Electronic Letters on Computer Vision and Image Analysis 13(1):18-32, 2014

Autonomous UAV for Suspicious Action Detection using Pictorial Human Pose Estimation and Classification

Surya Penmetsa*, Fatima Minhuj*, Amarjot Singh+, S.N. Omkar−

* *Department of Electronics and Communication Engineering, National Institute of Technology, Warangal, India.*
+ *School of Engineering Science, Simon Fraser University, Burnaby, Canada.*
− *Department of Aerospace Engineering, Indian Institute of Science, Bangalore, India.*

Received 3rd January 2014; accepted 26th May 2014

Abstract

Visual autonomous systems capable of monitoring crowded areas and alerting the authorities in occurrence of a suspicious action can play a vital role in controlling crime rate. Previous attempts have been made to monitor crime using posture recognition but nothing exclusive to investigating actions of people in large populated area has been cited. In order resolve this shortcoming, we propose an autonomous unmanned aerial vehicle (UAV) visual surveillance system that locates humans in image frames followed by pose estimation using weak constraints on position, appearance of body parts and image parsing. The estimated pose, represented as a pictorial structure, is flagged using the proposed Hough Orientation Calculator (HOC) on close resemblance with any pose in the suspicious action dataset. The robustness of the system is demonstrated on videos recorded using a UAV with no prior knowledge of background, lighting or location and scale of the human in the image. The system produces an accuracy of 71% and can also be applied on various other video sources such as CCTV camera.

Key Words: Unmanned Aerial Vehicle (UAV), Pose Estimation and classification, Pictorial structures, Image parsing, Human detection, Hough Transform.

1 Introduction

Meetings and huge gatherings often attract mischief from anti-social elements. The increasing rate of crime, vandalism and terrorism in densely populated areas has motivated security organizations to develop robust video surveillance systems. Aerial surveillance systems allow effective monitoring of populated areas due to their mobility. In contrast, mundane work is required for humans to manually identify criminal activities from live video. Besides, such a job is tedious and prone to human error, critical in these circumstances. With the increasing security threats, there is dire need of a system capable of automatically identifying suspicious action without any human monitoring.

A number of video surveillance systems have been developed in the past in an attempt to detect and report criminal activities [1], [2], [3]. Joyce *et al* [1] proposed a face detection system able to identify suspicious

Correspondence to: <p.surya@ieee.org>

Recommended for acceptance by <Angel Sappa>
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

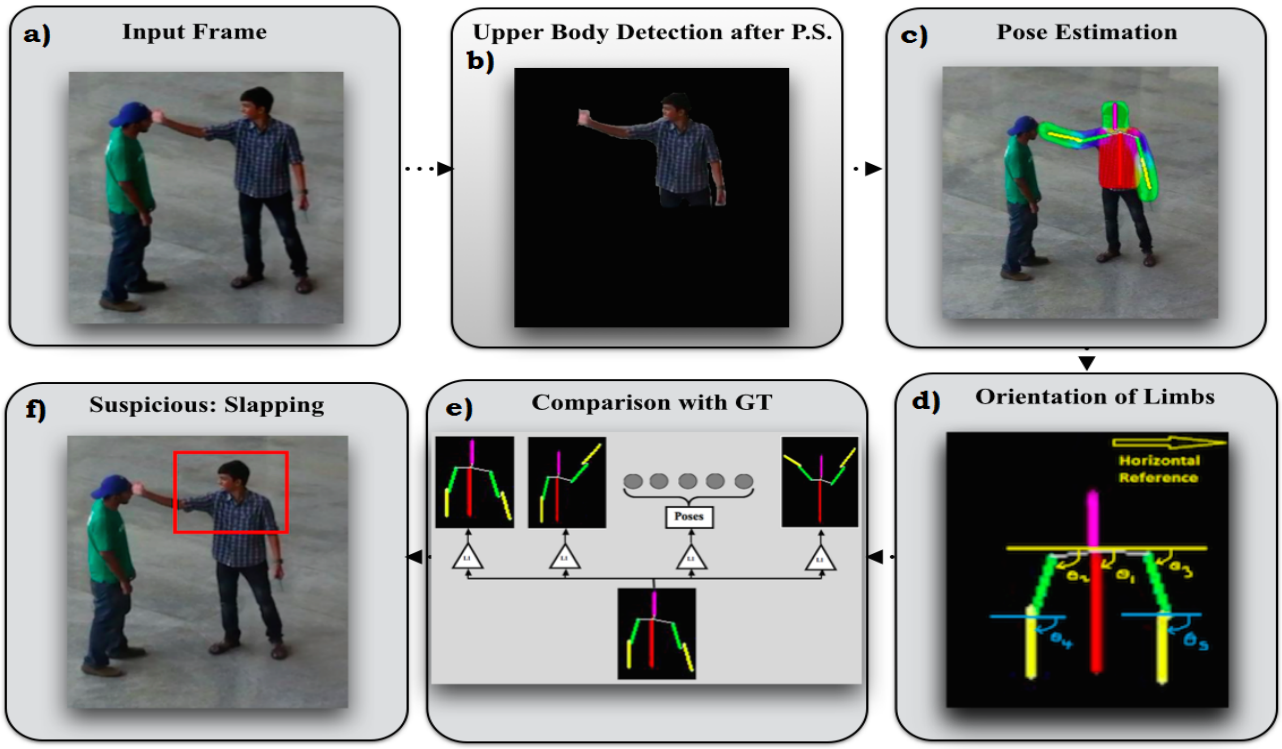

Figure 1: Block Diagram of the Proposed System. (a) Input image (b) Upper body detection after progressive search space reduction (c) Pose estimation represented as a CRF pictorial structure (d) Orientation of pose limbs to be used by the pose classifier (e) Comparison of pose with suspicious action dataset using the orientation of limbs (f) Suspicious action flagged with a red bounding box.

people by matching the face, and other personal details such as name, medical history, tax records, criminal arrest records, voting records, etc., with a criminal database. Bocchetti *et al* [2] proposed an algorithm that detects suspicious activities from video streams. Goya *et al* [3] introduced a Public Safety System (PSS) for identifying criminal actions such as purse snatching, child kidnapping, and fighting. The proposed system used three main characteristics (distance, velocity and area) to determine the human behaviour. Despite the efforts made to develop video surveillance systems, no attempts has been made yet to detect suspicious actions using aerial imagery. We propose a system that is able to detect suspicious actions using an unmanned aerial vehicle. To the best of our knowledge, this is the first surveillance system that uses still images to identify suspicious actions. The use of still images becomes extremely useful in case of aerial videos as the suspicious action is recorded only for a small period of time.

Our goal is to automatically detect and estimate the 2D pose of humans in images recorded under uncontrolled environments. Our work emphasis on detecting the upper body (head, arms and torso) of humans, as it contains enough information to identify the suspicious actions performed by a person. Moreover, an unmanned aerial vehicle usually only captures the upper body of the humans in a frame as the legs may be occluded by nearby humans. The proposed system may be used with general aerial still images as it requires no prior information of the background, lighting, or size of the humans in the image. The only assumption is that the humans in the aerial frame are standing upright and in a near frontal view. Most of the humans stand upright during large gathering, hence this is a weak assumption. The primary stage of the system is to detect and estimate the appearance of the person that is further used to locate his body parts (head, arms, torso). Once, we have part locations, we can narrow down the search domain to estimate the pose and further model it as a conditional random field (CRF) pictorial structure. The human pose is compared with the poses in the suspicious action dataset and it is flagged with the action which matches the best. The system can identify multiple suspicious

actions ranging from slapping, punching, hitting, shooting, chain snatching and choking. A block diagram of the proposed system is shown in Fig. 1. Before understanding the proposed system, it is essential to briefly understand the present state of the art algorithms involved in action recognition.

The paper is divided into the following sections. Section 2 summarizes relevant work on image parsing and pose estimation. Section 3 presents the main contributions of our work while Section 4 elaborates the pictorial model used for pose estimation. Section 5 describes the steps involved in model fitting further used for pose estimation and Section 6 compares the estimated pose with the suspicious action dataset. Section 7 presents the experimental results. Finally, the conclusion of our work is presented in Section 8.

2 Background and Related Works

State of the art methods for automatic identification of suspicious actions essentially involves (a) human detection and (b) human pose estimation. A number of attempts [4], [5] have been made towards 2D human pose estimation in still images and videos. Human pose estimation algorithms for both bottom-up [6], [7], [8] and top-down approaches [9] [10] have been proposed. Researchers have also retrieved the spatial configuration of humans by matching the holistic human shape [11], [12], aggregating poses from segmentations [8], or using contours to model the shape of human body parts [13], [14]. In addition, inference based methods have been introduced depending upon the model structure. Tree structured graphs have been used for exact inference [9], whereas loopy belief propagation [15] and linear programming [16] algorithms have been applied for approximate inference. Ferrari *et al* [19] introduced a novel technique, which uses GrabCut algorithm to estimate the pose of a near-frontal or near-back human image.

The methodologies mentioned above are not able to process single images as they require a video. Ramanan *et al* [17] proposed the pictorial structures framework. A person-specified appearance model can be used in still images. This method doesn't require any prior knowledge of the background for image parsing and pose estimation. This algorithm is human-centric as it detects and presents spatial configuration of the body parts for humans in the image frame.

The algorithm used in this paper is build on the work of Ramanan *et al*. [17] for human parsing. The original applications of the pictorial methods focused on naked human parsing [5]. Multiple appearance models have been proposed to estimate human poses in these environments [9]. The strike-a-pose algorithm [18] inspects all frames for a predefined indicative pose in which all parts are visible without overlap, allowing the learning of good appearance models, further used for pose estimation in all other frames.

Most surveillance systems use ground cameras making them incapable of monitoring large crowds due to the limited field of view. Monitoring the data received from multiple cameras requires large human effort and attention. An economical and convenient alternative to the problem is to use UAV for surveillance systems capable of analyzing large crowds which have not been cited in the literature so far. Such a system can be used to trigger warnings, sent to the security officials as soon as any violence is detected. Despite the wide field of view for videos recorded from UAV, one may face complications analyzing the videos. Video recorded using UAV are much more complex when compared to ground video as (i) the size of the people is smaller; (ii) complex background; (iii) there are multiple people in the same frame; (iv) tilt of the camera. Hence all these discrepancies have to be kept in mind before designing a UAV surveillance system.

3 Our Contributions

The main contribution of this paper is the coalesce and extension of available work- to perform surveillance using a video captured from UAV. Given that, there is no previous attempt to classify human suspicious action using aerial imagery. Our initial contributions include a combination of face detector with upper body detector [20] to improve the efficiency of human detection. Next, a cascade filtering mechanism is used to speed-up the face detection by discarding the non-human subwindows. Three prefiltering stages were employed to dis-

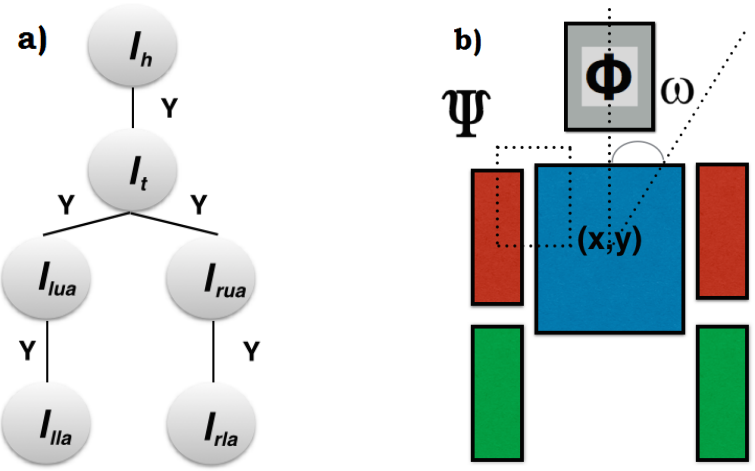

Figure 2: The figure shows the pictorial structure model where each node represents a body part (h: head, t: torso, left/right upper/lower arms lua, rua, lla, rla). (a) The kinematic tree includes edges between every two body parts, physically connected in the human body. (b) Cardboard representation where body parts are rectangular patches, parametrized by location (x, y) and orientation ω, with uninary potential of Φ, connected by kinematic priors Ψ.

card extra features, (i) area of the bounding box: extra-large and extra small boxes are eliminated, (ii) standard deviation: low standard deviation value depicts [20] background/leaves/shadow, and (iii) overlaping boxes: boxes that had more than 90% of the area in common. As the next contribution of this paper, Hough orientation calculator is proposed for pose classification. This technique uses the linear Hough transform technique to extract the exact orientation of each limb from the CRF image. These orientations are used as features for comparison with the suspicious ground truth action dataset.

4 Pictorial Structure Model

In this section, we describe the pictorial structure framework for representing the body parts of humans using a tree structured CRF as shown in Fig. 2(a). A human body part l_i is represented as rectangular patches with location (x, y), scale s and orientation Θ. The posterior configuration of the parts represented by $L = li$ is given by:

$$P(L|I, \Theta) \propto \exp\left(\sum_{(i,j)\epsilon E} \Psi(l_i, l_j) + \sum_i \Phi(l_i, \Theta)\right) \tag{1}$$

where $\Psi(l_i, l_j)$ denotes the pairwise potential representing the spatial prior on the relative position of the parts, taking the kinematic constraints into consideration. $\Phi(I|l_i, \Theta)$ denotes the unary potential of the local image evidence for a body part. The unary potential depends on the appearance model Θ describing the body part. The model computes the dissimilarity between the appearance model part i and the image patch at l_i.

The image patches are analysed using Ramanan *et al* [17] work. The body parts l_i are oriented patches of fixed size, with position parameterized by location (x, y) and orientation ω. The body parts are grouped together into a tree structure E as shown in Fig. 2(b), with edges $\Psi(l_i, l_j)$, defined by:

$$\Psi(l_i, l_j) = \alpha_i^T bin(l_i - l_j) \tag{2}$$

where $bin(.)$ represent a vectorized count of the spatial and angular histogram bins. $l_i - l_j$ represents the relative distance between part l_i and l_j. Selected spatial and angular bins are favoured by the model parameter

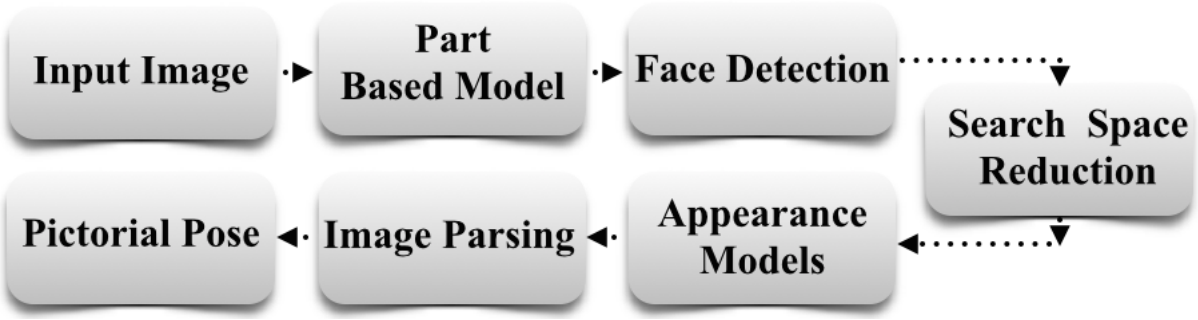

Figure 3: Processing pipeline for calculating pictorial pose

α_i for body part l_i with respect to parent of the body part. The model parameter α_i enable us to capture more complex distributions than gaussian priors. Edges are used as features to extract the soft estimates of the body parts. The soft estimates are computed by using an updated value of Φ including both the image edges and appearance of the body parts. The first stage of this machinery aims to capture a good appearance model. As the appearance models depend upon the edge based models, this stage typically fails in case of cluttered backgrounds resulting into bad appearance models and eventually incorrect pose classification.

The posterior configuration in Equation. 3 is restricted to detect only near frontal and near rear view humans. $\Upsilon(l_{head})$ and $\Upsilon(l_{torso})$ priors are added to PS model enforcing the orientations of head and torso to be near vertical. The extended model is defined below:

$$P(L|I,\Theta) \propto \exp\left(\sum_{(i,j)\epsilon E} \Psi(l_i,l_j) + \sum_i \Phi(l_i,\Theta) + \Upsilon(l_{head}) + \Upsilon(l_{torso})\right) \tag{3}$$

where $\Upsilon(l_.)$ has a probability greater than zero for only few vertical orientations resulting into a reduced search space for both head and torso. This step also improves the efficiency of pose estimation for arms due to the kinematic constrain of torso on the position of arms Ψ.

5 Pose estimation using Pictorial Structure Model Fitting

Pose is estimated by training the Pictorial Structure Model on aerial images. In the first step, humans are detected using the sliding window upper body detector on the input frame. In order to improve the probability of successful pose estimation, progressive search space reduction is used to reduce the search space. The location and scale of the human is used to estimate the person specific appearance model. Finally, person-specific appearance models and generic appearance models (edges) are used to run inference to estimate an articulated pose. The processing pipeline has been presented in Fig. 3.

5.1 Upper Body Detection

We start by detecting humans in each aerial image frame using the sliding window part based detector proposed by Felzenszwalb *et al.* [20]. The proposed technique presents a star structured part model to detect human upper body in still images. The detector uses a HOG [25] feature pyramid that captures coarse and finer gradients. It using multiple filters (root filter and a set of part filters) in conjunction with deformable models to compute the score for a particular body part from image frames. Once humans are detected, we attempt to remove false detections by applying Eigen face detector [26] in the proximity of the face. The algorithm is explained in detail below.

The input image frame is scanned by a sliding window scan fashion to detect humans. The window is divided into sub-windows further used to estimate the score for each limb of the body. The root filter is used to estimate

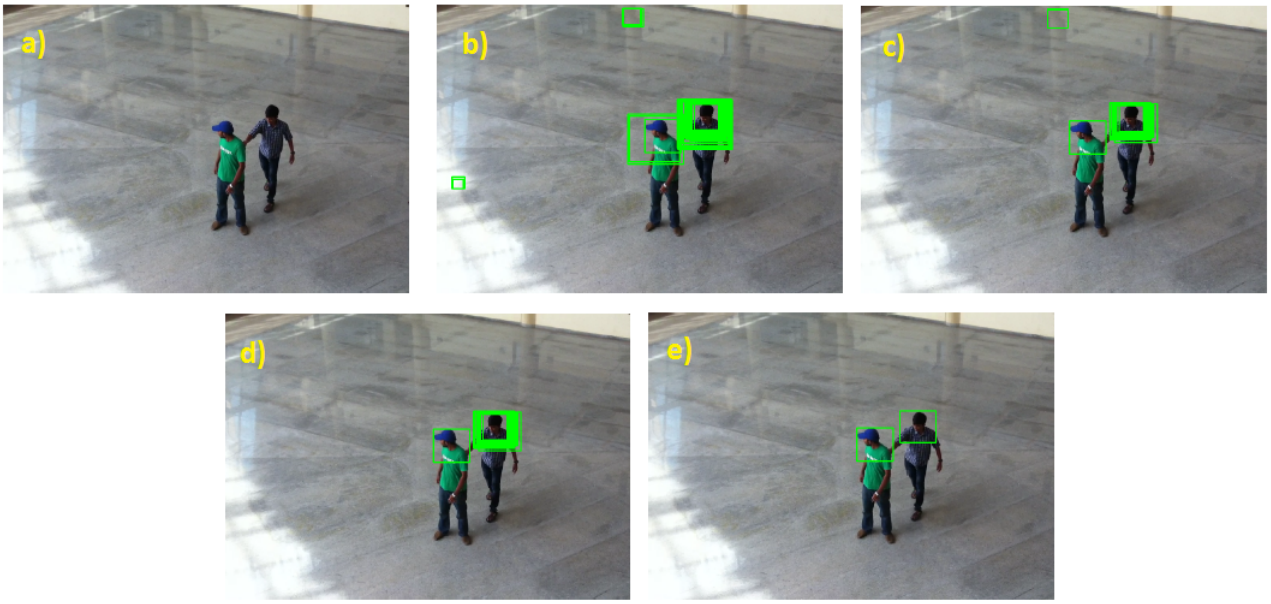

Figure 4: Demonstration of the upper body detection on an image taken from the UAV (a) Original Frame (b) Result after applying the standard algorithm explained above. The detected boxes are highlighted in green (c) Result after applying size filtering (d) Result after limiting the value of standard deviation (e) Result after removing the overlapping bounding boxes using non-max suppression.

the torso and the part filters are responsible for detecting the limbs. The score for a filter is given by the dot product of the filter and its respective sub window. Error can occur in accurately measuring the score for limbs due to the variation between sub-window and the filter, termed as the deformation cost. The total score of the model is calculated by adding the score of root filter all part, placements of part in the window and subtracting the deformation cost. The model is trained using latent SVM.

The score of each sample x (to be classified) can be calculated by the following function:

$$f_{\beta}(x) = \max_{z \epsilon Z(x)} \beta \cdot \Phi(x, z) \tag{4}$$

where, β represents a vector containing the parameters of the model, z stores the latent values, and $\Phi(x, z)$ is a feature vector. For a specific model, β is the linking between root filters, part filters, and deformation cost weights, z specifies the object configuration, and $\Phi(x, z)$ concatenates the features extracted from sub-windows to form a feature pyramid. The score of a mixture model at a particular position and scale is the maximum score of that component model. Detections carry information about the rough position and scale of people in the image.

This algorithm when applied to complex environments such as videos of a crowd recorded from a UAV returns some erroneous bounding boxes. These detections may comprise of some parts in the background mistaken as humans or may cover two or more people within the same bounding box. Fig. 4 represents one of such examples. Fig. 4(b) shows the result when the above algorithm is applied to Fig. 4(a). To get rid of this imprecision, additional steps of filtering have been introduced, to improve the standard algorithm.

5.1.1 Improved Algorithm

Face detectors applied in the vicinity of the location obtained by method proposed by Felzenszwalb *et al.* [20] successfully detects human in most of the cases but often misclassify patches as humans in videos acquired from a far distance as in case of a UAV. In addition, the algorithm produces unsatisfactory results if the color of

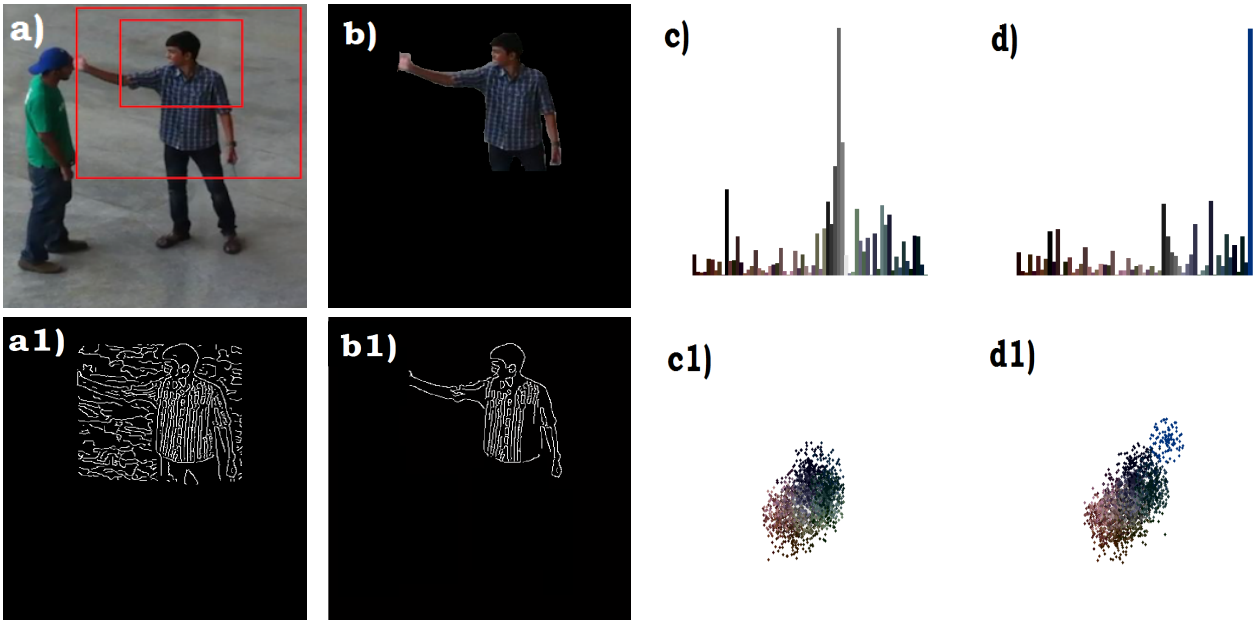

Figure 5: The figure presents the steps involved in model fitting. a) Upper body detection: The detected person (inner rectangle) and enlarged window where further processing is applied (outer rectangle). a1) The edge model for the enlarged window is presented in the same column b) The foreground region output by Grabcut; (b1) The edges corresponding to the foreground region;c) Estimating appearance models: Initial color histogram obtained from SP; (c1) Color cloud corresponding to color histogram presented in (c); (d) final Color histogram after appearance transfer stage; (d1) Color cloud corresponding to color histogram presented in (d).

human clothing is identical to the background color. Due to these limitations only a few percentage of people were recognized from the crowd.

In order to improve the accuracy of the human detection from aerial videos, the classification threshold used to decide if the window is classified as positive (in Eq. 4), is reduced to detect missing human subjects. This increases the sensitivity of the algorithm allowing it to detect all human subjects in the crowd by compromising the precision of bounding box detection. Relaxing the threshold value makes the algorithm more sensitive to detect human subjects but increases the false positive rate. In order to remove false positives, a three step cascading mechanism is introduced below.

A size box filter eliminates false boxes using the size of the bounding boxes. Depending on the height of the UAV, an approximation of the distance to crowd from the UAV is computed. Using this distance, a dynamic upper and lower limit for the area of the detected bounding box is set. All the sub-windows having area lower than the low limit or greater than the high limit are rejected. In other words, the extra-large and extra-small boxes without the possibility of human are eliminated as shown in Fig. 4(c).

A pixel standard deviation box filter eliminates false boxes using the pixel deviation within the box. The bounding boxes enclosing only the background are eliminated as the background has constant intensity and colour. The bounding box containing humans have higher standard deviation which helps the algorithm inselection as shown in Fig. 4(d).

Non-max suppression is used to eliminate overlapping bounding boxes where a single person is detected several times and marked by bounding boxes of nearly the same size and position. These boxes are replaced by an average of all the enclosing bounding boxes as shown in Fig. 4(e). The upper body detected for the person of interest is shown in Fig. 5(a) while the edge model for the detection window is shown in Fig. 5(a1).

5.2 Foreground Highlighting

The progressive search space reduction technique, proposed by Ferrari *et al.* [19] is used to reduce the search space for body parts (head, torso or limbs). From upper body detection mention in Section 5.1, a weak representation of the human upper body is obtained in terms of its location and scale. This information is used to initialize GrabCut [24] that reduces the search space by removing part of the background clutter. This stage further constrains the search space by limiting the locations to lie within the area output by Grabcut. Fig. 5(b) shows the result after the progressive search space reduction algorithm while the edge model after space reduction is shown in Fig. 5(b1).

5.3 Estimating Appearance Models

Once an approximate location (x, y) of the person present in the frame is obtained, appearance models are estimated and parsed to obtain posterior marginal distribution for person's body parts. This approach is based on two primary observations: (i) the location of torso and head is stable relative to the detection window $W = (x; y; s)$, where (x, y) is the location and s represents scale of the person, while others parts can be located almost anywhere in the window (e.g. lower arms); (ii) the appearances of the upper arms often have the same color as the torso and have close relation to each other. The conditions are used to develop appearance models for body parts that are located w.r.t W (e.g. torso) and are further used to determine the appearance model for more deformable parts (e.g. lower arms). The appearance models are obtained by (i) learning a segmentation prior (SP) that acquire the distribution of the body part locations w.r.t to W; (ii) improving the models derived from the segmentation prior by combining models for different body parts using an appearance transfer mechanism. The initial models learnt are estimated given W and the learnt segmentation priors. The models learnt are further refined using the appearance transfer mechanism. After learning, this method is ready to estimate appearance models on new, unannotated test images.

5.3.1 Training: learning segmentation priors

Segmentation Prior $SP_i(x, y)\ \epsilon\ [0, 1]$ for a body part is defined as the prior probability for a pixel (x, y) to be covered by that body part, before actually considering the image data. This is helpful in estimating the initial appearance models before running a pictorial structure inference. Here, the appearance models are color histograms $P_i(c|fg)$ obtained by weighting pixel contributions by the SPs $SP_i(x, y)$. In this common coordinate frame, the SPs are learnt in maximum likelihood fashion for every pixel in the enlarged detection area over a fraction of training images.

5.3.2 Training: transferring appearance models between body parts

In this section, the system learns a transfer mechanism that estimates the new appearance model of a part as the linear combination of the input appearance models of all parts. The parameters of the appearance transfer mechanism are the mixing weights of each body part. The new appearance model Af_i^{TM} for body part t is given by

$$Af_t^{TM} = \sum_i w_{if} Af_i^{SP} \tag{5}$$

where w_{if} is the mixing weight of part i, in the combination for part t, and Af^{SP} is the initial appearance model captured from SP. The weights are learnt by minimizing the squared difference between the appearance models produced by the transfer mechanism and those derived from the ground truth annotation. Upper arm models are improved by appearance transfer from the torso. Lower arms with highest localization uncertainty get strong contribution from all other parts.

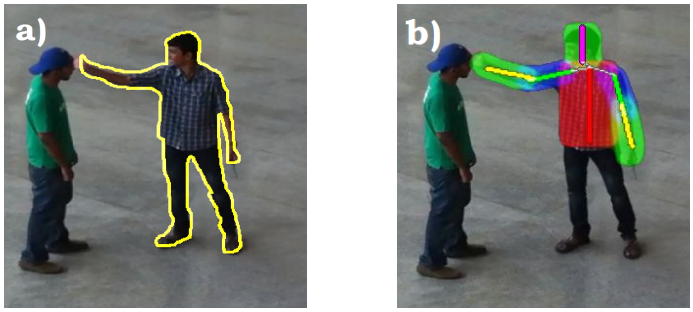

Figure 6: The figure shows (a) The person of interest in the aerial image frame (b) Pose estimated as pictorial structure for the person of interest.

5.3.3 Test: estimating appearance models for a new image

After learning the SPs and the mixing weights, color models are obtained by weighting color contributions according to SP values and further refining them by applying appearance transfer leading to final color models. The color models estimated characterize the appearance of the body parts. The foreground and background models are used to calculate the posterior probability for a pixel belonging to a part. The posterior foreground probabilities are then used to derive a color soft-segmentation of the image for each body part used as cue in the unary term of the pictorial structure. The color histogram and cloud for the initial appearance model and the final refined appearance model after inference are shown in Fig. 5(c), Fig. 5(c1) and Fig. 5(d), Fig. 5(d1) respectively.

5.4 Image Parsing

Finally, person-specific appearance models and generic appearance models (edges) are used to run inference to estimate an articulated pose. The output of the parsing stage is the posterior marginal distribution $Pi(x, y, \omega)$ the body parts of each detected person. The final result of image parsing algorithm is the CRF pictorial structure representation as shown in Fig. 6.

6 Pose Classification

The estimated pose is labelled as suspicious depending upon the similarity with a suspicious activity dataset. It is estimated using proposed Hough Orientation Calculator. The pose classification steps are explained in detail below.

6.1 Hough Orientation Calculator

The CRF representation of the pose estimated from above section is further used to classify the pose as suspicious or not. With the location of various joints and orientation of all limbs, each pose can be classified. Exploiting the fact that the head, torso and each limb are represented by straight lines in the pose representation, linear hough transform is applied on them. This transforms the image into a hough space graph with the straight lines represented by sharp peaks while angles at peaks represent the orientations of the limbs. The intersection of the lines in pose representation help us estimate the locations of the joints. The orientation of each body part, namely head, torso, arms and forearms, is the angle formed with a horizontal reference, measured in the anticlockwise direction in the range of -180 to 180 degrees as shown in Fig. 7(a). Fig. 7(a) represents the direction in which the angles are measured, from the horizontal. The five angles stored are the orientations of back, left upper arm, right upper arm, left lower arm and right lower arm. Fig. 7(b) shows a soft threshold representation of a pose. Fig. 7(c), Fig. 7(d), Fig. 7(e) shows the backbone, the upper arms and lower arms,

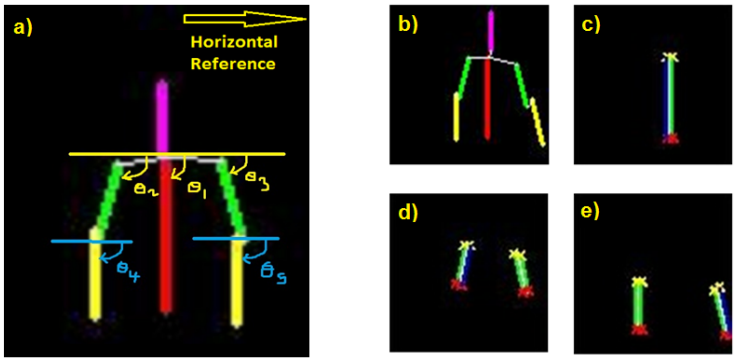

Figure 7: Different body parts extracted from the CRF image (a) References and Angles (b) A random CRF figure (c) Body separated and hough transform applied (d) Upper limbs separated and hough transform applied (e) Lower limbs separated and hough transform applied.

respectively, extracted from Fig. 7(b), when this technique is applied. Then, the angles obtained are computed with respect to spinal cord of the soft label representation. The tilt of the UAV can induce an unwanted bias in the angle of the body parts if computed with respect to some reference point in the frame.

6.2 Comparison with Ground Truth

The angles stored in the previous step are compared with the ones stored in the ground truth to estimate similarity between the two. The similarity for a pose with the ground truth is estimated by computing an L_1 norm between the orientation of the limbs of the input test pose and each pose in the ground truth database. By setting a threshold to the value of the obtained similarity, it can be decided if the two compared postures can be considered similar or not. The similarity values can be calculated by:

$$\tau = \sum_i |\theta_i - \phi_i| \tag{6}$$

where τ is the similarity, θ_i represents the orientation of the i^{th} limb of the test image, normalized with respect to the backbone, and ϕ_i stands for the corresponding angles in the ground truth.

The proposed system classifies the person performing a suspicious action as suspicious if the the orientations

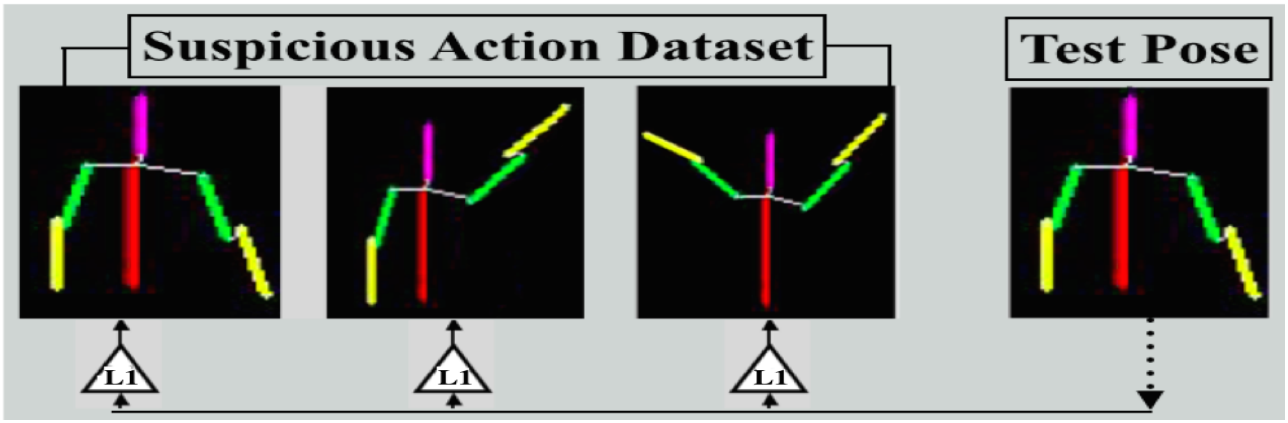

Figure 8: Comparison of test pose with all the poses in ground truth suspicious action dataset.

Figure 9: Results of the algorithm on ad frame recorded from the UAV (a) Result after removing overlaped bounding boxes (b) CRF figure representation of each person (c) Result marking the suspicious person (in red).

of the suspicious action performed matches with the orientations of any suspicious action in the ground truth as shown in Fig. 8. The ground truth consists of a variety of commonly encountered poses, that includes slapping, punching, shooting, choking and chain snatching. Hence, the system not only detects the person as suspicious, but also gives information on which suspicious action is being performed.

7 Results

The proposed algorithm is tested on three different videos recorded in various environments, including different suspicious actions. We applied the proposed algorithm to 170 images containing 430 poses for video-1, 310 images with 873 poses for video 2 and 152 images with 367 poses for video-3 respectively. The images depicted in the paper have been captured using a camera mounted on an UAV. The camera was fixed at 25 degree with respect to the ground in order to capture maximum number of people in the frame. The camera is capable of recording high definition videos at a frame rate of 60fps. The sequence recorded was about two minutes long summing to 7200 frames per sequence. Many frames in the sequence are redundant. Hence, instead of applying the algorithm to each frame in the sequence, frames at a temporal distance of one fourth of a second were considered. In the recorded sequence, 5-10 people were captured in one frame randomly performing the suspicious actions. Each of their actions are parsed and compared to the poses in the suspicious action dataset. An example of our qualitative results is shown in Fig. 9. A single set of parameters are used for all the images. Qualitative and quantitative evaluations, as well as performance aspects in terms of implementation and portability, are presented next.

7.1 System Parameter Settings

The proposed system has a few parameters that are set according to the characteristics of the input images. In this section, we explain the reasons and the values chosen for such parameters. This might help to port our system for input images from different sources. We consider these parameters to be the most important ones to adjust for a different dataset.

a) Eigen Face Detector: As explained earlier, Eigen face detector is used to detect the face and further the window surrounding the human in the aerial frame. We use 20 signatures with 10 largest Eigen vectors for each image frame.

b) Edge Model: We used canny edge operator in order to generate the edge model for the selected window for an aerial image. For all the images presented in this work, only one set of parameters are used for the Canny operator. These parameters include a lower threshold of 0.12 and a higher threshold of 0.3 that were found empirically.

c) Pictorial Structure Inference: The posterior marginals in a tree model are computed using belief propagation (BP) for 6 body parts and the number of states h as $|x|$,$|y|$=150, θ =24.

d) Appearance transfer weights: Table 1 shows the mixing weights learnt based on the segmentation prior. According to the learnt model, the appearance of refined stationary parts (torso and head) are identical to the

Figure 10: Example of qualitative results. Each row represents a specific suspicious action. The first row represents (a) 9^{th} person from the left choking. He is identified as suspicious and marked in red. (b) 9^{th} person from the left choking and 8th person from the left is shooting. Both of them are identified as suspicious and marked in red. (c) 8^{th} person from the left shooting. He is identified as suspicious and marked in red. (d) 9^{th} person from the left choking and 8th person from the left is shooting. Both of them are identified as suspicious and marked in red.

input model from segmentation prior. Upper and lower arms models use the appearance transfer from the torso. The results prove that exploiting relations between the appearance of different body parts leads to better appearance models.

7.2 Qualitative Evaluation

Fig. 10 represents typical results for suspicious action detection at different aerial frames. In this figure, the first column, represents the human upper body detected for actual aerial frame captured by the UAV. The second column, shows the pictorial structure representation of the pose for each person. Finally, the person performing the suspicious action is highlighted with a red bounding box.

7.3 Quantitative Evaluation

The accuracy of the recognition is measured on the dataset of 632 images and 1670 poses using the following three measures:

$$accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{7}$$

$$precision = \frac{t_p}{t_p + f_p} \tag{8}$$

$$recall = \frac{t_p}{t_p + f_n} \tag{9}$$

Where t_p represents true positive, f_p stands for false positive, t_n denotes true negative and f_n is false negative. A true positive represents suspicious action classified as suspicious by the classifier; a false negative represents the classification of suspicious action as non-suspicious; a false positive corresponds to the classification of non-suspicious action as suspicious and a true negative stands for non-suspicious action classified correctly. The accuracy, precision and recall on these videos have been depicted in Table. 2.

Table 1: Learned appearance transfer weights.

Body Parts	Torso	Upper Arms	Lower Arms	Head
Torso	1	0.13	0.15	0
Upper Arms	0	0.87	0.29	0
Lower Arms	0	0	0.37	0
Head	0	0	0.19	1

Table 2: Accuracy (in %) for various Suspicious Poses

Action	Precision	Recall	Accuracy
Slapping	95.65	62.86	77.78
Punching	87.10	58.70	76.67
Shooting	81.82	46.15	79.59
Choking	88.46	63.89	73.47
Snatching	91.97	61.11	78.26

The system finds it difficult to detect the suspicious action as the number of people in the image frame increases. The increase in number of people leads to mis-detections. The suspicious action recognition results with respect to the number of people (i.e. represented with frequency) in an image frame are shown in Fig. 11 (a). The decreasing accuracy represented using the trend line substantiates our claim.

7.4 Performance

The implemented proposed system is a mix of C++ and Matlab code. The simulations are performed on an Intel Core2 Quad 2.26 GHz with 2 GB RAM CPU. The images tested are of 1280 by 720 in pixels. The suspicious action detection process takes an average time of time of 50.9-98.4 sec for 5-10 persons respectively in the image frame. The average time complexity for individual steps is as follows: (1) Human detection: 3.4 sec; (2) foreground highlighting 2.8 sec.; (3) estimating appearance models: 0.9 sec.; (4) parsing: evaluating unary terms 1.9 sec., inference 1.3 sec. (5) Orientation calculation: 0.3, comparison with GT: 0.5 (6) overhead of loading models including image resizing and other operations: 1.8 sec. The average total time for an image is 3.4+9.5P sec., with P the number of persons. The average time complexity across number of people in the image frame is presented in Fig. 11 (b). The increasing average time complexity increases as the number of people in the image frame increase as the represented using the trend line.

8 Conclusion

A novel video surveillance system has been proposed to monitor suspicious activities performed by people in crowded areas. The pose of each person is extracted and matched to the poses in the suspicious action dataset. If a match is found, the person is marked in red. The system can be used during events conducted in

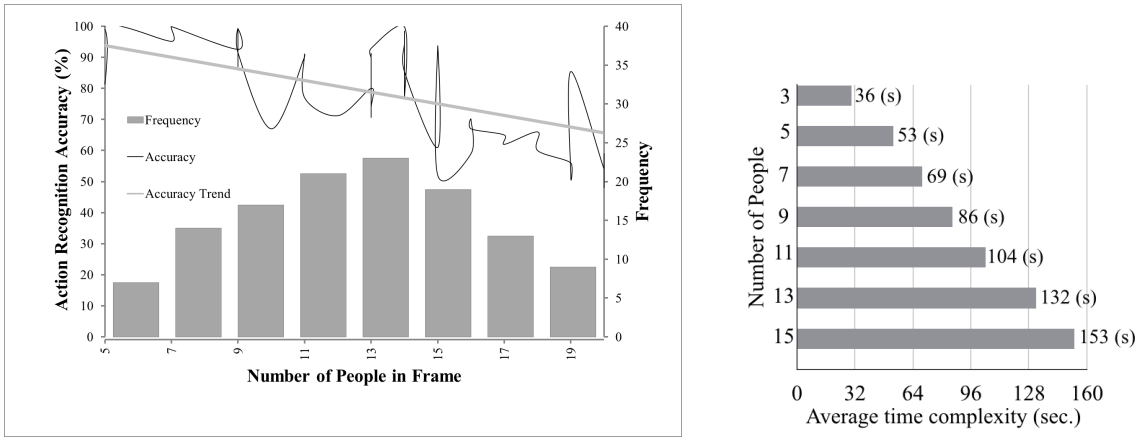

Figure 11: The graphs shows (a) Action Recognition accuracy and (b) Average time complexity.

huge grounds where the conventional ground camera is not enough to capture the whole crowd. The system was tested on three video recorded from a quadcopter that captured scenes where a group of volunteers were exhibiting suspicious actions. The system has been tested on different real scenarios offering promising results. The future work of this paper will include the detection of suspicious objects such as unmoved bags to make the system more effective. Reducing the computational time is also an important factor that can be improved.

References

[1] W.L Joyce, “Identifying Terrorists: Privacy Rights in the United States and the United Kingdom”, Hastings International and Comparative Law Review, 25, 2001.

[2] G. Bocchetti, F. Flammini, A. Pappalardo, “Dependable integrated surveillance systems for the physical security of metro railways”, ACM/IEEE International Conference on Distributed Smart Cameras, 1:1-7, 2009.

[3] K. Goya, X. Zhang, K. Kitayama, and I. Nagayama, “A method for automatic detection of crimes for public security by using motion analysis”, International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 1:736-741, 2009.

[4] D. Forsyth, M. Fleck, “Body plans”, IEEE Conference on Computer Vision and Pattern Recognition,, 1:678-683, 1997.

[5] S. Ioffe, D. Forsyth, “Finding people by sampling”, IEEE Conference on Computer Vision and Pattern Recognition, 2: 1092 - 1097, 1999.

[6] G. Hua, M.H. Yang, Y. Wu Y, “Learning to estimate human pose with data driven belief propagation”, IEEE Conference on Computer Vision and Pattern Recognition, 2: 747- 754, 2005.

[7] M.W. Lee, I. Cohen, “Proposal maps driven mcmc for estimating human body pose in static images”, IEEE Conference on Computer Vision and Pattern Recognition, 2: 334-341, 2004.

[8] X. Ren, A. Berg, J. Malik, “Recovering human body configurations using pairwise constraints between parts”, IEEE Conference on Computer Vision and Pattern Recognition, 1:824-831, 2005.

[9] P. Felzenszwalb, D. Huttenlocher, “Pictorial structures for object recognition”, International Journal of Computer Vision, 60(1):55-79, 2005.

[10] K. Mikolajczyk, C. Schmid, A. Zisserman, “Human detection based on a probabilistic assembly of robust part detectors”, ECCV: Lecture Notes in Computer Science, 3021: 69-82, 2004.

[11] D.M. Gavrilla, “Pedestrian detection from a moving vehicle”, ECCV: Lecture Notes in Computer Science, 2:3749, 2000.

[12] G. Mori, J. Malik, “Estimating human body configurations using shape context matching”, European Conference on Computer Vision (ECCV), 3: 666, 2002.

[13] M. Andriluka, S. Roth, B. Schiele, “Pictorial structures revisited: People detection and articulated pose estimation”, IEEE Conference on Computer Vision and Pattern Recognition, 1:1014 - 1021 , 2009.

[14] M.P. Kumar, P.H.S Torr, A. Zisserman, “Efficient discriminative learning of parts-based models”, IEEE International Conference on Computer Vision, 1:552-559, 2009.

[15] L. Sigal, M. Black, “Measure locally, reason globally: Occlusion-sensitive articulated pose estimation”, 2:2041-2048 , IEEE Conference on Computer Vision and Pattern Recognition, 2006.

[16] H. Jiang, “Human pose estimation using consistent max-covering”, IEEE International Conference on Computer Vision, 1:1357 - 1364 , 2009.

[17] D. Ramanan, “Learning to parse images of articulated bodies”, Neural Info. Proc. Systems (NIPS), 1:1129-1136, 2006.

[18] D. Ramanan, D.A. Forsyth, A. Zisserman, “Strike a pose: Tracking people by finding stylized poses”, IEEE Conference on Computer Vision and Pattern Recognition, 1:271278, 2005.

[19] V. Ferrari, M. Marin-Jimenez, A. Zisserman, “Progressive search space reduction for human pose estimation”, IEEE Conference on Computer Vision and Pattern Recognition, 1:1-8, 2008.

[20] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, “Object Detection with Discriminatively Trained Part-Based Models”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627,1645, 2010.

[21] Chia-Feng Juang, Chia-Ming Chang, “Human Body Posture Classification by a Neural Fuzzy Network and Home Care System Application”, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 37(6):984-994, 2007.

[22] S. Shahbudin, A. Hussain, Ahmed El-Shafie, N. M. Tahir, S. A. Samad, “Adaptive-Neuro Fuzzy Inference System for Human Posture Classification Using a Simplified Shock Graph”, Visual Informatics: Bridging Research and Practice, Lecture Notes in Computer Science, 5857:585-595, 2009.

[23] S. Andrews, I. Tsochantaridis, T. Hofmann, “Support Vector Machines for Multiple-Instance Learning,” Advances in Neural Information Processing Systems, 2002.

[24] C. Rother, V. Kolmogorov, A. Blake, “GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts”, ACM Transaction on Graphics, 23(3):309-314, 2004.

[25] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[26] Turk, M.A.; Pentland, A.P., “Face recognition using eigenfaces,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1991.