Multimed Tools Appl (2010) 49:195–212
DOI 10.1007/s11042-009-0391-8

Performing content-based retrieval of humans using gait biometrics

Sina Samangooei · Mark S. Nixon

Published online: 22 October 2009

Abstract In order to analyse surveillance video, we need to efficiently explore large datasets containing videos of walking humans. 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 however, at surveillance-image resolution, the human walk (their gait) can be analysed automatically. We explore the content-based retrieval of videos containing walking subjects, using semantic queries. We evaluate current research in gait biometrics, 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. We perform experiments on a dataset of 2000 videos of people walking in laboratory conditions and achieve promising retrieval results for features such as *Sex* (mAP = 14% above random), *Age* (mAP = 10% above random) and *Ethnicity* (mAP = 9% above random).

Keywords Content based video retrieval · Latent semantic indexing · Gait biometrics · Anthropometry · Semantic enrichment

S. Samangooei (✉) · M. S. Nixon
School of Electronics and Computer Science, Southampton University,
Southampton, SO17 1BJ, UK
e-mail: ss06r@ecs.soton.ac.uk

M. S. Nixon
e-mail: msn@ecs.soton.ac.uk

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 [10, 23], 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 [23, 53] as an open area of interest in surveillance. This 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 [56] 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 by automated means. Unfortunately, the manual annotation of videos is a laborious [10, 23] 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 [38] (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 LSA (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 LSA, the technique chosen to bridge the gap between semantic physical descriptions and gait signatures. In Section 3 we introduce the semantic physical *traits* and

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

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 Background

LSA (Latent Semantic Analysis) or LSI (Latent Semantic Indexing) was initially developed by Deerwester et al. [12] in their seminal work to address the inherent problems with direct lexical comparison for text retrieval. The assumption is that documents in a corpus and their associated terms are in fact correlated artefacts generated by a set of underlying concepts. It follows that a set of documents and terms can be represented as a weighted sum of these concepts. Furthermore, it is argued that by choosing only the most important concepts to represent the space of documents and terms, retrieval rates can be improved. Therefore, the goal of LSI is to determine an optimised set of underlying concepts, a goal achieved using SVD (Singular Value Decomposition). Initial experiments using LSI [13] showed improvements of around 30% when compared to simple lexical analysis, promising results which inspired the use of LSI in a variety of text retrieval applications [5]. LSI has been adapted to tackle Content Based Image retrieval [18, 42] and more recently, the automatic annotation of un-annotated images [20, 35], displaying competitive precision and recall [21] to other contemporary approaches. We use LSI in a similar way to retrieve gait videos of humans using semantic queries.

2.2 The singular value decomposition

An $n \times m$ occurrence matrix $\mathbf{O}$ is constructed whose values represent the *presence* of m terms in n documents. In our scenario documents are videos of subjects walking. Semantic features and automatic features are considered terms. The "occurrence" of a video feature represents the intensity of a grayscale or colour pixel where The "occurrence" of a semantic term signifies its relevance to the subject in the video (see Section 5 for further details regarding the datasource). The initial goal of LSI is to determine the concepts which underpin this document-term space. It can be shown [41, 42] that this concept space can be efficiently calculated using a SVD (Singular Value Decomposition) of $\mathbf{O}$ and selecting the left- and right-singular vectors associated with the highest singular values of $\mathbf{O}$. The decomposition:

$$\mathbf{O} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{1}$$

results in an $n \times r$ matrix $\mathbf{U}$, an $r \times r$ diagonal matrix $\mathbf{\Sigma}$ and an $r \times m$ matrix $\mathbf{V}^T$. Where r is the estimated rank of $\mathbf{O}$ and therefore the number of underlying concepts in the space.[2] The diagonal matrix $\mathbf{\Sigma}$ contains the r largest singular values of $\mathbf{O}$ ordered along its diagonal. The rows of $\mathbf{U}$ represent positions of the n documents against the set of top r left-singular vectors (the eigenvectors calculated from the document co-occurence matrix $\mathbf{O}\mathbf{O}^T$) while the columns of $\mathbf{V^T}$ represent the position of the m terms against the set top r right-singular vectors (the eigenvectors calculated from the term co-occurence matrix $\mathbf{O}^T\mathbf{O}$).

In our work the first step is the selection of appropriate semantic terms and visual features to construct $\mathbf{O}$ as explored in Sections 3 and 4. Once constructed, a fully observed training matrix $\mathbf{O}_{train}$ can be decomposed resulting in $\mathbf{O}_{train} = \mathbf{U}_{train}\mathbf{\Sigma}_{train}\mathbf{V}^T_{train}$. Content based retrieval by semantic query can be achieved by projecting semantic queries as partially observed[3] vectors o_{query} into the eigenterm space $\mathbf{V}^T$ and comparing them against the projections of partially observed[4] visual signatures $\mathbf{O}_{test}$ into the same space.

$$\mathbf{U}_{test} = \mathbf{O}_{test}\left(\mathbf{\Sigma}_{train}\mathbf{V}^T_{train}\right)^T, \tag{2}$$

$$u_{query} = o_{query}\left(\mathbf{\Sigma}_{train}\mathbf{V}^T_{train}\right)^T. \tag{3}$$

If a query and a visual signature are related, they should have similar weightings to each of the eigenterms in $\mathbf{V}^T$ and therefore share a similar position once projected into $\mathbf{V}^T$ according to some distance metric. By ordering the projected test visual documents $\mathbf{U}_{test}$ by their cosine distances[5] to a projected query u_{test}, we achieve an ordering of the visual test documents based on their relevance to a semantic query and therefore retrieval.

3 Human physical descriptions

The description of humans based on their physical features has been explored for several purposes including medicine [44], eyewitness analysis and human identification [24]. 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 [6] 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* [37], 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

[2] In practice several r values are attempted to choose an optimal number of concepts for a given dataset.

[3] i.e. only semantic terms, visual terms set to 0

[4] i.e. only visual terms, semantic terms set to 0

[5] chosen to disregard scaling effects Papadimitriou et al. [41]

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.

3.1 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 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 [50] with limited visual cues. Other factors such as the target's perceived somatotype [34] (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 [54]. Koppen and Lochun [51] 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, but accuracy was 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 [55]. 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. [33] performed a unique analysis on whole body descriptions using bipolar scales to define traits. Initially, whole body traits often described by people in freeform annotation 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 principal components have been incorporated into our final set of traits.

Jain et al. [25] 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. [33] we generated a list of visual semantic traits which is given in Table 1.

3.2 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

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]

cognition [34, 49, 50]. 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 [11]. Perhaps the most difficult trait for which to find a limited set of terms was *Ethnicity*. There is a large corpus of work [2, 17, 43] 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.[6]

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 [26, 50]; 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 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 [38]. 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 [30] 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 [36] 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 [19, 31, 52], 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

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

tackle the problem in 3D. Values for model parameters are estimated which most faithfully represent the sensed video data. An elegant early approach by [39] 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.

A current challenge in gait biometrics is how it should be put to use in real world applications. The fusion of gait with existing (and more established) biometrics has been shown to be a viable approach towards taking advantage of gait's abilities while overcoming its inaccuracies due to covariates (exploratory variables). Recent studies have shown identification improvements when a gait signature is fused with a face signature [27, 32]. Another challenge in gait biometrics is the viewpoint dependent nature of the vast majority of early gait signatures. To date, much of the data analysed has been from subjects walking in the plane normal to the view of the camera; achieving viewpoint independence allows identification in a greater range of scenarios [8, 16]. Three-dimensional models derived by multiple cameras are currently being considered which go towards pose invariance. This includes approaches such as 3D model fitting [7] and arbitrary viewpoint generation [47] techniques as well as the development of novel 3D gait datasets [46].

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 [28]. 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. To complement the semantic terms chosen which describe colours of particular human traits, we have also generated a set of average colour silhouettes signatures. Both signatures are used in the experiments below.

5 Semantic and automatic data source

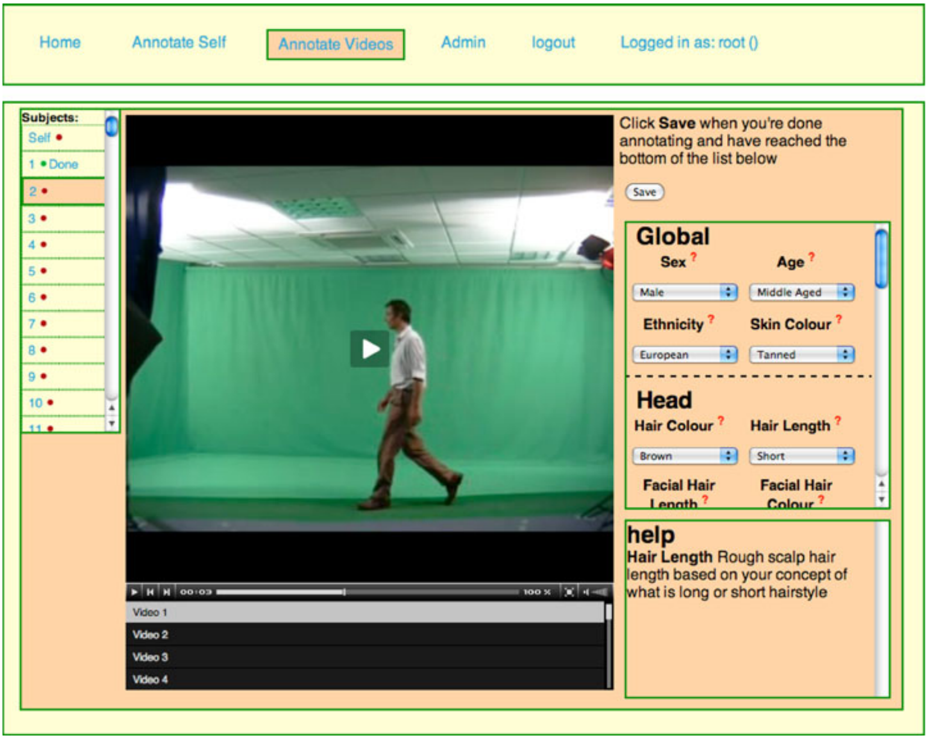
In this section we describe the procedures undertaken to extract automatic and manual data sources describing our gait videos. Our videos are of 115 individual subjects each with a minimum of 6 video samples from the Southampton University Gait Database [48] . In our experiments, the videos used are from the camera set-up wherein 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, and 5 sets of 10 have also been annotated by 5 sets of 10 individuals each. These extra annotations were made as part of a previous, more rigourous, though smaller scale experiment [45].

5.1 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 [14]—**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.



Click **Save** when you're done annotating and have reached the bottom of the list below

Global

Sex ? Male

Age ? Middle Aged

Ethnicity ? European

Skin Colour ? Tanned

Head

Hair Colour ? Brown

Hair Length ? Short

Facial Hair Length ?

Facial Hair Colour ?

help

Hair Length Rough scalp hair length based on your concept of what is long or short hairstyle

Fig. 1 Example of GAnn interface

- **Defaulting [29]—**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 [15, 40]—**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 [9]—**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 [55]. 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 made a judgement of a particular term for a particular subject. Each of the n annotators produces 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 j^{th} term for the k^{th} subject is calculated such that:

$$t_{jk} = \frac{1}{n}\sum_{i=1}^{n} r_{ijk} \tag{4}$$

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 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 113 × 115 (113 semantic terms, 115 subjects) matrix which is concatenated with the automatic feature matrix described in the following section.

5.2 Automatic gait features

Two automatic gait features are used in these experiments, the average (monochrome) silhouette gait signature and, for comparison, the average colour silhouette.

5.2.1 Standard average gait signature

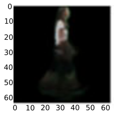
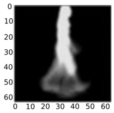
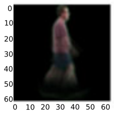
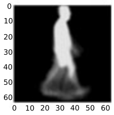
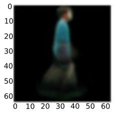
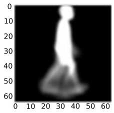
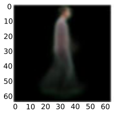
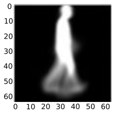
For each gait video, firstly the subject is extracted from the scene with a median background subtraction and transformed into a binary silhouette. This binary silhouette is resized to a 64 × 64 image to make the signature distance invariant. The gait signature of a particular video is the averaged summation of all these binary silhouettes across one gait cycle. For simplicity the gait signature's intensity values are used directly, although there have been several attempts made to find significant features in such feature vectors, using ANOVA or PCA [52] and also the a symmetry analysis [22].

5.2.2 Colour average gait signature

The binary silhouettes extracted during the first stage of the standard average gait signatures are used to mask the original full colour videos on a frame by frame basis. From these masked colour images the subject is extracted and normalised to 64 × 64. A colour signature is generated by averaging the colour components in all the masked images, separately for each colour, across the images from the same gait cycle as the standard average gait signature.

This two techniques result in two automatic feature vectors of size 4096 (64 × 64) and 12288 (64 × 64 × 3) (See Fig. 2) respectively which describe each sample video

Fig. 2 Signature examples

	Color Signature	Grayscale Signature
Subject 098		
Subject 112		
Query Age Pre Adolescence		
Query Age Young Adult		

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 a semantically unobserved set to construct $\mathbf{O}_{test}$ as described in Section 2.

6 Experiments

For both the monochrome and colour retrieval experiments it was necessary 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 a set semantic queries o_{query}, one for each semantic term in isolation.

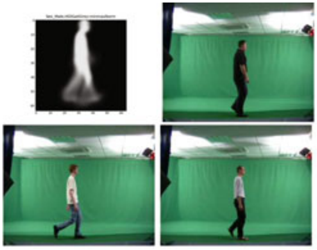
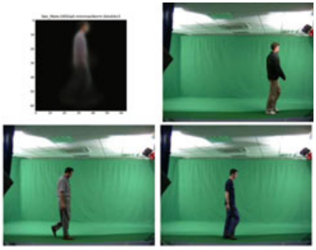
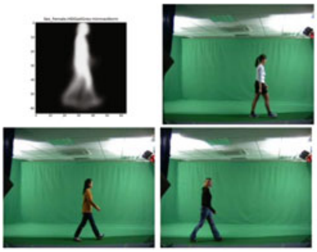
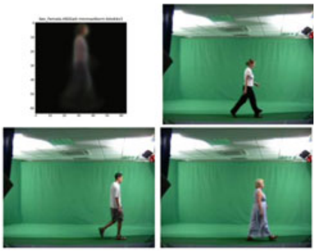
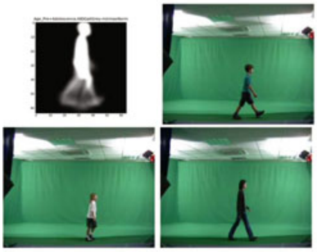
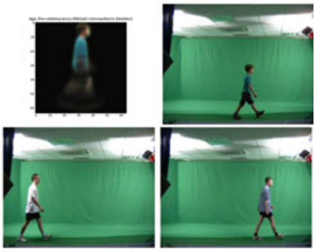
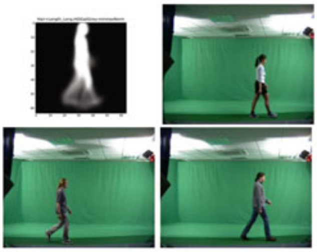
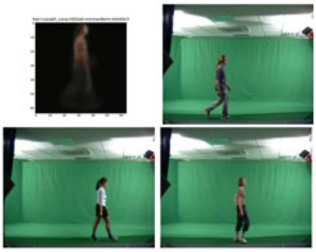
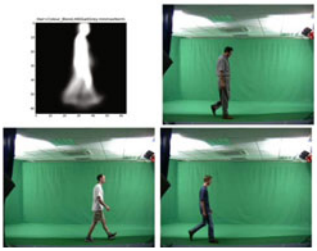
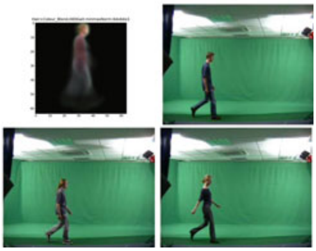
The documents in the training stage are the samples (and associated semantic annotations) of a randomly selected set of half of the 115 subjects, the test documents are the other subjects with their semantic terms set to zero. For analysis, 20 such sets are generated and matrix decompositions $\mathbf{U}_{train}$,$\mathbf{\Sigma}_{train}$ and $\mathbf{V}^{T}_{train}$ 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 Table 2 along with the signature automatically generated from the projection of the query. 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 completely random orderings for each semantic query and average their mAP. Figure 3 shows the sum of the differences of each physiological trait as a sum of it's semantic terms for both experimental configurations. These results give some idea of which traits our approach is most capable of performing queries against, and which it is not. Finally, in Table 3 we present p-values for each semantic trait as generate in a one-way ANOVA where monochrome and colour signatures are taken as separate groups, their mAPs for each of the 20 experiments as their group samples and therefore the p-value as a measure of significance of the difference between the two experimental configurations.

Our results show some merit and produce both success and failure, as expected. It has been shown in previous work for example that *Sex* ($\text{mAP} = 0.14$ and $\text{mAP} = 0.12$) is decipherable from average silhouettes alone [28], achieved by analysing the separate parts of the human silhouette. It is also expected that physical metrics such as *Height* ($\text{mAP} = 0.089$ and $\text{mAP} = 0.087$), *Figure* ($\text{mAP} = 0.051$ and $\text{mAP} = 0.059$) and *Neck Length* ($\text{mAP} = 0.056$ and $\text{mAP} = 0.065$) were also likely to be relatively successful as the average silhouette maintains a linear representation of these values in the overall intensity of pixels.

Table 2 Some example retrieval results

Query	Average silhouette	Average colour silhouette
Sex: Male		
Sex: Female		
Age: Pre Adolescence		
Hair Length: Long		
Hair Colour: Blond		

The first image in each set is the image generated for a semantic query as part of the method explained in Section 2.2. The next 3 images are video keyframes of the 3 top ranked subjects from a particular experiment

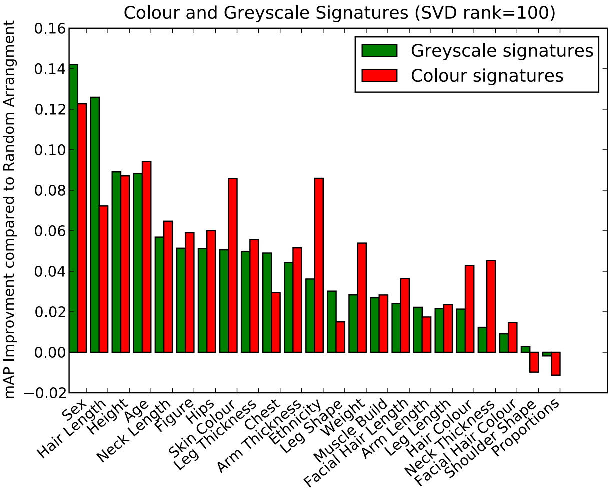


Fig. 3 The mean average precision improvement for each semantic trait. Each trait's mAP is the average summed difference its associated semantic terms

In Table 2 we see example orderings provided by our scheme and an anecdotal comparison of the ability of colour signatures against monochrome silhouettes. The examples aid to show the potential merits and pitfalls of using the different signatures. Both configurations perform well with *Sex*, though for our example *Sex Female* query, colour signatures incorrectly correlate light coloured clothing with

Table 3 The mAP p-values treating grey and colour signatures as seperate classes for each physiological trait. Here we use the significance value of $p \leq 0.1$

Trait	p-value
Significant features	
Hair length	2.3e–06
Ethnicity	2.9e–04
Hair colour	0.001
Neck thickness	0.002
Skin colour	0.002
Weight	0.006
Leg shape	0.055
Sex	0.087
Insignificant features	
Shoulder shape	0.103
Chest	0.148
Proportions	0.189
Hips	0.308
Facial hair length	0.388
Neck length	0.450
Figure	0.521
Facial hair colour	0.533
Leg thickness	0.561
Arm thickness	0.658
Arm length	0.734
Age	0.759
Muscle build	0.861
Leg length	0.873
Height	0.937

gender. The colour of clothing is ignored by the standard average silhouettes as the whole body silhouette of the individual is used and the internal detail ignored. The average colour signature has a similar problem with the example *Age* query. The tables turn on queries which inherently correlate with colour. In Table 2 we see that for the *Hair Colour* the average colour silhouette achieves more favourable results, correctly finding a correlation with light shades in the head area with blond hair (as can be seen on the automatically generated *Hair Colour* query signature).

Figure 3 also show the relative merits of the two approaches. It can be seen that whilst performing relatively poorly in both configurations, *Hair Colour* ($p = 0.001$); *Ethnicity* ($p = 0.0003$) and *Skin Colour* ($p = 0.002$) perform significantly better when colour average silhouettes are used. It should be noted however that, for *Sex* ($p = 0.087$) and *Hair Length* ($p = 0.00002$), all mAPs are significantly lower on the average colour silhouettes. This result was expected as the colour signature allows for misleading correlations with clothing, a failure which can be seen in the example query projections of *Sex Female* and *Hair Length Long* in Table 2 both showing correlation with light coloured clothing. This failure in the average colour silhouette could feasibly be avoided if only pertinent regions such as the head are taken into consideration for correlation, but which are not avoided using the holistic signatures currently used.

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 analysis confirm the results of previous works with regards to traits such as *Sex* and we also note the capability of retrieval using other traits, previously unexplored, such as *Age*, *Hair* and some build attributes. We also compare the capabilities of average (monochrome) silhouette gait signatures with a new average colour silhouette signature, exploring their respective advantages and limitations.

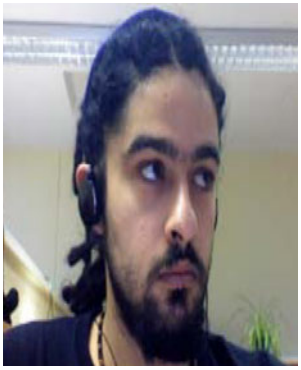
An inherent limitation of the current approach is that in using the SVD we extract linear structures for correlation while non linear correlations remain to be studied. Exploring the use of non-linear machine learning techniques would no-doubt extend the abilities of this new technique. Further, at present we handle laboratory data only. An exploration into subject retrieval in real world surveillance data would require handling of colour and low resolution video data as well as variation in illumination with complex background scenery. This remains an open area of research in gait biometrics.

There are several interesting avenues of research suggested by this work. To enrich this process, we can of course collect more manual labels allowing for a clearer notion of the value of a given trait for an individual subject. A further exploration into other important semantic traits 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 JK, Cai Q (1999) Human motion analysis: a review. Comput Vis Image Underst 73(3):428–440
2. Barbujani G (2005) Human races: classifying people vs understanding diversity. Current Genomics 6(12):215–226
3. BenAbdelkader C, Cutler R, Davis L (2002) Stride and cadence as a biometric in automatic person identification and verification. In: Proc. IEEE FG, pp 372–377
4. Bennetto J (2006) Big brother Britain 2006: we are waking up to a surveillance society all around us. In: The independent
5. Berry MW, Dumais ST, O'brien GW, Berry MW (1995) Using linear algebra for intelligent information retrieval. SIAM Rev 37:573–595
6. Bertillon A (1896) Signaletic instructions including the theory and practice of anthropometrical identification. The Werner Company
7. Bhanu B, Han J (2003) Human recognition on combining kinematic and stationary features. In: Proc. AVBPA, pp 600–608
8. Bouchrika I, Goffredo M, Carter JN, Nixon MS (2009) Covariate analysis for view-point independent gait recognition. In: Proc. ICB
9. Chapman GB, Johnson EJ (2002) Incorporating the irrelevant: anchors in judgments of belief and value. In: Heuristics and biases: the psychology of intuitive judgment. Cambridge University Press, Cambridge, pp 120–138
10. Davies AC, Velastin, SA (2005) A progress review of intelligent CCTV surveillance systems. In: Proc. IEEE IDAACS, pp 417–423
11. Dawes RM (1977) Suppose we measured height with rating scales instead of rulers. Appl Psychol Meas 1(2):267–273
12. Deerwester SC, Dumais ST, Landauer TK, Furnas GW, Harshman RA (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407
13. Dumais SI (1991) Improving the retrieval of information from external sources. Behavior research methods. In: Instruments and computers, pp 229–236
14. Ellis HD (1984) Practical aspects of facial memory. In: Eyewitness testimony: psychological perspectives, section 2. Cambridge University Press, Cambridge, pp 12–37
15. Flin RH, Shepherd JW (1986) Tall stories: eyewitnesses' ability to estimate height and weight characteristics. Hum Learn 5
16. Goffredo M, Seely RD, Carter JN, Nixon MS (2008) Markerless view independent gait analysis with self-camera calibration. In: Proc. IEEE FG
17. Gould SJ (1994) The geometer of race. Discover 65–69
18. Grosky W, Zhao R (2001) Negotiating the semantic gap: from feature maps to semantic landscapes. In: Proc. SOFSEM, pp 33–52
19. Han J, Bhanu B (2004) Statistical feature fusion for gait-based human recognition. In: Proc. IEEE CVPR, vol 2, pp II–842–II–847
20. Hare JS, Lewis PH, Enser PGB, Sandom CJ (2006) A linear-algebraic technique with an application in semantic image retrieval. In: Proc. CIVR, pp 31–40
21. Hare JS, Samangooei S, Lewis PH, Nixon MS (2008) Semantic spaces revisited: investigating the performance of auto-annotation and semantic retrieval using semantic spaces. In: Proc. CIVR, New York, NY, USA. ACM, New York, pp 359–368
22. Hayfron-Acquah JB, Nixon MS, Carter JN (2003) Automatic gait recognition by symmetry analysis. Pattern Recogn Lett 24(13):2175–2183
23. Hu W, Tan T, Wang L, Maybank S (2004) A survey on visual surveillance of object motion and behaviors. IEEE Trans SMC(A) 34(3):334–352
24. Interpol (2008) Disaster victim identification form (yellow). Booklet
25. Jain AK, Ross A, Prabhakar S (2004) An introduction to biometric recognition. IEEE Trans CSVT 14:4–19
26. Johansson G (1973) Visual perception of biological motion and a model for its analysis. Percept Phychophys 14(2):201–211
27. Kale A, Roychowdhury AK, Chellappa R (2004) Fusion of gait and face for human identification. In: Proc. IEEE ICASSP, vol 5, pp 901–904
28. Li X, Maybank SJ, Yan S, Tao D, Xu D (2008) Gait components and their application to gender recognition. IEEE Trans SMC(C) 38(2):145–155

29. Lindsay RCL, Martin R, Webber L (1994) Default values in eyewitness descriptions. Law Hum Behav 18(5):527–541
30. Little J, Boyd J (1995) Describing motion for recognition. In: Proc. ISCV, p 5A, Motion II
31. Liu Z, Sarkar S (2004) Simplest representation yet for gait recognition: averaged silhouette. In: Proc. ICPR, vol 4, pp 211–214
32. Liu Z, Sarkar S (2007) Outdoor recognition at a distance by fusing gait and face. Image Vis Comput 25(6):817–832
33. MacLeod MD, Frowley JN, Shepherd JW (1994) Whole body information: its relevance to eyewitnesses. In: Adult eyewitness testimony, chapter 6. Cambridge University Press, Cambridge
34. Macrae CN, Bodenhausen GV (2000) Social cognition: thinking categorically about others. Annu Rev Psychol 51(1):93–120
35. Monay F, Gatica-Perez D (2003) On image auto-annotation with latent space models. In: Proc. Multimedia, pp 275–278
36. Murase H, Sakai R (1996) Moving object recognition in eigenspace representation: gait analysis and lip reading. Pattern Recogn Lett 17(2):155–162
37. Nandakumar K, Dass SC, Jain AK (2004) Soft biometric traits for personal recognition systems. In: Proc. ICBA, pp 731–738
38. Nixon MS, Carter JN (2006) Automatic recognition by gait. Proc IEEE 94(11):2013–2024
39. Niyogi SA, Adelson EH (1994) Analyzing and recognizing walking figures in XYT. In: Proc. CVPR, pp 469–474
40. O'Toole AJ (2004) Psychological and neural perspectives on human face recognition. In: Handbook of face recognition. Springer, New York
41. Papadimitriou CH, Raghavan P, Tamaki H, Vempala S (1998) Latent semantic indexing: a probabilistic analysis. Comput Syst Sci 61:217–235
42. Pecenovic Z (1997) Image retrieval using latent semantic indexing. Master's thesis, AudioVisual Communications Lab, Ecole Polytechnique, F'ed'erale de Lausanne, Switzerland
43. Ponterotto JG, Mallinckrodt B (2007) Introduction to the special section on racial and ethnic identity in counseling psychology: conceptual and methodological challenges and proposed solutions. J Couns Psychol 54(3):219–223
44. Rosse C, Mejino JLV (2003) A reference ontology for biomedical informatics: the foundational model of anatomy. Journal of Biomedical Informatics 36(6):478–500
45. Samangooei S, Guo B, Nixon MS (2008) The use of semantic human description as a soft biometric. In: Proc. IEEE BTAS
46. Seely RD, Samangooei S, Middleton L, Carter JN, Nixon MS (2008) The University of Southampton multi-biometric tunnel and introducing a novel 3D gait dataset. In: Proc. IEEE BTAS
47. Shakhnarovich G, Lee L, Darrell T (2001) Integrated face and gait recognition from multiple views. In: Proc. IEEE CVPR, pp 439–446
48. Shutler J, Grant M, Nixon MS, Carter JN (2002) On a large sequence-based human gait database. In: Proc. RASC, pp 66–72
49. Tajfel H (1982) Social psychology of intergroup relations. Annu Rev Psychol 33:1–39
50. Troje NF, Sadr J, Nakayama K (2006) Axes vs averages: high-level representations of dynamic point-light forms. Vis Cogn 14:119–122
51. Van Koppen PJ, Lochun SK (1997) Portraying perpetrators; the validity of offender descriptions by witnesses. Law Hum Behav 21(6):662–685
52. Veres GV, Gordon L, Carter JN, Nixon MS (2004) What image information is important in silhouette-based gait recognition? In: Proc. IEEE CVPR, vol 2, pp II–776–II–782
53. Vrusias B, Makris D, Renno J-P, Newbold N, Ahmad K, Jones G (2007) A framework for ontology enriched semantic annotation of cctv video. In: Proc. WIAMIS, p 5
54. Wells GL, Olson EA (2003) Eyewitness testimony. Annu Rev Psychol 54:277–295
55. Yarmey AD, Yarmey MJ (1997) Eyewitness recall and duration estimates in field settings. J Appl Soc Psychol 27(4):330–344
56. Zhao R, Grosky W (2002) Bridging the semantic gap in image retrieval. IEEE Trans Multimedia 4:189–200



Sina Samangooei was awarded his MEng in Computer Science by the Department of Electronics and Computer Science at the University of Southampton in 2007. He is currently studying towards his PhD degree in Semantic Biometrics under Prof. Mark Nixon. His research interests include content based image retrieval, biometrics, anthropometry and semantic augmentation. Sina is a student member of the IEEE.

Mark S. Nixon is the Professor in Computer Vision at the University of Southampton UK. His research interests are in image processing and computer vision. His team develops new techniques for static and moving shape extraction which have found application in automatic face and automatic gait recognition and in medical image analysis. His team were early workers in face recognition, later came to pioneer gait recognition and more recently joined the pioneers of ear biometrics. Amongst research contracts, he was Principal Investigator with John Carter on the DARPA supported project Automatic Gait Recognition for Human ID at a Distance.

He chaired BMVC 98 and with Josef Kittler he chaired the Audio Visual Biometric Person Authentication (AVBPA 2003), was Publications Chair for the International Conference on Pattern Recognition (ICPR 2004) and the IEEE 7th International Conference on Face and Gesture Recognition FG2006. His vision book, co-written with Alberto Aguado, Feature Extraction and Image Processing was published in 2002 by Butterworth and with Tieniu Tan and Rama Chellappa, his new book Human ID based on Gait which is part of the new Springer Series on Biometrics, was published in 2005. Dr. Nixon is a member of the IEEE.