

Detecting and Tracking of Multiple People in Video based on Hybrid Detection and Human Anatomy Body Proportion

Amr El Maghraby
Computers &
Systems Eng.
Zagazig University

Mahmoud Abdalla
Communication Eng.
Zagazig University

Othman Enany
Computers &
Systems Eng
Zagazig University

Mohamed Y. El
Nahas
Computers &
Systems Eng.
Elazhar University

ABSTRACT

This paper addresses problems of detection and tracking of moving multiple people in a video stream. Detecting and tracking are fundamental tasks for future research into Human Computer Interaction (HCI). Detecting and Tracking multiple people in video are considered time consuming processes due to the amount of data a video contains, illumination changes, complex backgrounds and occlusions that occur as soon as people change orientations over time. This study focus on developing a fully automated system aims to Detecting and tracking multiple people in video, by analyzes sequential video frames based on hybrid detection algorithm, and tracking based on human body structure. The performance of the proposed system is tested through a series of experiments and human computer interaction application based human detection, tracking and identification. Identification is based on new clustering method mentioned in this paper.

General Terms

Video Processing, Computer vision systems, Human detection and tracking, Clustering.

Keywords

Video Processing, Human detection and tracking, Viola- Jones upper body, Skin detection, Computer vision systems, Biometrics

1. INTRODUCTION

Video has become an important element of multimedia computing and communication environments. As the years go by and people use technology more and more it's easy to make impressive videos in no time due to the rapid advancements in digital devices starting from capture, store, upload and download. Human detection and tracking are considered a critical technology for machine/human interaction and fundamental tasks of computer vision processes based on video analysis. Wide video domain related to Human Computer Interaction (HCI) such as computer graphics, artificial intelligence and computer vision. Some examples of applications with reliable human motion detection and tracking are: surveillance for security, biometrics, generating natural animation, human interaction for mobile robotics, pedestrian detection on vehicles and automatic motion capture for video. Human detection is the process techniques for locating human beings present in an image or video. Human tracking is the process techniques of analyzing video frames sequences that represent human movement in video and it requires a human detection mechanism in every frame. The algorithm applies primary detector on input video sequence based on Viola-Jones cascade object detector algorithm, to detect human upper body and return bounding-boxes that directly feeds into human

anatomy body proportion to locate the head and face positions of humans in individual video frames. The result of these steps returns a face position for human and false positive non human entity as detections from an upper body detector. To successfully classify a given face to be as human or non-human in nature, the proposed algorithm check skin color information of the face area as secondary skin detectors for enhancing the separation between skin and non skin pixel. Tracking process is done using repeating detection process for each progressive frame of the video which changes dynamically.

2. PREVIOUS WORK

The work in this thesis is a completion of the research that we have started and published two scientific paper. The first paper titled: Hybrid Face Detection System using Combination of Viola - Jones Method and Skin Detection [1]. The second paper titled: Detect and Analyze Image face parts information using Viola- Jones and Geometric approaches[2] . The first paper have a main objective to improves the performance of face detection systems in terms of increasing the face detection speed and decreasing false positive rate in still images with complex background . The algorithm based on three hybrid detector. The primary detector use Viola Jones upper body model for high probability of finding face in this region instead of searching the entire image. In order to find an accurate face in that region of interest, Viola-Jones face detector is used as a secondary detector to increase accuracy and reduces false negatives. Third detector pixel-based skin detection methods applied on the upper body region of interest which is not detecting a face using the secondary detector. The third detector classifies each pixel as skin or non-skin individually. Figure1 show a comparison between three face detection algorithms.




A) Viola – Jones face detection	B) Viola - Jones facial after skin color detection area	C) Upper body then viola Jones if not found face apply skin detection (proposed)
		
0 face found	0 face found	1 face found

Figure1: A comparison between three face detection algorithm

The summary of experimental results on 50 test images representing faces at different imaging conditions The accuracy is obtained by using the following equation on 100 :% Accuracy = 100 – (False positive Rate + False negative Rate)

Table 1: Comparison of Face Detection Accuracy for three methods

Criteria	Viola& Jones face detection	Viola& Jones face detection on skin region	Proposed Hybrid Face Detection System
False Positive Rate	3.738318	20.09345794	7.943925
False Negative Rate	20.09346	55.14018692	6.074766
Accuracy	76.16822	24.76635514	85.98131

The second paper focused on develop fully automated human face detection to locate eyes, mouth, nose and suratip in an image with complex backgrounds, and covers the detection tasks, landmark localization and measure facial part physical location by applying different techniques as in Figure2.

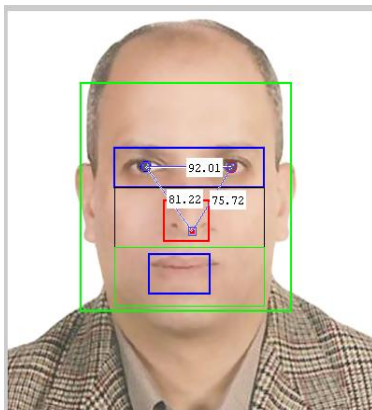


Figure2: detection, landmark localization and measure facial part eyes, mouth, nose and suratip in an image

3. PROPOSED ALGORITHM

The main objective of the study is to develop a novel unsupervised robust system which analyzes sequential individual video frames detecting and tracking multiple human in video. The system detect track and store the human movement from separate sequence of pictures and different between them by construct cluster table containing a point feature of human motion position and human centroid. The algorithm applies primary detector on input video sequence based on Viola-Jones cascade object detector algorithm to detect human upper body and return bounding-boxes that directly fed into human anatomy body proportion to locate the head and face position of human in video frames. The result of these steps may return a face position for Human and Non Human entity as false positive detections from upper body detector, the secondary detector check skin color information of the face area as skin detectors for enhancing the separation between skin and non skin pixel which could successfully classify a given face to be as human or nonhuman in nature. Tracking process is done using repeating detection process for each progressive frame of the video which changes the proposed algorithm consists of three main components.

3.1 Video Object Detection

Human detection has been studied in a number of multimedia applications such as face tracking, face recognition, and video surveillance. most detection algorithms reported so far have been human face detection. Face detection is the essential first step towards many advanced computer vision, biometrics recognition and multimedia applications. Paul Viola and Michael Jones presented a fast and robust method for face detection. The disadvantage of this method is to focus on detecting frontal faces with good lighting conditions. Other detectors are specifically trained at detecting the upper body as big object can be easily detected. Viola- Jones upper body model uses Haar features to encode the details of the head and shoulder region. Because it uses more features around the head [3], this model is more robust and detects the human upper body walking, standing and sitting for different illuminations and different position frontal or by side also the model is stable against pose changes, e.g. head rotations/tilts. The disadvantage of Viola Jones upper body detection is that the false positive rate is high. Each detection approach have their advantages and disadvantages. Most conventional approaches for object detection are background subtraction, optical flow and spatio-temporal filtering method; it is evident that the integration of multiple techniques can provide better results. This study uses a complete different technique. The proposed primary detector cascade object system uses the standard Viola Jones detector as first step to detect the upper-body region, which is defined as a head and shoulders areas. The following block diagram shows the Viola Jones upper body

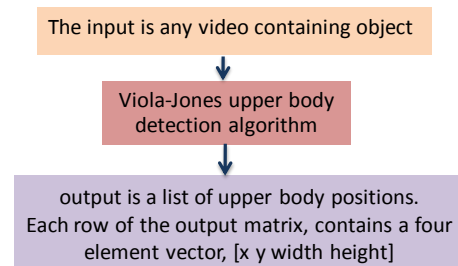


Figure3: block diagram show the Viola Jones upper body

Proposed primary detector system aims to making human distinctions among moving objects in a video sequence. The output of upper body detection process is a list of upper body positions. Each row of the output matrix contains a four element vector, [X Y Width Height] represented as green rectangle around the upper body human detection in video frame as in Figure 4.

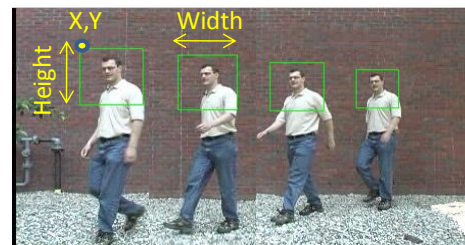


Figure4: video screen snap shoot represent primary detector upper body detector

3.2 Human Anatomy Body Proportion

Anatomy is the structure of the human body. It consists of muscles, skeleton, and proportions. Ancient Egyptian art used a canon of proportions based on the "fist"[4] measurement across the knuckles, with 18 fists from the ground to the hairline on the

forehead. This was already established by the Narmer Palette from about the 31st century BC, and remained in use until at least the conquest by Alexander the Great some 3,000 years later. An average person, is generally 7-and-a-half heads tall (including the head) [5].

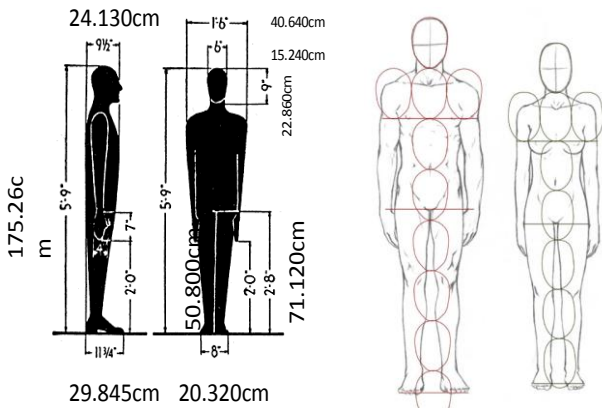


Figure 5: Human body proportion

The body width = 3 heads .The body height =7.5 heads .Distance between nipples on the chest = 1 . Head Bottom of the knees = 2 heads from ground level .The proposed model used the body proportion dimension as attractive reference to detect the human body by calculating the upper body detection by the primary detector which uses the same width and multiply the height * 3.5 to determine the whole body which is represented in Figure 6 by red rectangle .

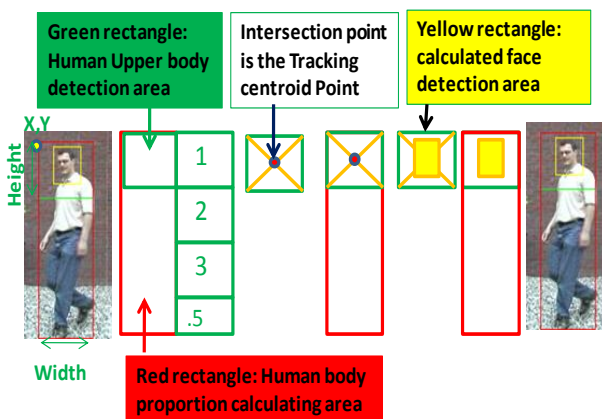
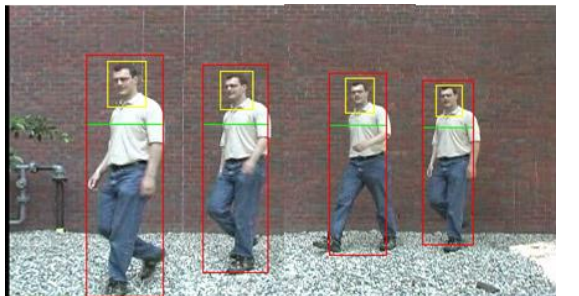


Figure 6: Human proportion calculating from upper body

Expected face width 1/3 body due to Human anatomy body proportion, Face width=(upper body width)/3).

We calculate the face position due to the following formula

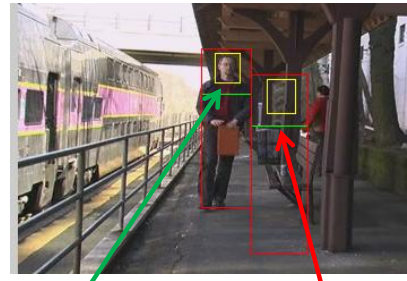
$X_position = X - (Face_width/2)$ where X is the position for the upper body

$$X_position = Y - (Face_width)$$

$$Width = Face_width$$

$$Height = Face_width * 1.2 ;$$

Experimental result considered this method as the fastest and most accurate to locate the human face area as region of interest



Upper body detection Upper body false positive

Figure 7: Primary Detection Result

3.3 Skin Color Threshold

Skin color is a distinguishing feature of human faces. Skin color is a popular parameter in the computer vision to detect humans. Skin detection in the proposed framework used to check the calculated located faces region in images from Human body proportion as explained. We use this feature as color processing technique to check limited number of pixels (which would reduce the detection rate) if we found one skin color pixel in this area. Human retina contains rod-shaped photoreceptors, which do not distinguish color, but are more sensitive to overall brightness. A color space is a mathematical representation of a set of colors example: RGB and HSV(used in computer graphics). YIQ, YUV, or YCbCr (used in video systems). CMYK (used in color printing). The proposed framework is based on color space transforming for the face calculate area from RGB to HSV (hue-saturation-value) and YCbCr Luma value (Y') represents the brightness in the region of interest Chroma values (Cb, Cr). The color space conversion is performed due to the following formula.

$$Y = 0.257R + 0.504G + 0.098B + 16$$

$$Cb = -0.148R - 0.291G + 0.439B + 128$$

$$Cr = 0.439R - 0.368G - 0.071B + 128$$

$$R' = R/255 \quad G' = G/255 \quad B' = B/255$$

$$Cmax = \max(R', G', B'), \quad Cmin = \min(R', G', B')$$

$$\Delta = Cmax - Cmin$$

Hue calculation:

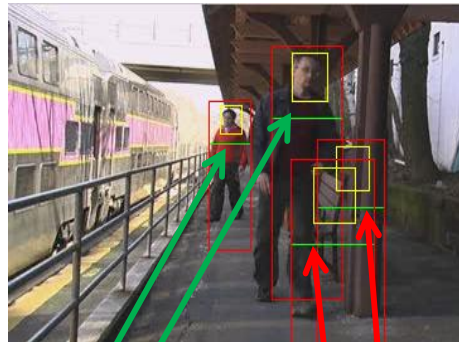
$$H = \begin{cases} 60^\circ \times \left(\frac{G'-B'}{\Delta} \text{mod} 6 \right), & Cmax = R' \\ 60^\circ \times \left(\frac{B'-R'}{\Delta} + 2 \right), & Cmax = G' \\ 60^\circ \times \left(\frac{R'-G'}{\Delta} + 4 \right), & Cmax = B' \end{cases}$$

Saturation calculation:

$$S = \begin{cases} 0, & Cmax = 0 \\ \frac{\Delta}{Cmax}, & Cmax \neq 0 \end{cases}$$

Value calculation: $V = Cmax$. Using matlab rgb2hsv converts an RGB colormap M to an HSV color map. Both colormaps are m-by-3 matrices. The elements of both color maps are in the range 0 to 1.

The proposed algorithm transforming every pixel from calculated and expected face region, converting RGB representation to chroma representation and determining the likelihood value based on the equation $140 < Cr < 165$ & $140 < Cb < 195$ a region of orange to red to pink in red-difference and blue-difference channels $0.01 < Hue < 0.1$ this means hue is basically reddish this define the skin pixel color



Upper body detection Upper body false positive

Figure 8: primary detector result and human body proportions appear the false positive result

Check skin threshold				
Skin	Yes	Yes	No	No
Action	Save	Save	Neglect	Neglect

Figure 9: Skin pixel threshold detection neglect false positive

The experiment show that we can only check the calculate face area and check for only one pixel which lie on the skin color region consider as a human body detect, Figure 10: summarize our proposed algorithm in block diagram. The proposed algorithm designed for real-time video so it is very light and smart able to save the whole body human detector and face in most common formats PSD, TIFF, JPEG, PNG and GIF as shown in Figure 11. The algorithm manipulating video frame by frame from a continuous stream processed one frame at a time.

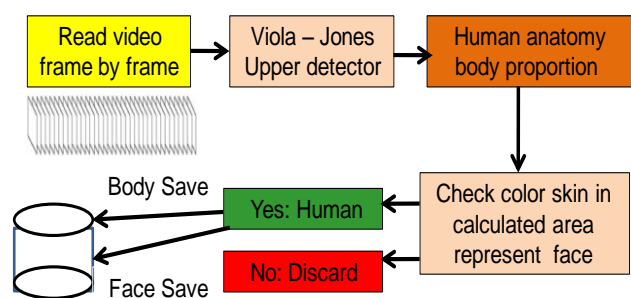


Figure 10: proposed algorithm block diagram

4. CLUSTERING

Clustering and classification are both fundamental tasks in learning and understanding data and have wide usage in a lot of fields, ranging from unsupervised learning neural network, Pattern recognitions, classification analysis, artificial intelligence, image processing, machine vision, and many others. Classification is used mostly as a supervised learning method, clustering for unsupervised learning[6].

Human body detect And track			
Mean	54.459886	88.423087	114.87625
Face detect			

Figure 11: Result of multiple person detecting and tracking

Since Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into clusters, so that the data in each subset share some common trait often according to some defined distance measure. the goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic. Clustering Traditional clustering methods were developed to analyze complete data sets. The k-means[7] clustering is an algorithm used to classify or to group cluster n objects based on attributes/features into K number of group, where K is positive integer known number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Weaknesses of K-Mean Clustering is, the number of cluster need to specify in advance before processing. This algorithm is not suitable in our case because k in this case represent number of human appear in the video scene, which is unknown because we don't know in advance how many people will appear in the video. The output of video processing is a set of characteristics and parameters related to each frame as shown in Table1. Where X and Y are the upper left pixel in the red rectangle around the human and width and height of red rectangle, frame # is the frame number which the human appear and mean value is done by computes the mean values of an image (human detect area). In this paper we discussed different approaches to determining the number of clusters (human) in a data set automatically and identify them by the mean value threshold. The proposed algorithm gives particular attention for the mean value of whole body area appears in each video frame. step1: starts from an initial value which is the first mean value in the list as reference value step2: Loop For each point, calculate the absolute difference between reference mean and the second mean row in the list replaces the reference value by the second mean value if it lies $<$ threshold then it is a member of cluster compute the average of the cluster. Repeat loop (until reach the end of list) step3: Loop take the average as reference and calculate the absolute difference for all the list which is not identified if it $<$ threshold then it is a member of the cluster Repeat loop (until reach the end of list) step4: get the minimum and maximum value of the cluster and the list mean value which is not recognize step5: loop For each point not recognized if the mean value of the item lies in this range it is a member of the cluster step 6: Repeat step 1 again by the first mean value in the list as reference value which is not recognized.

Table 1: sample result parameter for human detection and tracking

X	Y	Width	Height	Frame #	mean value	Id
203	30	57	51	66	51.46127075	1
208	32	53	48	67	52.16221053	1
207	32	53	49	68	51.16424035	1
165	72	27	24	68	85.91414566	2
207	31	57	52	69	51.97286603	1
163	72	27	25	69	88.42308721	2
210	30	56	51	70	52.36578947	1
160	69	33	30	70	93.79587495	2

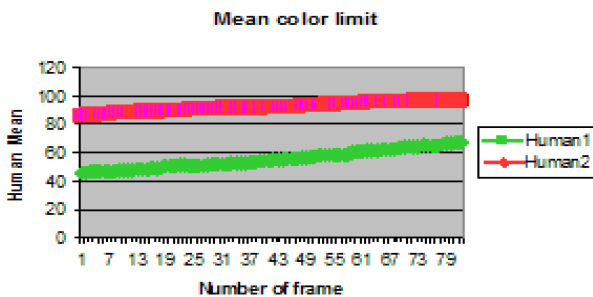


Figure 12: cluster chart for human detection and tracking for 80 video frames

The numerical example below is given to understand this iteration

Step 1: First element in the list as reference = 51.46127075 id =1

Step 2: $|51.46127075 - 52.16221053(\text{next element})| < 3$ id =1

Average = $(51.46127075 + 52.16221053) / 2 = 51.81174$

reference = 52.16221053 repeat with next

$|52.16221053 - 51.16424035| < 3$ Yes id =1

Average = $(51.46127075 + 52.16221053 + 51.16424035) / 3 = 51.595907$

reference = 51.16424035 repeat with next

$|51.16424035 - 85.91414566| < 3$ No id =0 skip to next

reference = 51.16424035

$|51.16424035 - 51.97286603| < 3$ Yes id =1

Average = $(51.46127075 + 52.16221053 + 51.16424035 + 51.97286603) / 4 = 51.690147$

reference = 51.97286603 m

$|51.97286603 - 88.42308721| < 3$ 3 No id =0 skip to next

reference = 51.97286603

$|51.97286603 - 52.36578947| < 3$ 3 Yes id =1

Average = $(51.46127075 + 52.16221053 + 51.16424035 + 51.97286603 + 52.36578947) / 5 = 51.82528$

Step 3: Repeat step 2 average 51.82528 as reference for all list where item not defined

Step 4: Get minimum value of cluster 52.16221053 get maximum value of cluster 52.36578947

Step 5: For the list any value between min and max cluster value and not recognized consider as the same cluster

Step 6 : Repeat step 1 where the first list have id =0

As a result for clustering algorithm multiple human by mean value can be identifying. The best image for each human is the maximum width and height for each cluster in the list which can be saved as best human position.

5. CONCLUSION AND FUTURE SCOPE

The main contribution of this paper is to detect and track multiple humans in video at real time. The algorithm is based on fast human detection based on calculating Human body proportions from upper body detection and calculating an expecting calculated face part area and check the skin color for pixel face, if one pixel found then it is human . We applied this algorithm to progressive video frames and obtained a very good experimental result based on time of detection and tracking, we also proposed a new clustering method by applying the quality which depends on both the similarity measure used by the method and its implementation and the ability to discover some or all of the hidden patterns. This research work was initiated as a part of research project for Human Actions Detection In Content-based Video Retrieval System. In the future this algorithm will be an essential part of a system which will identify human presence in video.

6. ACKNOWLEDGMENTS

Practical Application was done at Laboratory of Image Processing, Zagazig University Egypt. With appreciation and gratitude to International Journal of Computer Applications Staff and research paper Referees.

7. REFERENCES

- [1] A.Maghraby M.Abdalla O.Enany,Hybrid Face Detection System using Combination of Viola - Jones Method and Skin Detection, International Journal of Computer Applications , May 2013 ISBN : 973-93-80875-36-7
- [2] A.Maghraby M.Abdalla O.Enany, Detect and analyze face parts information using Viola-Jones and Geometric approaches, IJCA September 2014,ISBN : 973-93-80883-64-3
- [3] Kruppa H., Castrillon-Santana M., and B. Schiele. "Fast and Robust Face Finding via Local Context". *Proceedings of the Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2003, pp. 157–164.
- [4] Smith, W. Stevenson, and Simpson, William Kelly. *The Art and Architecture of Ancient Egypt*, pp. 12-13 and note 17, 3rd edn. 1998, Yale University Press (Penguin/Yale History of Art), ISBN 0300077475..
- [5] Frederick, D. A. et al. (2010). The influence of leg-to-body ratio (LBR) on judgments of female physical attractiveness: Assessments of computer-generated images varying in LBR In *Body Image*. 7(1):51-55
- [6] Strehl A. and Ghosh J., Clustering Guidance and Quality Evaluation Using Relationship-based Visualization, *Proceedings of Intelligent Engineering Systems Through Artificial Neural Networks*, 2000, St. Louis, Missouri,
- [7] USA, pp 483-488.H. Zha, C. Ding, M. Gu, X. He and H.D. Simon. "Spectral Relaxation for K-means Clustering", *Neural Information Processing Systems vol.14 (NIPS 2001)*. pp. 1057-1064, Vancouver, Canada. Dec. 2001.