# Video tracking and person categorization system

***Dr.Kashif Qureshi***

Department Of Computer Science, tel: +966 (0)535598451,Jazan University, Jazan, Saudia

Srk1521@gmail.com

**ABSTRACT**

Video tracking is the process of locating a moving object (or multiple objects) over time using a camera. It has a variety of uses, some of which are: human-computer interaction, security and surveillance, video communication and compression, augmented reality, traffic control, medical imaging[1] and video editing.[2][3] Video tracking can be a time consuming process due to the amount of data that is contained in video. Adding further to the complexity is the possible need to use object recognition techniques for tracking.

To perform video tracking an algorithm analyzes sequential video frames and outputs the movement of targets between the frames. There are a variety of algorithms, each having strengths and weaknesses. Considering the intended use is important when choosing which algorithm to use. There are two major components of a visual tracking system: target representation and localization and filtering and data association.

Target representation and localization is mostly a bottom-up process. These methods give a variety of tools for identifying the moving object. Locating and tracking the target object successfully is dependent on the algorithm. For example, using blob tracking is useful for identifying human movement because a person's profile changes dynamically.[4] Typically the computational complexity for these algorithms is low. The following are some common target representation and localization algorithms:

Blob tracking: segmentation of object interior (for example blob detection, block-based correlation or optical flow)

Kernel-based tracking (mean-shift tracking): an iterative localization procedure based on the maximization of a similarity measure (Bhattacharyya coefficient).

Contour tracking: detection of object boundary (e.g. active contours or Condensation algorithm)

Visual feature matching: registration

Filtering and data association is mostly a top-down process, which involves incorporating prior information about the scene or object, dealing with object dynamics, and evaluation of different hypotheses. These methods allow the tracking of complex objects along with more complex object interaction like tracking objects moving behind obstructions.[5] Additionally the complexity is increased if the video tracker (also named TV tracker or target tracker) is not mounted on rigid foundation (on-shore) but on a moving ship (off-shore), where typically an inertial measurement system is used to pre-stabilize the video tracker to reduce the required dynamics and bandwidth of the camera system.[6] The computational complexity for these algorithms is usually much higher. The following are some common filtering algorithms:

Kalman filter: an optimal recursive Bayesian filter for linear functions subjected to Gaussian noise.[7]

Particle filter: useful for sampling the underlying state-space distribution of nonlinear and non-Gaussian processes.

We present a prototype video tracking and person categorization system that uses face and person soft biometric features to tag people while tracking them in multiple camera views. Our approach takes advantage of sequential aspect of video by extracting and accumulating feasible soft biometric features for each person in every frame to build a dynamic soft biometric feature list for each tracked person in surveillance videos. We developed algorithms for extracting face soft biometric features to achieve gender and ethnicity classification and session soft biometric features to aid in camera hand-off in surveillance videos with low resolution and uncontrolled illumination. To train and test our face soft biometry algorithms, we collected over 1200 face images from both genders and three ethnicity groups with various sizes, poses and illumination. These soft biometric feature extractors and classifiers are implemented on our existing video content extraction platform to enhance video surveillance tasks. Our algorithms achieved promising results for gender and ethnicity classification, and tracked person re-identification for camera hand-off on low to good quality surveillance and broadcast videos. By utilizing the proposed system, a high level description of extracted person's soft biometric data can be stored to use later for different purposes, such as to provide categorical information of people, to create database partitions to accelerate searches in responding to user queries, and to track people between cameras.

**Keywords: soft biometry, video surveillance, gender ethnicity, person categorization, person detection and tracking, face detection, session soft biometry, camera hand-off.**

**Section-1**

# 1. INTRODUCTION

Since the first deployment of video surveillance systems in mid-1960s [16], these systems have been used for security, safety and forensic evidence collection. Many of these systems are currently placed in public places and most of the data is collected by the law enforcement agencies for security purpose. However, despite the large amount of collected surveillance data, there is still little evidence supporting that such systems prevent the crime [1]. Rather, the collected data is found to be useful for identifying the offenders in post event analyses. The manual search of any evidence in previously collected video surveillance data is a tedious task due to the large volumes of video in archives.

In recent years, intelligent surveillance systems have been introduced to help to reduce this search time, however, current intelligent video surveillance systems can extract very limited knowledge about objects by classifying them into categories of person, vehicle, group, etc. with no further detail on the people. When human observers, on the other hand, see a person in a video, they quickly extract categorical information, such as gender and ethnicity, even from poor quality images without having to recognize and identify the individual. Soft biometric features are the characteristics that provide categorical information about an individual while they are insufficient to identify an individual reliably and uniquely due to its lack of distinctiveness and permanence [7]. Soft biometric traits help in filtering large surveillance databases by limiting the number of entries to be searched in for each query. Thus, extracting these features automatically from surveillance video will provide significant help to security personnel while protecting the privacy of individuals. In this study we investigate the use of soft biometric traits, such as gender and ethnicity, as well as session soft biometric features, such as clothing and hair color in video surveillance for automatic categorization of people.

Face images have been extensively used to identify the gender, ethnicity and other soft biometric traits in several studies [5,10,14,18]. However, current literature focuses on soft biometric features from still images or video sequences collected under controlled environments, several researchers have studied gender and ethnicity classification from still images using datasets containing well aligned frontal face images taken under controlled lighting. In [5], a mixture of experts (along with an SVM classifier for gating the input) is used for the classification of gender, ethnicity, and pose from face images. In [14] an appearance based gender classifier is used along with non-linear SVM. Both of these studies report over 90% accuracy on good quality face images from the FERET [24] database which contains very little expression and pose changes. In [18], a Perceptron based demographic classification scheme is used to extract faces from unconstrained video sequences and classify them based on gender and ethnicity (Asian or non-Asian). They use a variant of the AdaBoost for learning and feature selection, and their classifier achieves more than 75% accuracy under unconstrained environments for both gender and ethnicity classification. Moghaddam and Yang [14] reported very good classification results under controlled lighting using simple pixel intensities as features in SVM classifier. In [10], HMAX features are used to make soft biometric feature extraction robust to change in illumination, and slight shifts in viewpoint and occlusions.

Session soft-biometry features, such as person's skin tone, hair color, color of upper and lower-body clothing, which are constant over a short period of time, i.e., a "session" and hence are fundamentally different from permanent soft biometric features, which remain unchanged over a person's lifetime, such as gender or ethnicity. These features are used for identifying people going out of a camera's view and entering into another camera field of view [13, 23] in an environment with a network of cameras. The performance of these session soft biometric feature matching algorithms for tracking people between multiple camera views depend on the viewing angles of different cameras and the differences of lighting in each scene and the density of people moving around.

Extracting robust soft biometric features from surveillance video is a challenging task due to lower quality of data and uncontrolled environment factors, such as changes in illumination, resolution, camera view angle, occlusion, and shadows which can dramatically degrade the performance of soft biometric feature extraction. Despite these challenges, one major advantage of video data is that the availability of multiple samples of the face, and the person's snapshot at each frame as they move through the camera field of view. Our proposed approach takes advantage of this data redundancy to build a dynamic soft biometric feature vector to overcome the challenges.

The rest of the paper is organized as follows: Section 2 describes the proposed system and provides background on the enabling person and face detection algorithms used in our implementation. Section 3 is devoted to soft biometric feature extraction and classification including the face based gender and ethnicity classification and session soft biometric features for camera views. In Section 4, the experimental datasets, namely soft biometric face database, surveillance and broadcast video datasets, are explained. The experimental results are presented in Section 5; finally, our conclusions and future work directions are presented in Section 6.

**Section-2**

## 2. PROPOSED SYSTEM

In this paper, we propose to use a set of soft biometrics such as gender and ethnicity, as well as session soft biometric features, such as clothing and hair color in an intelligent video surveillance system for person categorization. The novel point in our approach to extracting soft biometric features from surveillance video is to exploit the dynamic nature of people moving in a cluttered scene. At different frames of video different biometric features of the people can be detected; in one frame the whole body of a subject can be detected to extract features such as height and clothing colors (while the face will be too small to detect); yet in another frame only the upper body might be visible allowing the subject's face to be detected reliably (to extract face based features) as illustrated in Figure 1. The proposed approach takes advantage of availability of a person in multiple frames of a video sequence by extracting and accumulating feasible soft biometric features from every frame, and building a dynamic soft biometry list for each tracked person. Our overall system (See Figure 2) includes detection and tracking of whole bodies and faces in video streams; modules to extract gender, ethnicity and session soft biometric features; representation and manipulation of soft biometry metadata to support entity matching between cameras and systems.


Figure 1. Each frame provides different soft biometric features for a tracked person.

We extract two categories of soft biometric features: (i) facial features for gender and ethnicity detection, and (ii) clothing and hair color (session soft biometry) features to facilitate camera hand-off.

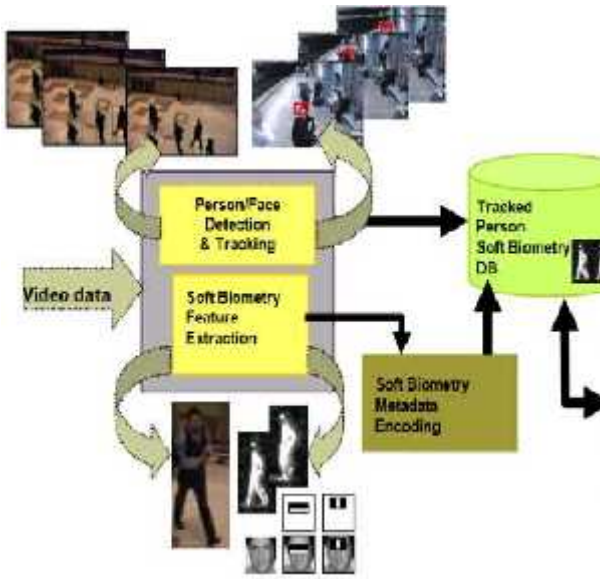



Figure 2. High-level diagram for the proposed Video Soft Biometric Feature Extraction System.

The proposed system is built on into Vision's intelligent video platforms .The detection, tracking and soft biometric feature extraction tasks are organized in layers as shown in Figure 3. First layer is responsible from detection and tracking people and their high level pose and faces, while the second layer extracts session and face soft biometric features and estimates their confidences to create the soft biometry metadata to be used in the other layers the system. The person and face detection and tracking in surveillance videos are discussed in Sections 2.1 and 2.2.

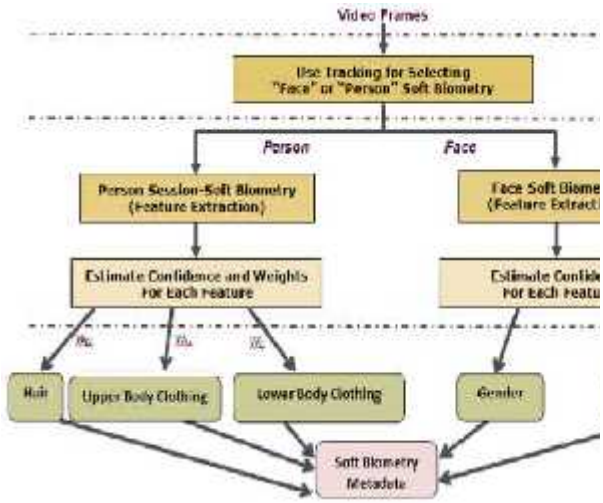



Figure 3. Conceptual organization of detection, tracking and soft biometric feature extraction phases.

### 2.1 Person Detection and Tracking

In busy scenes with a large number of people moving around and frequently occluding each other it is difficult to reliably estimate a scene background model for moving object detection. Single frame detection methods based on transform cascades or recognition have shown promising results for detection and subsequent tracking in such scenarios. While several other algorithms have been proposed and used for detecting people in surveillance scenarios, most of them are too slow for real time applications. For example, in Dalal and Triggs use histogram of

gradients features to robustly detect pedestrians, which takes around 0.5 seconds to process a 128x64 pixel image. Typical surveillance video frame resolution is 480x640 pixels and it would require several seconds to process one frame. Similarly, biologically inspired model used in [3] takes approximately 80 seconds per frame. Our person detection algorithm is based on the object detection algorithm using Haar features proposed in [21]. This algorithm provides a high recall rate with almost real-time performance and has been successfully used in detecting human faces and other types of objects, such as human hands [2], people [15], and upper body [9].

In our algorithm, Haar-features are extracted to build an over complete set of appearance features, which are then utilized to train a cascade of weak classifiers using Adaboost. Although the training period is long, a trained classifier detects people at real-time speeds. Once a person is detected at a given frame, we use motion information (as a correspondence base filtering) to track people in consecutive frames as detailed in [25]. Figure 4 shows some people detection results at a crowded airport environment by using this approach.

Figure 4. People detection results in busy indoors scenes.

#### 2.1.1 Estimation of Person Size and Directional Pose

For people detected in a given frame, we estimate their size and directional pose with respect to the camera to determine which soft biometric features can be most reliably extracted. For instance, if in a given frame a person is detected as moving towards the camera (indicates a frontal view), his/her face would be visible for face detection and extraction of *soft biometry* features such as gender and ethnicity; whereas if a person is moving away from the camera (observed from the back), the hair-color can be extracted as a *session soft biometric* feature. To determine the directional pose of a person, we utilize features such as the height to width ratio of the bounding box enclosing the person, velocity direction, and the change in the size of the person in consecutive frames. The height to width ratio information is also used to determine if the person's body is fully or partial visible and whether the person is being viewed from side or front/back. We also use object motion (velocity) and Recurrent Motion (a measure of articulate motion of the object as developed in [8]) features to estimate the pose angle of a person in the scene to estimate the side versus front or back views of a person's pose- angle with respect to the camera. Figure 5 shows examples showing Recurrent Motion Image (RMI) features for two people - one moving across camera field of view (side view) and the other moving towards the camera (front view). As seen in accumulated RMIs, side and front views are very different; especially in the lower parts of RMIs where there are high motion regions due to the leg movements.

### 2.2 Face Detection and Registration

Once a person is detected and passes the criteria of including a face (frontal view and sufficient resolution), we further check for the presence of face within the detected bounding box. Our face detection system uses an Adaboost classifier trained on Haar like features [22]. To improve computation efficiency, we use only the top half of the bounding box as an input to our Face classifier (See Figure 6). After a face is detected, it is aligned according to the defined common

Figure 5. Person detection and RMI for two body poses.

coordinate system, and a scale dependent face-mask is used to extract central face region (Region of Interest - ROI) devoid of hair or background. These pixel values of central region are used for the gender and ethnicity categorization.

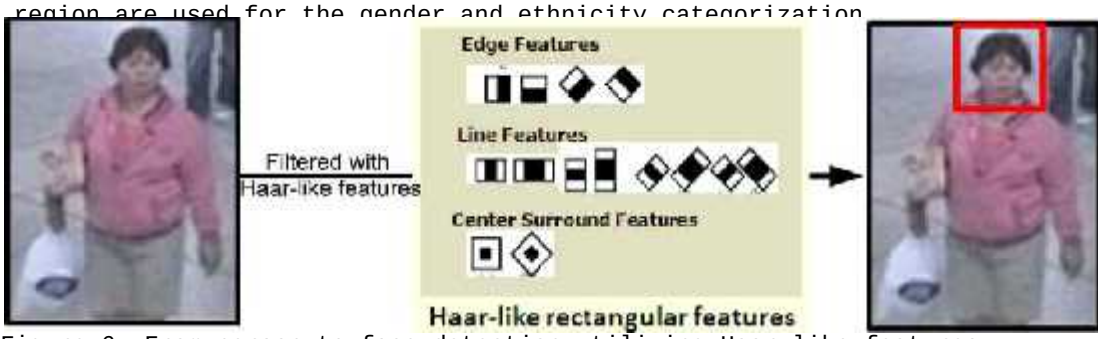



Figure 6. From person to face detection utilizing Haar-like features.

**Section-3**

## 3.SOFT BIOMETRIC FEATURE EXTRACTION

As mentioned earlier, this work focuses on two types of soft biometric features that are useful in video surveillance applications: (i) face-based soft biometric features for gender and ethnicity classification, and (ii) session soft biometric features, which consist of features such as clothing and hair color. Session soft biometric features are quite stable over shorter durations and are often used in camera hand-off.

### 2.3 Face-Based Gender and Ethnicity Classification

Human faces can be easily categorized into different types of gender and ethnicity based on the facial features and their geometry. Previous literature has studied gender and ethnicity classification only from still images using datasets containing well aligned frontal face images taken under controlled lighting. Our work extends the previous work by developing algorithms for gender and ethnicity recognition that not only work well on still images but also performs well on good to low quality video images. We present two algorithms. The first algorithm utilizes pixel intensity values while the other one uses Biologically Inspired Model (BIM) features for soft biometry computation. Both of these methods use SVM for classification.

#### 2.3.1 Pixel Intensity-Based Gender and Ethnicity Classification

In the first algorithm, pixel intensity-based features were used as a simple and fast method to implement real time gender and ethnicity detection. Detection of the intensity-based facial features firstly requires the registration of the faces. Thus, we define a common coordinate system for the face images in the database: the middle point between the right and left eye is set to be the origin of the new coordinate system. Automated pupils detection is achieved utilizing **Haar features**[Haar-like features are digital image features used in object recognition. They owe their name to their intuitive similarity with Haar wavelets and were used in the first real-time face detector.

Historically, working with only image intensities (i.e., the RGB pixel values at each and every pixel of image) made the task of feature calculation computationally expensive. A publication by Papageorgiou et al.[1] discussed working with an alternate feature set based on Haar wavelets instead of the usual image intensities. Viola and Jones[2] adapted the idea of using Haar wavelets and developed the so called Haar-like features. A Haar-like feature considers adjacent rectangular regions at a specific location in a detection window, sums up the pixel intensities in each region and calculates the difference between these sums. This difference is then used to categorize subsections of an image. For example, let us say we have an image database with human faces. It is a common observation that among all faces the region of the eyes is darker than the region of the cheeks. Therefore a common haar feature for face detection is a set of two adjacent rectangles that lie above the eye and the cheek region. The position of these rectangles is defined relative to a detection window that acts like a bounding box to the target object (the face in this case).

In the detection phase of the Viola–Jones object detection framework, a window of the target size is moved over the input image, and for each subsection of the image the Haar-like feature is calculated. This difference is then compared to a learned threshold that separates non-objects from objects. Because such a Haar-like feature is only a weak learner or classifier (its detection quality is slightly better than random

guessing) a large number of Haar-like features are necessary to describe an object with sufficient accuracy. In the Viola–Jones object detection framework, the Haar-like features are therefore organized in something called a classifier cascade to form a strong learner or classifier.

The key advantage of a Haar-like feature over most other features is its calculation speed. Due to the use of integral images, a Haar-like feature of any size can be calculated in constant time (approximately 60 microprocessor instructions for a 2-rectangle feature).

]. In our experiments some face images, with low contrast or eyes closed/occluded, were annotated manually. The angle between the eye points is used for rotating all the face images to a common horizontal line whereas the eye locations are utilized to translate the image according to a common coordinate system (See Figure 7a). Once all of the faces are rotated for horizontal alignment, the distance between eyes is used for mapping all the face images to a common scale. Based on the inspection of the eye distance histogram (See Figure 7b), the mean and the median are 31.66 and 30 pixels, respectively; thus, the common eye distance is empirically set to 30 pixels.

After the registration phase, we generate a face-mask that outlines the face region of interest (ROI) utilizing the aligned average faces. Example masked faces, each of which has the size of 61x66, are as shown in Figure 8. We used these masked face images for feature extraction for the gender and ethnicity classification. After masking, so defining the ROIs, images are converted to gray-scale and normalized using min-max normalization rule, which are later used as features. We chose 206 pixels as features and scaled the values between [-1, +1]. SVM classifiers were used to perform the classification. As explained in Section 5, we trained separate SVMs (using the SVM implementation of [12]) to do gender and ethnicity classification for six gender-ethnicity classes at once or separately, i.e. classification for gender (2- class) and classification for ethnicity (3-class). Our results, as presented in the Section 5.1 with this simple but effective algorithm, are promising for good to low resolution broadcast and surveillance video sequences.

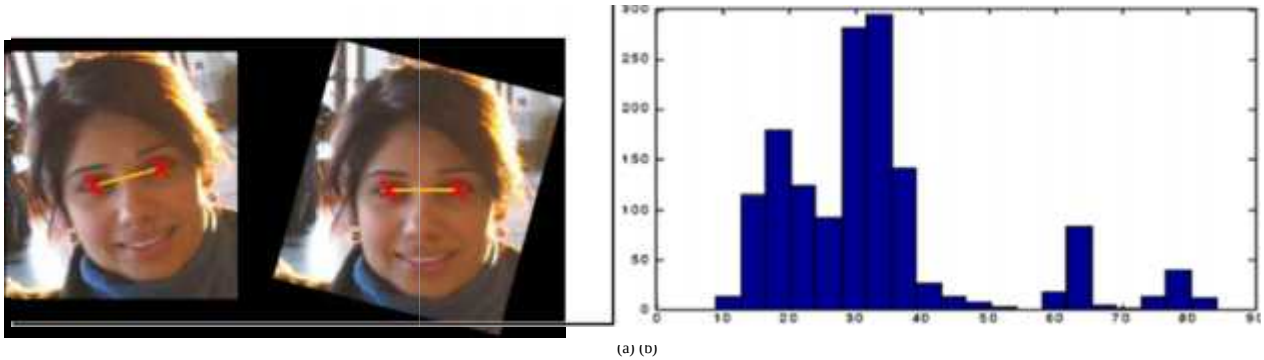



Figure 7. (a) Face registration (red marks show annotation points, the yellow line shows the angle between from the horizontal), (b)
eye distance histogram of into Vision face database.

Figure 8. Aligned, scaled and masked face image ROIs. (Images are shaded to protect the anonymity of the subjects.)

### 2.3.2 Biologically Inspired Model Feature Based Gender-Ethnicity Classification

Our second algorithm uses the Biologically Inspired Model (BIM) to extract illumination, scale, and shift invariant features from face images. These features are also known as **HMAX features**[Research in the lab is based on a computational model of object recognition in cortex (Riesenhuber & Poggio, *Nature Neuroscience*, 1999), dubbed HMAX ("Hierarchical Model and X") by Mike Tarr (*Nature Neuroscience*, 1999) in his News & Views on the paper. Since we didn't think of a better name beforehand, HMAX stuck. Oh well...

The model summarizes the basic facts about the ventral visual stream, a hierarchy of brain areas thought to mediate object recognition in cortex. It was originally developed to account for the experimental data of Logothetis *et al.* (*Cerebral Cortex*, 1995) on the invariance properties and shape tuning of neurons in macaque infer temporal cortex, the highest visual area in the ventral stream. In the meantime, the model has been shown to predict several other experimental results and provide interesting perspectives on still other data and claims. The model is used as a basis in a variety of projects in our and other labs.

The goal is to explain cognitive phenomena in terms of simple and well-understood computational processes in a physiologically plausible model. Thus, the model is a tool to integrate and interpret existing data and to make predictions to guide new experiments. Clearly, the road ahead will require a close interaction between model and experiment. Towards this end, this web site provides background information on HMAX, including the source code, and further references:

The "Standard Model" of object recognition

] .We select the model since the BIM features can handle the partial occlusions, slight scale variation, and changes in illumination, which are common problems in video surveillance. The BIM as the name suggests, emulates object recognition in human cortex and has experimentally shown to perform better than some current state of art features extractors for several object detection/classification tasks .

BIM is an advanced method that computes features that exhibit a better trade-off between invariance and selectivity than template-based or histogram-based approaches. Template based approaches are too closely tied to the geometry and the pixel values of an image. On the other hand histogram approach, are too general and do not encode explicit shape and spatial information. BIM provides much better

representation. It computes an intermediate representation of image called 'C1'. This intermediate representation is computed by applying Gabor filter to an image and then taking the maximum value of the filtered output within a local spatial region and two slightly different image scales. Using high frequency Gabor filters, reduces BIM sensitivity to variation in illumination, while taking a max over spatial neighbor makes it invariant to small shifts in position and scale. BIM can be thought of as a more sophisticated way of template matching with position and scale invariance built in. For more details on the BIM features please refer to [19].

In this work we compute BIM features along with SVM classifier to obtain BIM based Gender-Ethnicity classifier. Here we briefly describe the BIM algorithm [19]. The BIM algorithm has two main phases - training and a testing phase. In the training phase, we first compute intermediate representation C1 for all the positive training images using Gabor filters, as described above. From this set of C1 images, template patches of different sizes are randomly selected. The set of these template patches create a dictionary defining a positive class. In other words, an image with higher number of these template patches will more likely be an image from positive class.

Once the dictionary is computed these template patches are used to compute features and train SVM. Each template patch gives a feature, thus the number of features is equal to the number of template patches in the dictionary. Each template patch is correlated over an entire image and the maximum correlation value defines the feature value. Features computed in this manner for positive and negative training dataset is used to train SVM. During run time, again the intermediate image is used and correlated with template patches to compute feature values, which are passed to SVM for prediction. For further details we refer the reader to [19].

### 3.2. Session-Soft Biometrics Features

Our session biometry algorithm extracts skin tone color, hair color, and color of clothing. These features are then used to compare if an object exiting from field of view from a camera matches to the object that enter in the field of view of an adjacent camera. Using these session-soft biometry features can help us handoff the tracking ids more accurately from one camera to the other cameras. Here we briefly discuss how the various session biometry features are computed.

**Skin tone** - The algorithm detects the skin pixels in an image and stores the color information of the corresponding skin pixels. We convert the given color image region into HSV space and then use thresholds in hue and chromaticity to detect skin tone pixels. The skin tone color is stored as a 2D histogram of hue and saturation.

**Hair Color** - For hair color we only look at the top portion of the face region (for people in frontal pose) or the whole head region. We use skin tone detector to reject skin regions at the top of face region. In addition, foreground mask further restrains from using color information from background regions. As before, image is converted into HSV space to make the color invariant to illumination. The color information is stored as a 2D histogram of hue and saturation.

**Shirt and Trouser Color** - For computing shirt and trouser colors, we first convert the color image into HSV image and then use hue and saturation to compute 2D histogram of middle (for shirt) and bottom (for trousers) region of the person's bounding box. A foreground mask is used to select color information from only foreground pixels.

### 3.2.1 Tracking People between Camera Views

## The Need for Viseum IMC

Random unpredictable crimes and anti-social behaviour usually happen where cameras are not pointing, yet evidence needed to deter and clear up crimes is only caught if the camera is pointing close up wherever it occurs. No matter how many PTZ cameras are installed in the many public open spaces, they are pointing or touring the wrong way when crimes occur unless staff are watching them at the right times and controlling them in the right directions.

## How Viseum IMCs work

Depending on the coverage required, one or more fixed cameras, used in conjunction with Viseum licensed PTZ cameras, and the Auto-intelligent Surveillance Unit “AiSU”, provides unrivalled CCTV protection for open space where random crimes occur.

The information gathered from the fixed cameras is used to detect anyone entering your areas, or pre-empt criminal or anti-social behaviour. This then automatically instructs the PTZ to zoom into and follow the subject(s) providing evidential quality recordings.

## Viseum-driven Auto-intelligent Camera Units "AiCU"

For complete coverage the Viseum-driven Auto-intelligent Camera Units “AiCU” is well proven to capture crucial evidence that would have otherwise been missed. This patented multiple camera unit was designed to provide large scale, practical, cost effective and aesthetic deployment of the Viseum IMC technology. Its versatility and appearance makes it suitable for a wide variety of environments and applications.

Thanks to up to 7 fixed cameras that constantly survey the surrounding areas, each AiCU offers up to 360° constant visibility protecting an area the size of almost 14 football pitches.

## Each 360° Viseum IMC Equivalent

It would take at least 80 fixed cameras costing at least £60,000 to provide the same level of PSDB recommended identification evidence without human intervention, or 40 camera operators to constantly monitor this many cameras proactively 24/7/365, costing over £1 million per year.

## Key Benefits

Viseum IMCs provide up to 25 times more court quality evidence than any other solution, by monitoring the many public open space areas where random crimes occur. This is achieved by providing greater coverage than multiple fixed camera installations, and better performance than intermittently manned moving cameras. In short, Viseum IMC technology is making CCTV surveillance less complex, more effective and at the same time more affordable.

- **More evidence of crime is produced - improving missed evidence statistics -** Wide view and close up recordings of the same event provides evidence of the lead up to the crime as well as the aftermath - as recommended by UK Home Office
- **Moving camera is consistently seen to be reacting intelligently to provide long term deterrent -** Proven to deter criminals from planning their crimes
- **The enhanced crime clear up and deterrent at your Viseum IMC sites will reduce the burden of ongoing cost and fear of crime -** Less crime to deal with and your communities feel safer
- **Each Viseum IMC installation frees your key staff to focus on other security issues e.g. hot spot/high activity/RIPA and revenue cameras -** Progressively creating greater performance for your entire CCTV network and making your key staff more productive and better motivated than ever before

- **The single up-front cost per camera site is less than two-thirds of the yearly costs of just one additional member of staff** - Ongoing savings mean investment will be recovered in a shorter period of time and provide life-time savings thereafter
- **With intelligence local to the camera installation, many areas benefit from reduced communication costs:**
  - Certain areas would not need any remote communications i.e. used as a standalone system
  - Certain areas would downgrade their fibre optics for a lower bandwidth solution

## Auto-intelligent Camera Unit (AiCU) includes:

Up to 8 high resolution cameras in one unit - extended dual evidence of any incident:

- Up to 7 fixed high resolution cameras for up to 360 degrees coverage - detection and overview evidence
- 1 x high resolution Pan, Tilt and Zoom camera - close up evidence

## Auto-intelligent Surveillance Unit (AiSU) includes:

- High Performance Dual Core Computer
- Virtual Operator Assistant (VOA) Software for intelligent camera control
- Optional Auto-intelligent Digital Video Recorder (AiDVR) Software + HDD's for 30 days local video storage

Alternatively all cables can be fed into any other DVR and patched though to the AiSU.

## Viseum IMC USPs

- The most effective method to automatically protect open space the size of almost 14 football pitches per single camera installation.
- Suspect recognition evidence up to 147 metres in every direction.
- PTZ switching provides automatic close up evidence even if several events occur at the same time.
- Preventing diversion tactics by remembering and predicting activity to enable accurate PTZ switching between any number of multiple events
- Viseum IMCs are managed over IP using any PC devise (laptop, PDA, or mobile phone).
- Viseum-driven camera installations enable analogue cameras to benefit from secure IP communications freedom
- Remote support from the Viseum support team
- Typically over 3G, no matter what your problem or need, the Viseum support team can be plugged into your products for extra support
- Can include recording or can be an add-on to existing recording systems
- Use your favourite or existing DVR, or benefit from the compact all-in-one-box recording detection and camera control system
- Very easy to install, use, and maintain
- All cameras on one structure makes it very easy to deploy, set up and operate. Software very simple to use

## CCTV Control Room Integrated

Providing the best of both worlds where machine-to-machine intelligent automation interacts with human vigilance.

The value of your CCTV infrastructure depends entirely on its effectiveness as a crime clear up tool and the resulting long term deterrent effect. By having some of your camera installations Viseum

IMC enabled your entire infrastructure benefits from the additional support. Your camera operators can always seamlessly override

Viseum's automation at any time, but when they are too busy or simply unavailable and crimes are committed, high quality close up evidence will still get captured, and the long term deterrent will be sustained.

## Re-deployable

The majority of re-deployable moving cameras are becoming Viseum IMC enabled. This is because missed opportunities due to operator fatigue are reduced, and significant man hours are saved in re-deploying and adjusting camera positions.

## Standalone

During tracking people in a network of surveillance camera views these session soft biometric features are used to compare people. If a match is found, the same identity tag (ID) is assigned to the person entering a view from an adjacent camera; otherwise a new ID number is assigned. Figure 9 shows handoff between two cameras on a stairwell. These are non-calibrated cameras with non-overlapping views. Since the system is un-calibrated, dimension measurements, unlike the work in [13], cannot be reliable. Thus, we only employ the color session soft biometric features described above. In calibrated systems, more features (height and built and anthropometric ratios) can also be computed. Use of additional features will make camera hand-off more robust especially in crowded scenes where several people will be moving between camera views.

**camera camera 2 camera 3 camera 4 camera 5 camera 6**



.

## DATASETS FOR TRANING TRAINING AND TESTING GENDER AND ETHNICITY CLASSIFIERS

We have collected a large database of face images and videos consisting of large variation in pose, illumination, resolution, and image quality to test the efficacy of our soft biometry classifier. Our database consists of still images as well as face images extracted from web and surveillance videos.

### 2.4 Still Image Face Datasets for Gender and Ethnicity Classification Training

We have combined images from several databases collected from different public face databases and WWW to creating a single database with various face image sizes, varying poses, different illuminations and large number of subjects in both genders and three ethnicity groups as illustrated in Figure 10. Since the ethnicity groups are somewhat loosely defined, we limited the ethnicity classes to three as Caucasian, Asian and African-American, and two (male and female) gender classes. This dataset is used for the six-class classification problem.

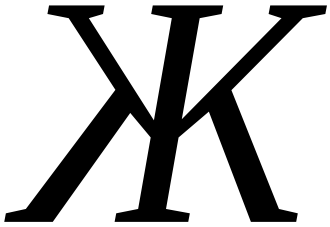
**Gender Database** contains 635 female (100 African-American, 335 White, and 200 Asian) and 823 male (181 African American, 486 Caucasian, and 156 Asian) unique face images.

**Ethnicity Database** contains a non-biased gender- ethnicity combination, with 200 African American (100 female, 100 male), 200 Caucasian (100 female,
100 male) and 200 Asian (100 female, 100 male) unique face images.

This dataset provides large variability, and is a representative basis for faces in surveillance or broadcast video. The figure shows sample faces from the database (shaded to protect the anonymity of the subjects) and the average face obtained after face registration. The last row shows faces from FERET [24] database with good quality, controlled lighting and face pose and the other rows show samples with varying image quality, pose and uncontrolled illumination.

### 2.5 Web and Surveillance Video Datasets

We have created two video databases; one with reasonable good quality (TV broadcast news and shows videos from WWW) shown in Figure11-a, and another with the low quality surveillance videos) (from subway video surveillance) as shown Figure 11-b.



(a)

(b)

Figure 11. (a) Good-quality broadcast video frames (image courtesy of ABC, Fox TV, HBO and MSNBC) , (b) Low-quality video surveillance frames.

## Section-3

## 3. EXPERIMENTAL RESULTS AND PROTOTYPE SYSTEM

Here we present the results of our pixel intensity-based and BIM feature based gender and ethnicity classifiers and use of these classifiers in our prototype system. In all our experiments we have used 4-fold cross validation to train and test the SVM classifiers [12]. We have tested the classification accuracy on two different face sizes images to observe the effect of resolution on the classification performance.

### 3.1 Pixel Intensity-Based Features: Experiments with Still Images

In Table 1, we present the classification results for three separate experiments: gender only, ethnicity only, and gender and ethnicity together. These experiments provided very promising results for the low-resolution face images of only size of 16*15, which is typical face size for surveillance videos.

To compare the performance of separate and combined gender and ethnicity classification, we also ran an experiment for gender independent ethnicity classification. These experiments indicate that with the sample sizes of our dataset it is more efficient to perform gender and ethnicity classification separately.

e 1: Gender, ethnicity, and one step gender-ethnicity recognition accuracies (%).

| **Image Size** (original size of 66x61) | **Image Scaling method** | **Gender (2Classes - 823 male, 635 female)** | **Ethnicity (3 Classes - 200 faces per class)** | **Gender&Ethnicity (6 Classes - 100 faces per** |
|---|---|---|---|---|
| [33 30] | Bi-cubic | 88.4 | 88 | 70.6 |
| [33 30] | Bi-linear | 90.4 | 94 | 74.6 |
| [16 15] | Bi-cubic | 89.3 | 87.3 | 78 |
| [16 15] | Bi-linear | 89.4 | 92.6 | 81.3 |

### 3.2 Pixel Intensity-based Features: Experiments with Web and Surveillance Video Data

For these experiments, we used the video databases described in Section 4.2. Based on reported cross validation results, we utilized the following parameter values for the experiments discussed below: face image resolution is set to 16x15; bi-linear interpolation is used; SVM parameter C and y are empirically set to 32 and 0.007, respectively. Furthermore, we utilized all the images in intuVision face database for the training purpose to obtain better accuracy. For the two different video-stream databases, we obtained the following results:

**Gender recognition experiment on good quality videos:** Using the pixel intensity features, we obtained the performance of 90% on 50 female and 50 male images that are extracted from broadcast videos (See samples in Figure 12 top row).

**Gender recognition experiment on low quality videos:** Using the same method, we achieved 70% performance on 50 female and 50 male images that are extracted from surveillance videos (See samples in Figure 12 bottom row).

The major reasons for degraded performance from 90% to 70% are: 1) very low image quality as shown above, 2) analog camera interlacing issue, and 3) due to the camera angle not having many frames where a person's face is observable.

Figure 12. Top row: Faces extracted from broadcast video; Bottom row: faces extracted from surveillance video. (Images are shaded to protect the anonymity of the subjects)

### 3.3 BIM Features: Experiments with Web and Surveillance Video Data

For BIM experiments, we selected 300 features from 200 images of size 66x61 from intuVision Face Database (See Figure 11). 100 images were used for training the SVM classifier. Testing was done on 100 images that are not included in the training set. The BIM classifier gave very high gender classification results (accuracy of 95%) for FERET [24] database images (See Table 2). However, the classification rate quickly degraded for good (broadcast) and low quality (surveillance video) face images. For good and low-resolution images we obtained classification accuracy (See Table 3).

Table 2: Gender Recognition for 2 Class -1094 faces, Test Set: 364 faces (4 fold cross-validation).

| Image Size | Optimized accuracy | Optimized SVM parameters |
|---|---|---|
| 66x61 good quality | 95% | C=8.0, y =0.0125 |
| 66x61 low quality | 60% | C=8.0, y =0.125 |

Table 3: Ethnicity Recognition for 3 Class - 200 faces for each class (4 fold cross-validation).

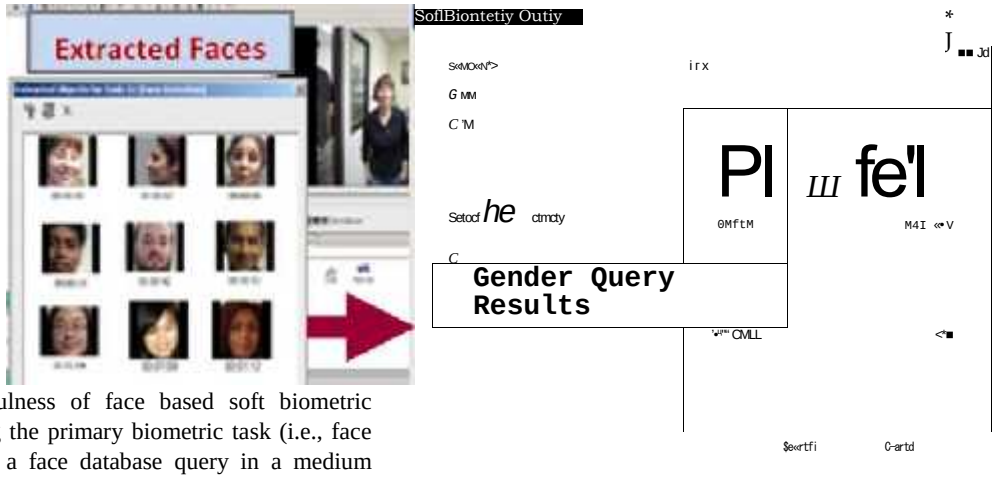
| Image Size | Optimized accuracy | Optimized SVM parameters |
|---|---|---|
| 66x61 (Asian vs. others) | 85.0% | **C=32.0,γ=0.007** |
| 66x61 (African vs. others) | 86.5% | C=8.0, **γ** =0.007 |
| 66x61 (Caucasian vs. others) | 78.5% | C=2.0, **γ** =0.031 |

lower than the pixel intensity-based approach (See Table 2). This degradation in performance in low-resolution images can be attributed to the lack of orientation features in low quality images. This issue can be resolved by using super-resolution to obtain higher quality images from multiple frames of person's face in a video [20]. On the other hand, we have observed that the ethnicity experiments on into Vision Face Database utilizing BIM features seems more promising compared to the gender recognition. We obtain minimum 78.5% accuracy for the comparison of different ethnicities

### 3.4 Video Indexing with Face-Based Gender and Ethnicity

We implemented a face-based gender and ethnicity indexing and query module into our existing video post event analysis tool VideoAnalyst™ (using the algorithms described in Section 3). VideoAnalyst has a built-in face detection module to extract faces from surveillance and broadcast videos. We use the extracted faces as input to the soft biometric gender-ethnicity classifiers. The trained gender and ethnicity classifiers then detect faces with the specified gender and/or ethnicity from all extracted faces as illustrated in Figure 13. Currently, we employ classifiers for detecting both gender (male and female) and three ethnicity classes (African-American, Asian and Caucasian).

Figure 13.




Soft Biometry query in Video Analyst for finding "male" faces in an indoors surveillance video.

### 3.4Face Database Partitioning with Soft Biometric Features

To evaluate the usefulness of face based soft biometric features in augmenting the primary biometric task (i.e., face recognition), we used a face database query in a medium scale face recognition task. We performed an experiment to quantify the efficiency of ethnicity-based database partitioning in face matching. In the first case, we matched an Asian male using the entire sample male dataset which consist of 600 images (200 images each from three different ethnicities and with equal number of males and females). We used Eigenfaces based face recognition algorithm written in MATLAB which takes 1 min for each pair of matching. The first run took 10 hrs for face matching. In the second case, we ran the same query against the ethnicity partitioned database using our ethnicity classifier. Then, we used soft- biometric tags to narrow down the recognition search to only Asian male faces within the database, taking only 1hr and 40 min compared to 10 hrs in the previous case. Using soft biometric features reduced the time performance of face recognition search tasks by almost a factor of 6. The database has to be tagged with soft biometric features and partitioned in advance, but this needs to be done once and can be done offline when database is not in active use.

## Section-4

## 4. SUMMARY AND FUTURE WORK

We developed algorithms for face and session soft biometric feature extraction to provide gender and ethnicity classification and to aid in camera hand-off of tracked people in surveillance video. We implemented these feature extractors in existing video content extraction platforms to enhance video surveillance tasks.

By utilizing the proposed system, a high level description of extracted person categorization data can be stored to provide categorical information and to create database partitions to accelerate searches in responding to user queries. In addition to the security or database search applications, this technology could be applicable to other fields, such as market research. For instance, a retailer may wish to determine how many women stop to view certain products displayed on a certain aisle or end-cap.

For face-based gender and ethnicity detection there seems to be a need in especially law enforcement applications and we would like to improve the performance accuracy to make this system viable. The simple pixel intensity-based features do not produce accurate results if the faces are not aligned and masked properly, BIM-feature-based approach does not require registration and masking but its performance depend on the type of video and image resolution. We believe that incremental learning can give better performance while not requiring a huge database for training. The idea is to train the classifiers to improve their performance s they are used on more videos. The system will provide a base classifier which produces good classification results without being specific to any particular video data set. To improve the performance users will incrementally train the classifier by selecting wrongly classified examples and giving them as input to the trainer. This way the users will improve the classifier as they apply it to new videos in their domain.

Our current algorithms use frontal faces and need to be expanded for gender and ethnicity classification from profile and % faces to test the limits of view-point invariant classification. Another area of future work is the use of face super resolution algorithms. Our initial experiments in this area were not conclusive and the time performance of the super resolution algorithm was prohibitive, hence more investigation is needed in this area. Currently, the work on improving gender and ethnicity classification in surveillance quality video is ongoing at intuVision.

**Acknowledgements** This work is partially supported by a US Government Small Business Innovation Research project grant. Authors would like to thank Dr. Arun Ross for his collaboration in the project. Thanks are also due to Dr. Tal Arbel and Dr. James J. Clark for their valuable suggestions on this manuscript.

## REFERENCES

1. Baram, M., "Eye on the City: Do Cameras Reduce Crime? " ABC News, 2007-07-10.
2. Barczak, A., L., C., Dadgostar, F., and Johnson M., J., "Real-Time Hand Tracking using the Viola and Jones Method," Signal and Image Processing, 2005.
3. Bileschi, S. and Wolf, L., "Image representations beyond histograms of gradients: The role of Gestalt descriptors," CVPR 2007.
4. Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," CVPR, 2005.
5. Gutta, S. , Huang, J.R.J. Jonathon, P. Wechsler, H., "Mixture of Experts for Classification of Gender, Ethnic Origin, and Pose of Human Faces," IEEE Transactions on Neural Networks 11, 948-960 (2000).
6. Huang, Y., Huang, K., Tao, D., Wang, L, Li, X., and Tan, T., "Enhanced Biologically Inspired Model," IEEE Int. Conference on Computer Vision and Pattern Recognition (CVPR), 2008.
7. Jain, A., K. and Lu, X., "Ethnicity Identification from Face Images," Proc. of SPIE Int. Symp. on Defense and Security : Biometric Technology for Human Identification, 2004.
8. Javed, O. and Shah, M., "Tracking and Object Classification for Automated Surveillance," ECCV, 2002.
9. Kruppa, H., Castrillon, M. and Schiele, S.B., "Fast and Robust Face Finding via Local Context," Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2003.
10. Lapedriza, A., Marin-Jimenez, M. and Vitria, J., "Gender Recognition in Non Controlled Environments," ICPR, (3), 2006.
11. Lienhart, R. and Maydt, J., "An Extended Set of Haar-like Features for Rapid Object Detection," ICIP, 2002.
12. Chang, C. and Lin, C., "LIBSVM: A Library for Support Vector Machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.
13. Madden, C. and Piccardi, M., "Height Measurement as a Session-based Biometric for People Matching Across Disjoint Camera Views," Image and Vision Computing New Zealand, 2005.
14. Moghaddam, B. and Yang, M., H., "Learning Gender with Support Faces," IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 707-711 (2002).
15. Monteiro, G., Peixoto, P. and Nunes, U., "Vision-Based Pedestrian Detection Using Haar-Like Features," Robotica, 2006.
16. Roberts, L., P., "The history of video surveillance," http://www.video-surveillance-guide.com.
17. Riesenhuber, M. and Poggio, T., "Hierarchical models of object recognition in cortex," Nature Neuroscience, 2, 1999.
18. Shakhnarovich, G., Viola, P. and Moghaddam , B., "A Unified Learning Framework for Real Time Face Detection and Classification, " Proc. of Int. Conf. on Automatic Face and Gesture Recognition, 2002.
19. Serre, T., Wolf, L. and Poggio,T., "Object recognition with features inspired by visual cortex, " CVPR, 2005.
20. Park, S., C., Park, M., K., and Kang, M. G., "Super-resolution image reconstruction: a technical overview, " IEEE Signal Processing Magazine, 20(3), 21-36 (2003).
21. Viola, P., A. and Jones, M., "Rapid object detection using a boosted cascade of simple features, " CVPR, 2001.
22. Viola, P., A. and Jones, M., J., "Robust Real-Time Face Detection," Int. Journal of Computer Vision, 2004.
23. Wang, Y., F., Chang, E., Y. and Cheng, K., P., "A Video Analysis Framework for Soft Biometry Security Surveillance," Int. Workshop on Video surveillance & Sensor Networks, 2005.
24. "Color FERET face database," www.itl.nist.gov/iad/humanid/colorferet.
25. Yarlagadda, P., Demirkus, M., Garg, K. and Guler, S., "Intuvision Event Detection Systems for TRECVID 2008," TREC Video Retrieval Evaluation Competition Conf., Nov. 17-18, Washington, DC.
26. www.wikipedia.com