HUMAN DETECTION IN VIDEOS

[1]Muhammad Usman Ghani Khan, [2]Atif Saeed

[1]The University of Sheffield
[2]Lecturer, Department of CS, COMSATS Institute of Information Technology, Lahore

Email: [1]usmanghanikhan@gmail.com, [2] i_atif7@hotmail.com

ABSTRACT

Extracting high level features is an important field in video indexing and retrieving. Identifying the presence of human in video is one of these high level features, which facilitate the understanding of other aspects concerning people or the interactions between people. Our work proposes a method for identifying the presence of human in videos. The proposed algorithm detects the human face based on the colour and motion information extracted from frames over wide range of variations in lightning conditions, skin colour races, backgrounds and faces' sizes and orientations. Experimental results demonstrate the successfulness of the algorithm used and its capability in detecting faces under different challenges. The proposed work is crucial in lots of applications whose concern is mainly human activities and can be a basic step in such activities.

Keywords: - *Video Processing, Computer vision, Human detection, Face recognition*

1. INTRODUCTION

Identifying the presence of human in video streams is one of the most important features that must be extracted. However, this task becomes more complicated in the presence of different variations in brightness, lightings, contrast levels, poses, and backgrounds. This work proposes a method to identify the presence of humans in a video sequence and differentiating them from non-human objects. The purpose of this work is to address the human detection problem and give a method by which a variety of face detection algorithms are used in a certain sequence towards achieving acceptable robust results. Each of these algorithms deals with the face from different angle of views and the goal behind that is to maximize the number of correctly detected faces by removing as much noise as possible after going through several tests. The proposed algorithm, can detect different faces of different skin races and sizes and under different lightning conditions. It represents the first step towards accomplishing other goals such as face recognition, analyzing facial expressions, finding shots of people shaking hands, having banners or signs, entering or leaving a building, or sitting in a meeting and others listed by TRECVID 2005 [18]. The rest of the paper is organized as follows; related work is discussed in section 2, section 3 describes our proposed methodology, section 4 is about results and discussions, then comes conclusion and at the end future recommendations and references are provided.

2. RELATED WORK

2.1 "People" Feature Identification

Most of the studies in this field use face detection algorithm as the key idea. Jin [15] proposed a method to identify video shots with people based on face detection. The category of the shot was considered to be "people", only if there is at least one image with more than one face within that shot. One of the three features chosen by Huang et al. [16] to be evaluated in the TREC video Evaluation (2003) was "People" feature, Huang et al., state that for a segment of video to have people feature it should contain at least three human faces. Huang et al. used a skin-tone filter to detect skin regions, followed by the omni-face detection algorithm which was proposed by Wei and Sethi [23].

2.2 Human Detection

From the literature reviews done, it can be concluded that most common way in human detection is via detecting human face. Human face is the most unique part in human body, and if it is accurately detected it leads to robust human existence detection.

2.3 Face Detection methods

Several studies were done in face detection field since 1970, and lots of surveys addressed the algorithms used in this field under different categories [5], [13], [14] but in general two main classes can be used to classify these algorithms namely, feature based (e.g. Bottom-Up) and image based (e.g. Appearance-Based and Template matching) approaches. Features based approaches extract facial features from an image and manipulate its parameters such as angles, size, and distances. Image base approaches rely on training and learning set of examples of objects of interest. However, dealing with video introduces other approaches for face detection such as motion based approach. A brief description of the most common approaches and examples of algorithms used in each of them is given in the rest of this section.

2.3.1Knowledge-based (Top-Down) approach

In this method the relationship between facial features is captured to represent the contents of a face and encode it as a set of rules. Coarse-to-fine scale is used in lots of algorithms classified under this category, in which the coarsest scale is searched first and then proceeds with the others until the finest scale is reached.

2.3.2 Feature invariant (Bottom-Up) approach

In this approach, the face's structural features which do not change under different conditions such as varying viewpoints, pose angles and/or lightning conditions. Common algorithms used under this category are:

Colour-based approach, or so called skin-model based approach. This approach makes

use of the fact that the skin colour can be used as indication to the existence of human

using the fact that different skins from different races are clustered in a single region.

Cezhnevets et al., [21], presented 4 pixel-based skin modelling techniques named as Explicitly defined skin region, Non-parametric skin distribution modelling, Parametric skin distribution modelling, Dynamic skin distribution modelling.

Facial features based approach This method, in which global (e.g. skin, size, and shape) and/or detailed (e.g. eyes, hose, and lips) features are used, has become popular recently. Mostly, the global features first are used to detect the candidate area and then tested using the detailed features.

Texture The human face differs from other objects in texture. This method, examines the likelihood of sub image to belong to human face texture, using Space Gray Level dependency (SGLD) matrix.

2.3.3 Template matching methods

These methods are based on measuring the degree of similarity between the candidate sub image and the predefined stored face pattern. The predefined image might be for the whole face pattern or the individual face features such as eyes, nose and lips. Common algorithms used under this category are:

Predefined face templates, in which several templates for the whole, individual or both (whole and individual) parts of a face are stored.

Deformable Templates in which an elastic facial feature model as a reference model where the deformable template mode of the object of interest, is fitted in.

2.3.4 Appearance-Based Method

Unlike template matching methods, where the templates are predefined by experts,

Appearance-Based method learns the templates from set of images, using statistical

analysis and machine learning. Examples of algorithms used by these approaches are:

Eigenfaces, or so called eigenvectors, in which different algorithms are used to approximate the eigenvectors of the auto correlation matrix of a candidate image. [27]

Distributed-Based, where the distribution pattern of an object is learned using the positive and negative image sets of that object.

Neural Networks, where networks of neurons (simple Elements) called nodes are used to perform function in parallel. The idea of neural networks comes from the central nervous system. However, these networks are trained to detect the presence of face by giving it face and no face samples.

Support Vector Machines, these are learning machines that make binary classifications. The idea here is to maximize the margin between positive and negative sets of vectors and obtain an optimal boundary which separates the two sets of vectors. They were first suggested by Vapnik in 1960 [4].
Hidden Markov Model is a statistical model used to model the statistical properties of a signal. The Markov process is used to model the processed system and the Markov parameters are taken from the observed parameters.

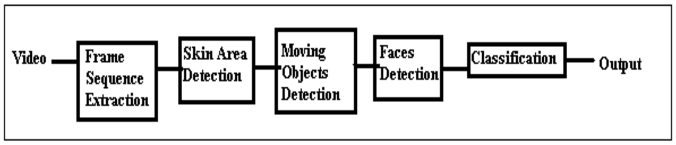
2.4 Movement Detection

Unlike still images, video sequences hold more details about the history of moving objects (foreground), which help in isolating the foreground from the background. Generally, the moving areas are detected by finding the changes that happen among the sequences of images [1], [2].Most of the research done in movement detection applied pre-processing steps before applying the change detection algorithms, [2]. Such pre-processing steps involve geometric and intensity adjustments. The problem of variation in light intensity is solved by intensity adjustment in which illumination effect is reduced to some degrees based on the method used. Elgammal et al. [1], state that transforming the RGB values, into chromatic colour space makes the module insensitive to the small changes in the illumination.
There are several ways for detecting a change in a video sequence [2]. Recent studies agree that Image differencing method is more effective than others in change detection [3].

3. PROPOSED METHODOLOGY

Our proposed algorithm comprises of following steps.

1. Converting a video sequence in to individual images.
2. Accessing the sequential images and detecting the important features.
3. Allocating those regions (if any) giving indications of human presence such indication is having a human skin like colour.
4. Applying movement detection test for all of the allocated regions.
5. Applying face detector to those detected moving objects to detect if it is a face or not.
6. Face detector needs to work as a series of filters that filter out the noise (non face regions) through different tests that utilize the spatial as well as the temporal information available.



Steps involved in human detection

Stage 1: Frame Sequence Extraction

In this stage video (MPEG formatted videos) is converted into individual JPG frames

.

Stage 2: Skin Areas Detection

In this stage colour information of the digital image is utilized to find those areas close to human skin colour. This stage helps in reducing the search space and therefore speeds up the simulation by consuming the processing time efficiently. However, skin test is not enough to detect human faces as it will also detect other parts of the body as well as other non face skin coloured objects. Thus, other tests to filter out those unwanted areas should be applied. Further stages in the proposed project are designed to gradually eliminate the false detected areas found at this stage. The first test to remove the unwanted skin like areas was chosen to be movement detection.

Stage 3: Moving Objects Detection

To minimize the errors in face detection we can utilize the human nature that human will have at least small amount of movements such as eyes blinking and/or mouth and face boundary movements. We can get this information easily because we are dealing with video sequence by which the whole sequence of the object's movements can be obtained.
Taking that point in to account we can reduce the error that occurs due to false detection of a human face and minimize the time of simulation. This step was designed to be implemented only across those skin regions found in the previous step. Giving those moving pixels different colour

than surrounding region (human face skin colour in case of face was detect), these pixels reshape the human face facial features which in turn helps in later stages. However it is important to take in to account that a change may occurs due to several sources such as moving objects, presence or absence of objects
camera movement and zooming, brightness changes This means that some changes are significant and others are not and this is varying with the application requirements. For example, the change detected in background is not significant in video surveillance whereas it has a great importance in remote sensing. Although, it is difficult to take decision whether a detected change is significant or not, it is an important step to remove unwanted changes and focus the processing only on those changes of interest, which reduced the processing time and false detected areas. Hence, movement detection was chosen as a vital stage in the proposed design.

Stage 4: Face Detection

To insure that the moving part is a face, additional tests are required. In this stage, the moving objects which were detected in the previous stage are examined to identify if any of them is a face by examining the pass of the following four tests:

Geometric Test

The candidate regions are tested here against some human face geometric features which are governed by the relation between the width and height of the human face. This test is important to eliminate some of those regions which contain non face objects whose colours are similar to the human face skin colour and experience some acceptable movement across the frames sequence.

Temporal Test

In this test the advantage of having the temporal information from the frame video sequence is being utilized to help in constructing additional verification step before applying the further tests which are more computationally expensive. The principle used in this step based on the fact that no face can occur or disappears suddenly in a certain sequence hence, comparison with the previous and next frames (if exist) gives indication weather or not the candidate region is a face. This additional step helps in maximizing the elimination of those detected skin areas that do not include a human face. Therefore, a reduction in the computational efforts as well as simulation time will be achieved.

Facial Feature Test

This stage will examine the existence of the facial features (mainly the moving areas (none skin) such as eyes, lips and face boundary) in the candidate skin areas filtered in the previous test. It is useful to use the fact that the face skin region must have at three separate spots as a test to be applied in order to verify the existence of a human face in the candidate areas. Only those candidates passing this test will be proposed for the next verification tests.

Template Matching Test

Those candidate regions that have passed the previous two tests will be compared (correlated) with a template model of a human face. Only those candidates that achieve a correlation value beyond a pre defined threshold and a distance (from the face space) value less than a pre defined threshold will be considered as a human face.

Stage 5: Classification

Only those allocated faces that have passed all of the verification stages successfully undergo the classification stage where the category of the detected candidate region is classified as either a face or not face in this stage based on the results of the last two tests namely correlation and distance from face space tests. The candidate region to be accepted as a face has to have a correlation value above the specified correlation threshold value and its distance from the face space should be below the specified distance threshold value.

4. RESULTS AND DISCUSSIONS

Following are the components of the test bed used for our work.
Although, YCbCr colour space is used to separate the luminance component from chrominance components [19], Kovač et al. [8] claim that RGB colour space achieves better performance. RGB colour space was chosen for skin detection stage. However, the case is different in the movement detection stage, where

the algorithm needs to be insensitive to small changes happened by the variations in the brightness component of the colours. Thus, Chromatic colour space was used in movement detection to make the process invariant to the illumination.

MPEG format has been used in implementing the proposed project, as it is the video standard format and used by most of the news video. The individual frames format corresponding to the MPEG was chosen to be JPG. Matlab is used as the Programming Language

Main sources of data are,

1- The data provided by TRECVID 2005, [18], for different channels sources. There are 169 hours of news video from Arabic, Chinese, and American sources, collected in November 2004, all in MPEG-1 format.

2- AT&T data base laboratories: consists of 400 images for male and female face in different views (frontal and side views with normal, happy or angry expressions). The faces' two eyes exist in all of the views. [33]

3- Productive Aging Laboratory, PAL face database. [35]

4- Internet images and videos.

5- Friends and family records.

Remaining part gives the details of the results of the proposed project associated with the discussions. At the end of this chapter different samples of both successful and failed detections are represented.

4.1 Frame Sequence Extraction Results

The frames were extracted perfectly in this stage, and the integration with system to be directly giving the out put in to the desired derives was successfully working.

4.2 Skin Areas Detection Results

Skin test was implemented using the results found by Franc Solina [8], [9]. Figure 4.1 shows the result of the software simulation of this stage. The first frame is the reference image and the second frame is that reference image after applying skin test on each pixel in it.

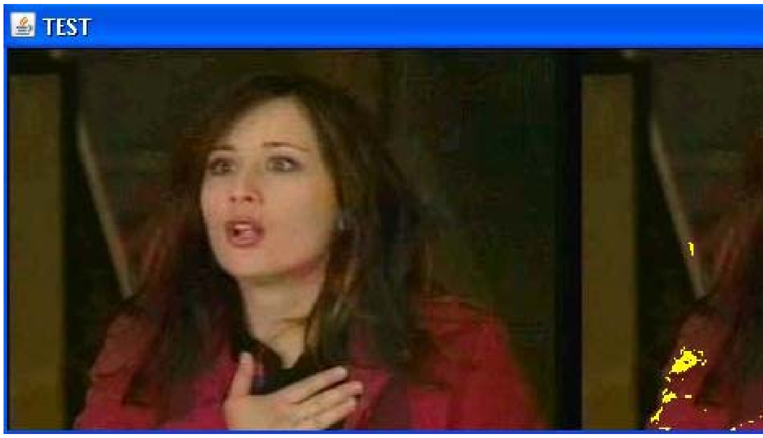


Figure 4.1: *Skin Area Detection*

4.3 Moving Objects Detection Results

The result of subtracting frames sequence is shown in Figure 4.2, the first frame is the reference frame, the second frame is the fifth frame of the 16 consecutive frames following the reference frame, (the full sequence of the (16 frames) is not shown), third frame represents the reference frame with black areas indicating the moving parts.

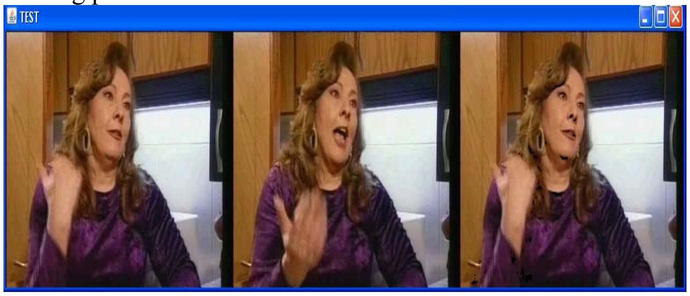


Figure 4.2: *Movement detection using 16 frames.*

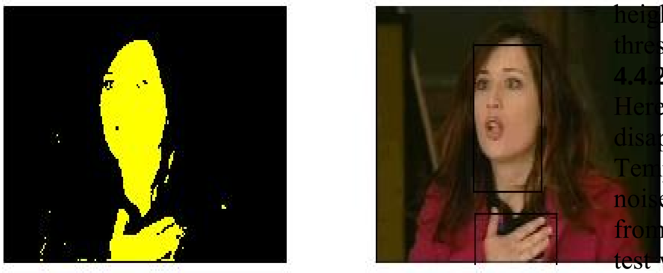
All other unwanted parts will be filtered out by applying the next stage, by which all moving areas that are not within a skin region will be filtered out (discarded).

4. 4: Face Detection

Implementing this stage involves implementing 4 different tests to verify if the input candidates coming from the pre-processing stages are faces or not. The results of implementing these 4 tests are discussed in this part of the report individually as well as in general after integrating the different stages.

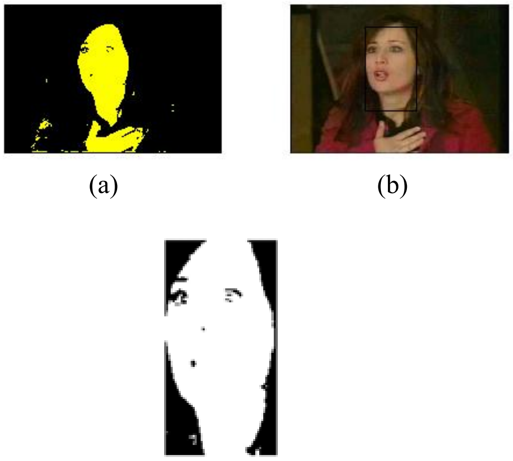
4.4.1 Geometric Test Results

Performing this test improves the out put as it discards lots of those detected areas that do not have specific human face geometry features. But not all of the false detected areas have been removed as lots of noises have same geometric dimensions as human face geometry.



(a)&(b) ***Figure 4.3:*** *(a) The corresponding skin regions. (b) The candidate areas before applying geometric tests.*

Figure 4.3 demonstrates the removal of the lots of noise areas and concentrating the search space in those areas that have passed the following test after skin areas detection stage has been applied: Motion test applied in the moving objects detection stage and Euler number test.



(a)

(b)

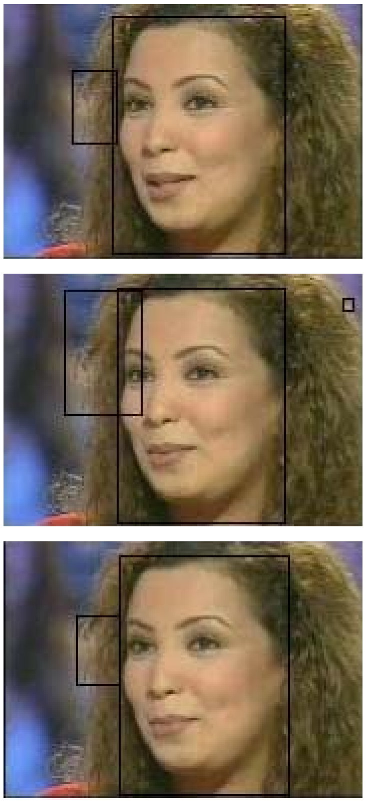
(c)

Figure 4.4: *(a) The corresponding skin regions. (b) The detected area after applying geometric tests.*
(c) The motion result of the candidate region.

In comparison with Figure 4.3, Figure 4.4 demonstrates the effect of passing through the geometric test. Motion test was unable to remove the hand as there was a significant motion among the frames in that area, Euler number was also not able to remove the hand as the number of the spots was more than three. Put, when the two candidates' areas reach the geometric test, the hand did not pass as its width was more than its height by an amount that was above the applied threshold.

4.4.2 Temporal Test Results

Here assumption is that no face can occur or disappears suddenly in a video sequence. Temporal test helps to high degree in eliminating noise areas. Hence, it prevents the noise it detects from going in to further tests to be verified. This test worked as a filter that filtered out as much as possible of those detected areas that include noise instead of a human face.



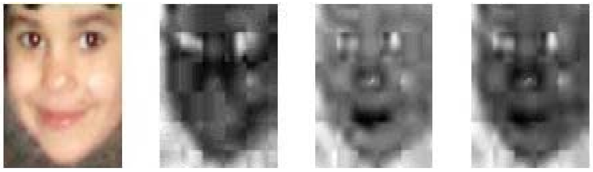
(a) & (b) & (c) Figure 4.5: *Three sequential frames before implementing the temporal test (a) The first frame (b) Second frame (c) Third frame*

Among the three sequential frames presented in Figure 4.5, the middle frame contains an additional detected area in the top left end that does not overlap with any of those detected areas in the previous or next frames. Such objects is rejected by the temporal tests based on no sudden faces can appear or disappear in a video sequence, whereas the other two rectangles are sent to be verified be the further tests in the algorithm. However, this test increased the time of processing for those correct detected areas because of the comparison stage between the previous and next frames. Therefore, temporal test will be chosen for those applications sacrifices the speed towards having more robust system especially with videos having frame with complex back ground.

4.4.3 Facial Feature Test Results

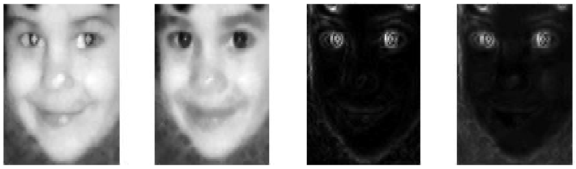
In this stage only the eyes map was implemented. Figure 4.6 and Figure 4.7 and Figure 4.8 show the result of implementing the proposed algorithm.

Eyes Map result:



(a) **(b)** **(c)** **(d)**

Figure 4.6: *Generating the eyes chrominance component. (a) Face (b) (Cb2) (c) (Cr) C (d) (Eyes_Map_C)*



(a) **(b)** **(c)** **(d)**

Figure 4.7: *Generating the eyes luminance component (a-c). Eyes Map (d). (a) Dilation**.** (b) Erotion. (c) Luminance component of the eyes. (d) Luminance & chrominance components of the eyes.*

Mouth Map result:

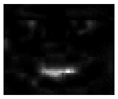


Figure 4.8: *Mouth Map.*

4.4.4 Template matching test

The generated face model (average face) using the proposed algorithm is shown in Figure 4.9. All of the candidate faces were correlated with the average face and based on the threshold value that have been chosen experimentally the faces were rejected or accepted. (See table 4.2)

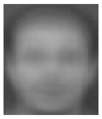


Figure 4.9: *(The average face of 400 training data (AT&T data base laboratories))*

General Results

To evaluate the robustness of the algorithm used in implementing the proposed project, it was applied to 24 video sequences each of these sequences consists of 10 frames. The total number of faces exists in these 240 frames is 470. Among these frames there were 20 frames without any human presence, and all others contain various numbers of faces. 30 frames were having very complex background (background with colour similar to the human skin colour). To make this evaluation more reliable, the test data that have been used were independent of the development data and were collected from different resources (discussed at start of section 4):

Further more, among the above listed sources the sub samples that have been used were carefully chosen to include different sequences, under different lighting conditions and have variety of back ground and faces' sizes. Table 4.1 shows the results of this project.

Correct Hit (TP)	Missed Faces (FN)	FP	PPV	TPR
360	110	148	71%	76.6%

Table 4.1: *Test results*

Threshold Field	**Threshold Value**
Skin	Solina [8, 9]
Movement	0.04
Golden ratio	1.618 0339 887
Height to width	0.59
Correlation	0.5
Distance from face	26.8649

space	

Table 4.2: *Related thresholds*

In the assessment process the result were evaluated using receiver operation characteristic (ROC), [36]. Where by the correct hit or called true positive (TP) is only given for those rectangles which have a face with mouth and eyes features. On the other hand, false negative (FN) is equivalent to miss which occurs if the algorithm fails to detect an existing human. More over, true negative (TN) is corresponding to the undetected areas which are not true (not a face). Whereas, those detected areas that do not include a face are called false positive (FP). However, we are interested in the ratio of the number of true positive out comes to the total number of the existing faces (TP + FN), which is called True Positive Ratio (TPR) or called Hit Rate Recall Sensitivity. In addition to another ratio which is the Positive Predicted Value (PPV), that is represented by the ratio of the number of true positive out comes (TP) to the total number of the detected regions (TP + FP). This ratio is also called precision value. From the implementation of every stage individually and in combination with others it was found that degree of robustness is mainly depending on the robustness of the first stage which is skin area detection stage. All of those faces that have not been detected by the skin area detection stage will not be detected by the further stages. On the other hand, most of those falsely detected skin areas will gradually be discarded through the algorithm's further stages. Hence, skin area detection stage is the most crucial stage in the proposed algorithm.

5. CONCLUSION

An algorithm has been proposed to detect the presence of human in video sequence. The main two techniques used in building the proposed algorithm are face and motion detection techniques. A series of stages were implemented in a certain order to promise maximizing the detection of existing faces and eliminating the other objects (noise). The proposed algorithm detects faces of different sizes under different lightning conditions.

6. FUTURE RECOMMENDATIONS

The proposed work handles the first step in all of those applications concerning human recognition or human activities such as identifying the shots that include different activities such as meeting, running and hand shaking. Further more, its contribution in more advance applications where integration between image processes and audio processing is required to understand the videos and find out different high level features such as identifying the speaker in a certain sense, widens its importance and its applied field.

REFERENCES

[1]. Elgammal.A., Duraiswami.R., Harwood.D., And Davis.L., (2002) "Background and Foreground Modeling Using Nonparametric Kernel Density Estimation for Visual Surveillance" Proceedings of the IEEE. vol 90, no. 7 pp.1151-1163.

[2]. Richard J. Radke, Andra.S., Al-Kofahi.O., and Roysam.B**, (2005)**" Image Change Detection Algorithms: A Systematic Survey". IEEE Transactions on Image Processing, VOL. 14, NO. 3.

[3]. http://www.cast.uark.edu/local/brandon_thesis/Chapter_IV_change.htm last accessed on [3th of September 2008]

[4]. Burbidge.R., Buxton.B., "An Introduction to Support Vector Machines for Data Mining", available from: http://www.martinsewell.com/datamining/BuBu.pdf

[5]. Peer .P., Solina.F., (1999). "An Automatic Human Face Detection Method", Project J2-8829, available from [http://lrv.fri.unilj.si/~peterp/publications/cvww99.pdf], last accessed on [1st of May 2007]

[6]. Kriegman.J., Yang.M., Ahuja.N., 2002 "Detecting Faces in Images: A Survey" IEEE Transactions On Pattern Analysis And Machine Intelligence, VOL. 24, NO. 1 pp. 34-58

[7]. Gonzalez, Rafael C. & , Woods.R.E. (1992), Digital Image Processing /Rafael C. Gonzalez, Richard E. Woods . - Reading, Mass.; Wokingham : Addison-Wesley, (1992)

[8]. Kovac. J., Peer. P. and Solina. F., (2004) "Human Skin Colour Clustering

for Face Detection" available from: [http://www.princeton.edu/~aabdalla/ele579/Human%20Skin%20Colour%20Clustering%20for%20Face%20Detection.pdf] last accessed on [1st of September 2008]

[9]. Peer. P., Solina. F., Batagelj. B., Juvan. S., & Kovaˇc. J., (2003) "Colour-Based Face Detection in The “15 Seconds Of Fame”

[10]. Clarke, R. J. (Roger John), Digital compression of still images and video / R. J. Clarke (1995). London: Academic Press, 1995.

[11]. http://en.wikipedia.org/wiki/Video_codec, last accessed on [5th of October 2008]

[12]. http://en.wikipedia.org/wiki/Video_compression, last accessed on [5th of October 2007]

[13]. Yow. K. C., Cipolla. R., (1997) "Feature-based Human Face detection", no. 15, pp. 713-735.

[14]. Low, B. K & Hjelm°as1, E. (2001), "Face Detection: A Survey"

[15]. Jin, R. & Hauptmann, A. G., "Learning to Identify Video Shots With People Based on Face Detection", July 6-9. 2003

[16]. Huang, X., Wei, G. & Petrushin, V. A., "Shot Boundary Detection and High level Features Extraction for the TREC Video Evaluation 2003"

[17]. http://java.sun.com/products/java-media/jmf/2.1.1/solutions/FrameAccess.html, last accessed on [9th of September 2008]

[18]. Over, P., Ianeva, T., Kraaijz, W., & Smeaton, F. A."TRECVID 2005 – An Overview", March 27, 2006.

[19]. Kuchi, .P, Gabbur, P., Bhat, P. S., & David, S. S. "Human Face Detection and Tracking using Skin Colour Modeling and Connected Component Operators", (2002)

[20]. http://java.sun.com/products/java-media/jmf/, last accessed on [6th of October 2008]

[21]. Vezhnevets, V., Sazonov, V., & Andreeva, A., "A Survey on Pixel-Based Skin Colour Detection Techniques"

[22]. Gomez, G., Morales, E.F. "Automatic Feature Construction and a Simple Rule Induction Algorithm for Skin Detection"

[23]. Wei, G. & Sethi, I. K., (2000) "Omni-Face Detection for Video/Image Content Description"

[24]. http://www.ucl.ac.uk/oncology/MicroCore/HTML_resource/PCA_1.htm last accessed on [29th of September 2008]

[25]. http://en.wikipedia.org/wiki/Karhunen-Lo%C3%A8ve_transform last accessed on [29th of October 2008]

[26]. http://en.wikipedia.org/wiki/Eigenface last accessed on [20th of October 2008]

[27]. M. Turk and A. Pentland. Eigenfaces for recognition. Journal of Cognitive Neuroscience,3(1), 1991a. URL http://www.cs.ucsb.edu/~mturk/Papers/jcn.pdf.

[28]. Ashish Tiwari. Face Recognition Eigen-faces with 99 PCA coefficients.ppt

[29]. Jon Krueger, Doug Kochelek, Marshall Robinson & Matthew Escarra.

[30]. Thresholds for Eigenface Recognition. Version 1.2: Dec 17, 2004

[31]. Hsu, R., Abdel-Mottaleb, M. & jain, A. (2002) "Face Detection in Colour Images"

[32]. http://goldennumber.net/ last accessed on [9th of October 2008]

[33]. http://www.cl.cam.ac.uk/research/dtg/attarchive/facesataglance.html last accessed on [9th of October 2008]

[34]. http://www.fourcc.org/fccyvrgb.php last accessed on [19th of November 2008]

[35]. http://agingmind.cns.uiuc.edu/facedb/ last accessed on [19th of November 2008]

[36]. "Receiver operating characteristic". Available from http://en.wikipedia.org/wiki/Receiver_operating_characteristic, last accessed on [19th of November 2008]