially presented with some default value or even seemingly unrelated information, the replies given are often weighted around those initial values. This is especially likely when people are asked for answers which have some natural ordering (e.g. measures of magnitude)

We have designed our semantic data gathering procedure to account for all these factors. Memory issues are addressed by allowing annotators to view videos of subjects as many times as they please, also allowing them to repeat a particular video if necessary. Defaulting is 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 [43]. Subject variables can never be completely removed so instead we allow the study of differing physical traits across various annotators. Users are asked to self annotate based on self perception, also certain subjects being annotated are themselves annotators. 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. We have accounted for anchoring of terms gathered for individual traits by setting the default term of a trait to a neutral "Unsure" rather than any concept of "Average".

To allow for inclusion of semantic terms of each trait in the LSA observation matrix, each semantic term is represented by its occurrence for each subject. This occurrence is extracted by finding a consensus between annotators which

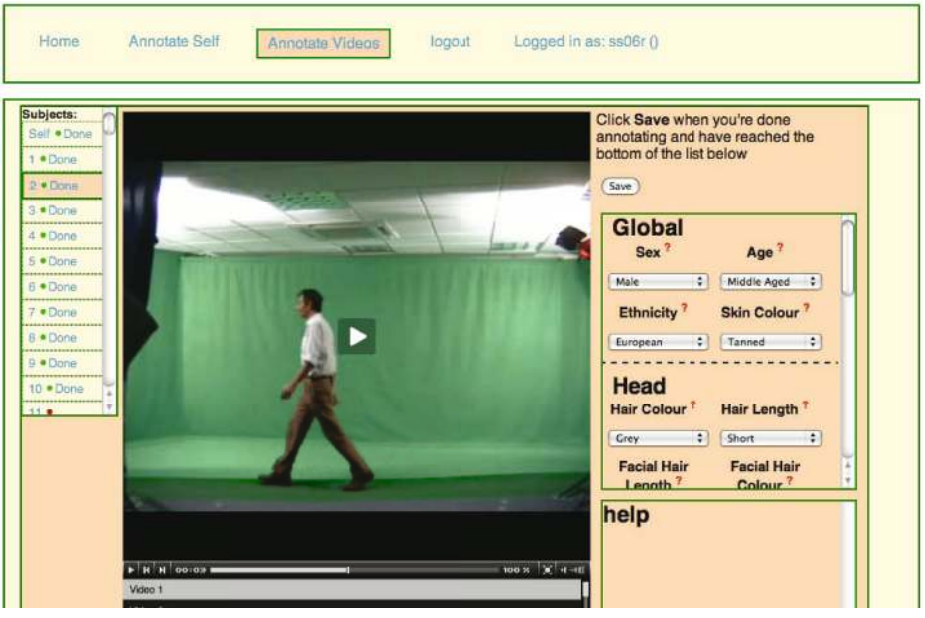

Fig. 1. Example of GAnn interface

made a judgement of a particular term for a particular subject. Each of the n annotators produces the the i^{th} annotation assigning the j^{th} term for the k^{th} subject, producing a response $r_{ijk} \in [0, 1]$. The value for each term t_{jk} for j^{th} for the k^{th} subject is calculated such that:

$$t_{jk} = \frac{1}{n} \sum_{i=0}^{n} r_{ij} \tag{11}$$

This results in a single annotation for each subject for each term which is a value between 0.0 and 1.0 which defines how relevant a particular semantic term is to a particular subject, i.e. its occurrence (see Section 2).

If an annotator responds as with "Unsure" for each trait, or does not provide the annotations at all, their response is set to the mode of that trait across all annotators across that particular subject. This results in a complete 113x115 (113 semantic terms, 115 subjects) matrix which is concatenated with the automatic feature matrix described in the following section.

Automatic Gait Features

The automatic feature vector used for these experiments were the average silhouette gait signatures. For each gait video, firstly the subject is extracted from the scene with a median background subtraction and transformed into a binary silhouette. This binary silhouette is resized to a 64x64 image to make the signature distance invariant. The gait signature of a particular video is the averaged

summation of all these binary silhouettes across one gait cycle. For simplicity the gait signature's intensity values are use directly, although there have been several attempts made to find significant features in such feature vectors, using ANOVA or PCA [40] and also a Fourier Decomposition [15].

This results in 4096 (64x64) automatic feature components which describe each sample video of each of the 115 subjects. The final observation matrix $\mathbf{O}$ is constructed by concatenating each sample feature vector with its subject's annotation feature vector as described in the previous section. This complete set of automatically and semantically observed subjects is manipulated in Section 6 to generate $\mathbf{O}_{train}$ and $\mathbf{O}_{test}$ as described in Section 2.

6 Experiments

For the retrieval experiment it was required to construct a training matrix $\mathbf{O}_{train}$, for which visual features and semantic features are fully observed, and $\mathbf{O}_{test}$ matrix such that the semantic features are set to zero. The retrieval task attempts to order the documents in $\mathbf{O}_{test}$ against some semantic queries.

The documents in the training stage are the samples (and associated semantic annotations) of a randomly selected set of half of the 115 subjects, the test documents are the other subjects with their semantic terms set to zero. For analysis, 10 such sets are generated and latent semantic spaces ($\mathbf{T}_{train}$ and $\mathbf{D}_{train}$) are generated for each.

6.1 Semantic Query Retrieval Results

We test the retrieval ability of our approach by testing each semantic term in isolation (e.g. Sex Male, Height Tall etc.). A few example retrieval queries can be seen in Fig. 3. Here, the Male subjects have been retrieved successfully as have the Female subjects. In Pre-Adolescence, the system selects two children but one adult, incorrectly. The Hair Length retrieval is consistently correct. To put our results in context we also measure the standard mean average precision (mAP) metric as calculated by TREC-Eval. The mAP of each semantic term is taken from the mAP of a random ordering for each query. To generate the random mAP we generate 100 random orderings for each semantic query and average their mAP.Fig. 2 shows the sum of the random order difference of each semantic terms for each trait. These results give some idea of which traits our approach is most capable of performing queries against, and which it is not.

Our results show some merit and produce both success and failure, as expected. It has been shown in previous work for example that Sex (mAP=0.12) is decipherable from average silhouettes alone [21], achieved by analysing the separate parts of the human silhouette. It is also expected that physical metrics of mass such as Weight (mAP=0.043), Figure (mAP=0.041) and Leg Thickness (mAP=0.044) were also likely to be relatively successful as the average silhouette maintains a linear representation of these values in the overall intensity of pixels. Also, the poor performance of Height (mAP=0.0026) is expected as the

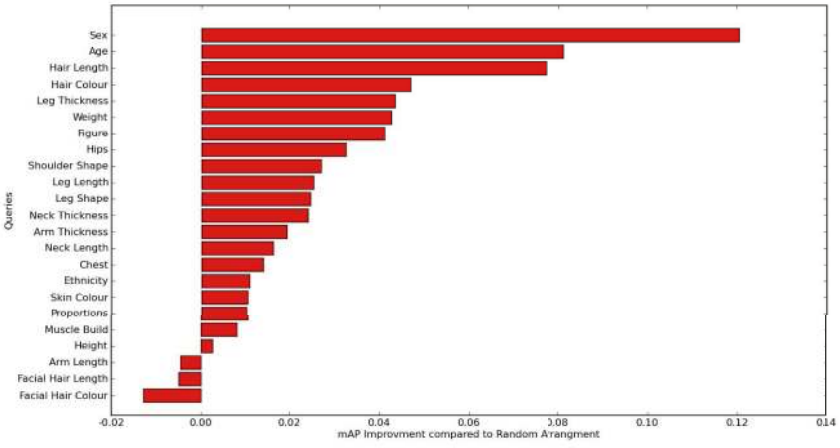

Fig. 2. Bar-graph showing the mean average precision improvement for each semantic trait. Each trait is the weighted sum of its substituent semantic terms.

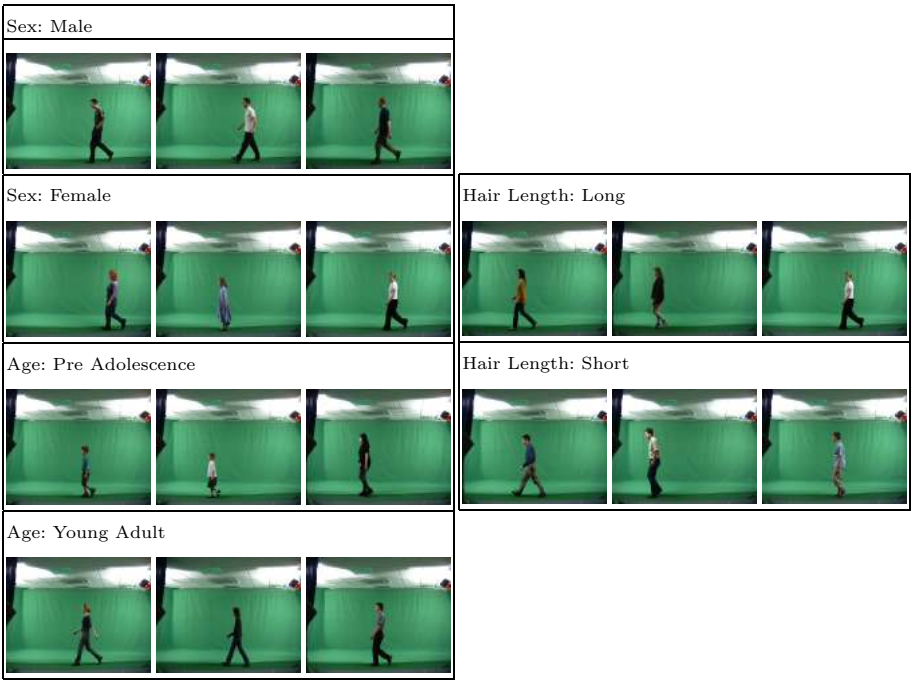

Fig. 3. Some Retrieval Results

average silhouette features have removed this feature by normalising silhouettes, making them the same height for comparison. Only a latent concept of Height in terms of the aspect ratio is maintained.

Perhaps most surprising is the relative success of Hair Colour (mAP=0.047) metric as, on first inspection it seems as though silhouette images maintain no colour information. However, the construction of binary silhouettes is undoubtedly affected by hair colour when compared to background, and as such the average silhouette images retain hair colour as brightness in the head region. Veres et al. [40] noted that this region was the most useful portion of the silhouette for recognition. It is also likely that this Hair Colour holds some significant relationship with another semantic feature, for example Sex, as most of our Female participants were indeed of East Asian origin and subsequently had Black Hair. In future work we discuss an exploration into other such points of interest.

A t-test was also performed to gauge the significance of each mAP difference from random in relation to the mAP standard deviation. A small p-value indicates higher significance. The p-value of each mAP difference shows that many of our retrieval precision rates were significant[5]. Sex, Hair Colour/Length and Age each shared a p-value of 10^{-5} where Facial Hair features had p-values > 0.2. This further demonstrates the merit of our approach. It should be noted that there is a correlation with traits that performed poorly and those reported by annotators to be confusing and difficult to decipher.

7 Conclusions and Further Work

We have introduced the use of semantic human descriptions as queries in content-based retrieval against human gait signatures. We carefully selected a set of physical traits and successfully used them return an ordered list of un-annotated subjects based on their gait signature alone. Our results confirm the results of previous work with regards to traits such as Sex and we also note the capability of retrieval using other traits, previously unexplored, such as Age and some build attributes.

There are several interesting avenues of research suggested by this work. A further exploration into other important semantic features would no doubt uncover a large range of useful terms for discovery of surveillance video. An exploration into other gait signatures would also improve the recall of certain semantic features. Using model based techniques to more directly extract Height and limb attributes would no doubt improve their retrieval rates.

References

[1] Aggarwal, J.K., Cai, Q.: Human motion analysis: A review. Computer Vision and Image Understanding: CVIU 73(3), 428–440 (1999)

[2] Barbujani, G.: Human races: Classifying people vs understanding diversity. Current Genomics 6(12), 215–226 (2005)

[3] BenAbdelkader, C., Cutler, R., Davis, L.: Stride and cadence as a biometric in automatic person identification and verification. In: Proc. 5th IEEEFG, pp. 372–377 (May 2002)

[5] To a threshold of 0.01.

[4] Bennetto, J.: Big brother britain 2006: we are waking up to a surveillance society all around us. The Independant (2006)
[5] Bertillon, A.: Signaletic Instructions including the theory and practice of Anthropometrical Identification. The Werner Company (1896)
[6] Chapman, G.B., Johnson, E.J.: Incorporating the irrelevant: Anchors in judgments of belief and value. In: Heuristics and Biases: The Psychology of Intuitive Judgment, pp. 120–138. Cambridge University Press, Cambridge (2002)
[7] Davies, A., Velastin, S.: A Progress Review of Intelligent CCTV Surveillance Systems. In: IEEEIDAACS 2005, pp. 417–423 (September 2005)
[8] Dawes, R.M.: Suppose We Measured Height With Rating Scales Instead of Rulers. App. Psych. Meas. 1(2), 267–273 (1977)
[9] Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. J. of the American Society of Information Science 41(6), 391–407 (1990)
[10] Ellis, H.D.: Practical aspects of facial memory. In: Eyewitness Testimony: Psychological perspectives, section 2, pp. 12–37. Cambridge University Press, Cambridge (1984)
[11] Flin, R.H., Shepherd, J.W.: Tall stories: Eyewitnesses ability to estimate height and weight characteristics. Human Learning 5 (1986)
[12] Gould, S.J.: The Geometer of Race. Discover, pp. 65–69 (1994)
[13] Han, J., Bhanu, B.: Statistical feature fusion for gait-based human recognition. In: Proc. IEEE CVPR 2004, vol. 2 , pp. II–842–II–847 (June-July 2004)
[14] Hare, J.S., Lewis, P.H., Enser, P.G.B., Sandom, C.J.: A linear-algebraic technique with an application in semantic image retrieval. In: Sundaram, H., Naphade, M., Smith, J.R., Rui, Y. (eds.) CIVR 2006. LNCS, vol. 4071, pp. 31–40. Springer, Heidelberg (2006)
[15] Hayfron-Acquah, J.B., Nixon, M.S., Carter, J.N.: Automatic Gait Recognition by Symmetry Analysis. Pattern Recognition Letters 24(13), 2175–2183 (2003)
[16] Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEETSMC(A) 34(3), 334–352 (2004)
[17] Jain, A.K., Ross, A., Prabhakar, S.: An Introduction to Biometric Recognition. Trans. CSVT 14, 4–19 (2004)
[18] Johansson, G.: Visual perception of biological motion and a model for its analysis. Percept. Phychophys. 14(2), 201–211 (1973)
[19] Koppen, P.V., Lochun, S.K.: Portraying perpetrators; the validity of offender descriptions by witnesses. Law and Human Behavior 21(6), 662–685 (1997)
[20] Landauer, T., Littman, M.: Fully automatic cross-language document retrieval using latent semantic indexing. In: 6th Annual Conference of the UW Centre for the New OED, pp. 31–38 (1990)
[21] Li, X., Maybank, S., Yan, S., Tao, D., Xu, D.: Gait components and their application to gender recognition. IEEETSMC(C) 38(2), 145–155 (2008)
[22] Lindsay, R., Martin, R., Webber, L.: Default values in eyewitness descriptions. Law and Human Behavior 18(5), 527–541 (1994)
[23] Little, J., Boyd, J.: Describing motion for recognition. In: SCV 1995, page 5A Motion II (1995)
[24] Liu, Z., Sarkar, S.: Simplest representation yet for gait recognition: averaged silhouette. In: Proc. ICPR 2004, vol. 4, pp. 211–214 (August 2004)
[25] MacLeod, M.D., Frowley, J.N., Shepherd, J.W.: Whole body information: Its relevance to eyewitnesses. In: Adult Eyewitness Testimony, ch. 6. Cambridge University Press, Cambridge (1994)

[26] Macrae, C.N., Bodenhausen, G.V.: Social Cognition: Thinking Categorically about Others. Ann. Review of Psych. 51(1), 93–120 (2000)
[27] Murase, H., Sakai, R.: Moving object recognition in eigenspace representation: gait analysis and lip reading. Pattern Recogn. Lett. 17(2), 155–162 (1996)
[28] Nandakumar, K., Dass, S.C., Jain, A.K.: Soft biometric traits for personal recognition systems. In: Zhang, D., Jain, A.K. (eds.) ICBA 2004, vol. 3072, pp. 731–738. Springer, Heidelberg (2004)
[29] Nixon, M., Carter, J.N.: Automatic recognition by gait. Proc. of the IEEE 94(11), 2013–2024 (2006)
[30] Nixon, M.S., Carter, J.N.: Automatic recognition by gait. Proceedings of the IEEE 94(11), 2013–2024 (2006)
[31] Niyogi, S., Adelson, E.: Analyzing and recognizing walking figures in XYT. In: Proc. CVPR 1994, pp. 469–474 (June 1994)
[32] O'Toole, A.J.: Psychological and Neural Perspectives on Human Face Recognition. In: Handbook of Face Recognition. Springer, Heidelberg (2004)
[33] Ponterotto, J.G., Mallinckrodt, B.: Introduction to the special section on racial and ethnic identity in counseling psychology: Conceptual and methodological challenges and proposed solutions. J. of Counselling Psych. 54(3), 219–223 (2007)
[34] Rosse, C., Mejino, J.L.V.: A reference ontology for biomedical informatics: the foundational model of anatomy. J. of Biomed. Informatics 36(6), 478–500 (2003)
[35] Samangooei, S., Guo, B., Nixon, M.S.: The use of semantic human description as a soft biometric. In: BTAS (September 2008)
[36] Shutler, J., Grant, M., Nixon, M.S., Carter, J.N.: On a large sequence-based human gait database. In: RASC 2006, pp. 66–72 (2002)
[37] Skillicorn, D.: Understanding Complex Datasets. In: Singular Value Decomposition (SVD), ch. 3. Chapman & Hall/CRC (2007)
[38] Tajfel, H.: Social Psychology of Intergroup Relations. Ann. Rev. of Psych. 33, 1–39 (1982)
[39] Troje, N.F., Sadr, J., Nakayama, K.: Axes vs averages: High-level representations of dynamic point-light forms. Vis. Cog. 14, 119–122 (2006)
[40] Veres, G., Gordon, L., Carter, J., Nixon, M.: What image information is important in silhouette-based gait recognition? In: Proc. IEEE CVPR 2004, vol. 2, pp. II-776–II-782 (June-July 2004)
[41] Vrusias, B., Makris, D., Renno, J.-P., Newbold, N., Ahmad, K., Jones, G.: A framework for ontology enriched semantic annotation of cctv video. In: WIAMIS 2007, p. 5 (June 2007)
[42] Wells, G.L., Olson, E.A.: Eyewitness testimony. Ann. Rev. of Psych. 54, 277–295 (2003)
[43] Yarmey, A.D., Yarmey, M.J.: Eyewitness recall and duration estimates in field settings. J. of App. Soc. Psych. 27(4), 330–344 (1997)
[44] Zhao, R., Grosky, W.: Bridging the Semantic Gap in Image Retrieval. IEEE Transactions on Multimedia 4, 189–200 (2002)