Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Smoking Detection in Video Footage

Author:

Éamon Dunne

Supervisor:

Dr. Kenneth Dawson-Howe

A Dissertation submitted in partial fulfilment
of the requirements for the degree of
MAI (Computer Engineering)

Submitted to the University of Dublin, Trinity College, May, 2018

Declaration

I, Éamon Dunne, declare that the following dissertation, except where otherwise stated, is entirely my own work; that it has not previously been submitted as an exercise for a degree, either in Trinity College Dublin, or in any other University; and that the library may lend or copy it or any part thereof on request.

Signature: ____________________ Date: ____________________

Summary

The goal of this project is to develop an automated system that can detect instances of smoking in video footage. A significant number of modern films depict some form of tobacco use, with potentially detrimental effects on the behaviour of young people. As such, an automated recognition system for this action may be useful for the purposes of film classification.

To identify instances of smoking, the system relies on the probability that a person smoking will be in close proximity to cigarette smoke. The system classifies a smoking event by searching for smoke near a person's hand or face.

The system first identifies faces within each frame. The system then takes a sample of each face's colour to find areas of the same skin colour within the frame. The largest area of skin outside of the identified face region is presumed to be the hand. The system then searches the face and hand areas it has identified for smoke. If smoke is found in these areas, that particular frame of the video is classified as a potential smoking frame. If a significant proportion of frames in a video are potential smoking frames, the video itself is deemed by the system to contain an instance of smoking.

The system was implemented in C++, using the OpenCV library. The final system was designed to work with short sequences, typically less than 10 seconds in length. Additionally, the system's usage of a Gaussian Mixture Model to detect motion limits it to footage shot with a stationary camera.

A selection of short films clips, indicative of modern films, were chosen to test the system's performance. The system's final results were mixed, partially due to the variety and complexity of the footage used (an issue inherent with testing on footage from modern films). The large variation present in the footage makes it difficult to optimise the system, as the range of input is extremely broad. Calibrating the system to properly classify as many positive samples as possible will introduce many false positives, and vice-versa.

There is scope for improvement, as a possible avenue of future work. In particular, refactoring the system to properly handle longer footage, as well as footage shot with a non-stationary camera, would be beneficial.

Acknowledgements

I would like to thank my supervisor, Kenneth Dawson-Howe, for his constant advice and encouragement over the course of this project.

I also wish to thank my family, in particular my parents, for their support over the last year.

Contents

Code from project is on attached CD.

1 Introduction

1.1 Aims

The aim of this project is to develop an automated system that can determine whether smoking occurs in a video sequence.

Given a series of short videos as input, the system should return a positive or negative result for each one, based on whether or not smoking is depicted. For the purposes of rigorous classification, a video is deemed to contain smoking if at any point, someone inhales and exhales from a cigarette. Figure 1.1 shows some typical examples of input.

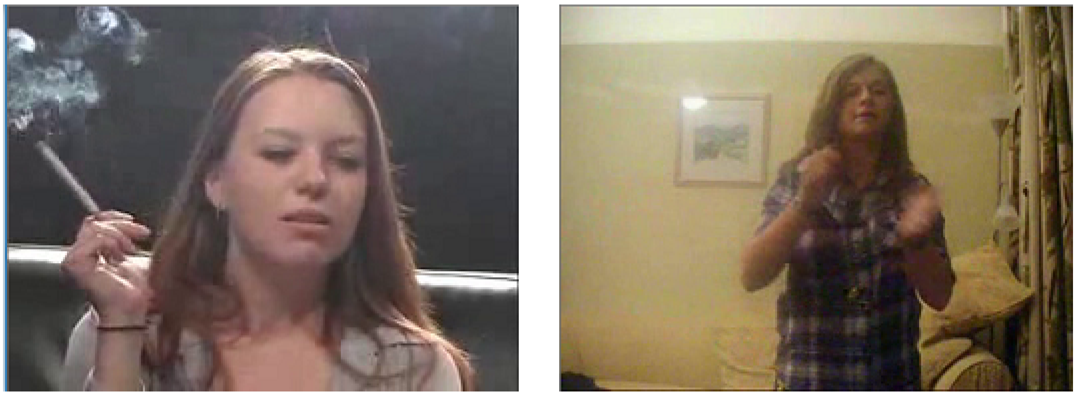


Figure 1.1: Stills from a smoking video (left) and a non-smoking video (right). Images source: [HMDB].

1.2 Motivation

Roughly half of the highest-grossing films in the US from 2010-2016 (across all age ratings) depicted some form of tobacco use [Tynan, 2017]. If this is any way reflective of worldwide trends, it suggests that depictions of smoking are being shown to very large audiences.

The potential issue with this is that separate studies by both the US Surgeon General [Surgeon General, 2012] and the World Health Organisation [WHO, 2015] have found a link between young people being exposed to on-screen tobacco use and taking up the habit themselves. In their 2015 report on tobacco use in films, the World Health Organisation went so far as to recommend that films depicting smoking be given an adult rating [WHO, 2015].

This suggests a potential need for film classification to take depictions of smoking into account. An automated system for detecting smoking in films would be of possible interest to both classification boards and tobacco advocacy groups.

2 Literary Review

Existing projects in the area of smoking detection typically rely on a combination of face detection, object recognition and smoke detection.

2.1 Face Detection

Face detection is used to find people within footage and hence, verify that a person is actually present. This helps to distinguish between footage of smoking and footage of smoke alone (e.g. wildfires, chimneys, etc.).

The use of Haar classifers, put forward by Viola and Jones [Viola, 2001], is very frequently used for this purpose. A trained cascade classifier identifies faces by searching for a particular set of features in every area of the frame.

A different approach to face detection was put forward by forward by Hsu et al. [Hsu, 2002], as an alternative to face detection algorithms that require large volumes of training data. Areas of skin colour in an image can be identified by their Cb and Cr values. Skin tone is normally dependent on luminance, but this can be altered by applying a nonlinear transformation to the image. Hence, a range of colour in the Cb-Cr colour subspace can be used as a model of skin colour. Furthermore, the section of an image corresponding to a mouth can be identified by its comparatively reddish colour. If a mouth is present in an area of skin, it can be isolated by comparing its distribution of Cr values to the rest of the area.

2.2 Object Recognition

Hsu et al.'s approach to face detection was used by Wu and Chen [Wu, 2011] to develop a system that could recognise smoking in still images. Given an image, the objective is to determine whether there is a person present with a cigarette in their mouth. First, any faces and corresponding mouths are identified within the image. Then, the system searches for a cigarette within the mouth itself. This is done by simply searching the area around the mouth for a white object in the HSV colour space. Any white object near the mouth is classified as a cigarette.

This system is capable of positively identifying cigarettes, but is vulnerable to false alarms from other white objects held near the mouth (straws, pens, etc.). Additionally, the system struggles to identify white objects in bright lighting or against white backgrounds.

Hsieh et al. [Hsieh, 2014] proposed a smoking detection system for video analysis that follows the same principle, but uses a more sophisticated method of object recognition. Again, faces are identified in the frame (using a cascade classifier) and the face region is isolated. The colour histogram of this region is then examined over a number of frames. If a person holds an object close to their face, the change in colour it introduces will be reflected in the histogram of the face region. Thus, the colour of a handheld object brought close to the face can be accurately determined. Once the colour of the object has been identified, it can be tracked if it moves away from the face in subsequent frames.

The system proposed by Hsieh et al. also searches for smoke itself as a visual indicator of smoking. For each potential smoking frame, three features are recorded: the size of the handheld object, its distance from the mouth of the person in frame and the density of smoke within the frame. Analysing these properties over time reveals if the object is moving towards or away from the mouth, and if the volume of smoke is increasing or decreasing. The system then uses a Hidden Markov Model (HMM) to match these findings to one of three actions: smoking, drinking or using a phone.

2.3 Smoke Detection

Smoke detection itself has been the subject of a considerable amount of research. There are a wide range of techniques in use, but as noted by Kaabi et al. [Kaabi, 2017], they generally follow the same three stages:

- Identify regions of movement within a frame
- Search for smoke features within each identified region
- Conclude whether smoke is present based on the extracted smoke features

2.3.1 Movement Detection

The smoke detection step used by Hsieh et al. uses a foreground segmentation technique proposed by Kim et al. [Kim, 2005] to identify regions of movement within an image. After identifying regions of movement, the second stage of smoke detection searches for pixels within a certain range of saturation and intensity, as well as blurred edges (another potential indicator of the presence of smoke). The extracted features are given as input to a Source Vector Machine (SVM) Classifier in order to finally determine if the region contains smoke.

Wu et al. [Wu, 2010] proposed a similar smoking detection system to Hsieh et al., albeit one that uses a far simpler smoke detection algorithm. Regions of movement are identified using a Gaussian Mixture Model, but the only features considered for smoke classification are the density of moving pixels in the region and the proximity of the region to an identified face.

2.3.2 Colour Analysis

As noted by Kaabi et al., colour is one of the most important features used for smoke detection. Smoke is characterised by a greyish colour and decreased chrominance relative to the surrounding area. In the RGB space, pixels corresponding to smoke will have R, G and B components that are quite close together. In the YUV space, the chrominance (U and V) values will decrease in the presence of smoke.

2.3.3 Energy Analysis

Another key indicator of smoke is that it tends to blur the edges of objects behind it, making them appear less distinct. This can be evaluated through energy analysis, as used by Toreyin et al. [Toreyin, 2006].

The discrete wavelet transform of an image contains its edge information: to be precise, the LH, HL and HH subimages contain the horizontal, vertical and diagonal edges of the original image, respectively. The squares of these subimages can be summed together to give the total energy of the original frame. If smoke has blurred the edges of the image, its energy will be reduced.

To check for smoke, each input frame is divided into blocks of uniform size. The energy is then calculated for each individual block and compared to the corresponding block in a slowly-updating background model. A reduction in energy within a certain threshold signifies the possibility of smoke.

Toreyin et al. used this method, noting that they are looking for edges that would gradually soften over time without disappearing completely. If an edge disappears instantaneously, it is due to a moving object in the foreground obscuring it, rather than smoke. Thus, it should be ignored.

Calderara et al. [Calderara, 2008] and Brovko et al. [Brovko, 2013] compare the energy values of the input frame and the background using a ratio:

$$\frac{\text{Image energy of current frame}}{\text{Image energy of background model}} \tag{1}$$

This has the advantage of normalising the energy values, allowing for a fair comparison. A decrease in this ratio in a particular image block denotes the possibility of smoke being present.

2.3.4 Disorder Analysis

Smoke constantly diffuses in the air, giving it an unstable shape and size. This characteristic is exploited by Chen et al. [Chen, 2004], who evaluate potential smoke areas over time to check they behave in this manner. A true smoke area will deform significantly between frames, while an area that merely looks like smoke will remain more or less the same. Chen et al. use a form of colour analysis to identify potential smoke areas, and then check for this condition to eliminate false positives.

2.3.5 Final Classification

Having extracted potential smoke features (e.g. colour, energy, disorder), these must then be used to intelligently judge whether or not smoke is present. Kaabi et al. [Kaabi, 2017] note that Source Vector Machines are a common choice for this final step, although Bayesian Classifiers and Markov Models are also used. Typically, the extracted features are encoded as a multidimensional feature vector, which is then given as input to the classifier.

3 Design

The aim of this project is to develop a smoking detection system for video footage. Given a short film clip as input, the system should conclude whether or not the footage contains a depiction of smoking.

The finished system was implemented using C++ and the OpenCV library.

This chapter gives a high-level overview of the developed system, and the various techniques used. It also outlines some of the issues with the system's overall design and the assumptions it makes.

A more detailed description of the system is given in the chapter following this one.

3.1 Overview

The system searches for the faces of people in each frame, and tries to find the hand of each person identified. The system then searches the identified face and hand regions for smoke. If smoke is found, the frame is deemed to depict smoking. The system reaches its final conclusion for a particular video clip based on the proportion of frames it has found to depict smoking. If the proportion is sufficiently high, the footage is classified as a smoking clip.

For each frame, the system first identifies all faces present. Haar-based cascade classifiers [Viola, 2001] provided by the OpenCV library are used for this purpose. To reduce the number of false alarms, faces are only considered for the next stage if they persist for three or more consecutive frames. The system checks for overlap between faces from multiple frames to determine whether an identified face has remained.

For each face obtained from the previous stage, the system extracts a small section from the centre as a sample of skin colour. The back projection method provided by OpenCV is then used to search for areas of a similar colour in the frame. The system then verifies that all of these areas are skin-coloured using a colour analysis technique put forward by Kolkur et al. [Kolkur, 2017].

The system now tries to identify a hand in the frame, corresponding to the identified face. The face region is masked out, then OpenCV's implementation of connected components is used to find all remaining contours within the image. The hand corresponding to this face is presumed to be the largest of the contours.

Having defined face and hand areas, the system searches for smoke in these regions. First, the system identifies areas of movement using a Gaussian Mixture Model (GMM) [Stauffer, 1999], provided by OpenCV. The system then identifies smoke-coloured pixels, using a colour analysis algorithm by F. Yuan [Yuan, 2008]. The system verifies that each moving pixel's RGB values fall within certain ranges, giving them a greyish colour. Pixels that fail this stage of analysis are discarded, giving a binary image.

To conclude whether these smoke-coloured pixels represent actual smoke, a disorder analysis algorithm used by Chen et al. [Chen, 2004] is employed. The output of the colour analysis stage from the current frame is compared to that of the previous frame, and that of the next frame. The smoke pixels in the previous frame are subtracted from the smoke pixels in the current frame. Likewise, the smoke pixels in the current frame are subtracted from the smoke pixels in the next frame. The result of each subtraction is a separate disorder image: the difference between these two disorder images indicates the level of disorder in the current frame. Since smoke constantly moves and diffuses in the air, a smoke region should create a relatively high level of disorder. Hence, if the level of disorder in the current frame exceeds a certain threshold, it is deemed to contain smoke.

If this final criteria is found in a frame (an identified hand or face region contains smoke), the frame itself is deemed to depict smoking. The system evaluates every frame of footage in this manner. If a sufficiently high proportion of frames are classified by the system as smoking frames, it concludes that the footage contains a depiction of smoking.

3.2 Design Problems

A major limitation of this system is that it requires the test footage to use a stationary camera. The smoking detection step uses GMM to identify moving objects, and GMM is unable to properly handle moving backgrounds. If the camera moves, GMM treats everything in the frame as moving, and the motion detection fails.

Additionally, the system assumes that in any depiction of smoking, the smoker's face will be clearly visible. This isn't always true: in some cases, the smoker's face may be obscured, turned away or simply not visible in the frame. The assumption that the smoker's hand will always be present is similarly problematic. Ideally, if the hand is not present in frame, the largest skin-coloured contour outside of the face region should be very small. Hence, the extra area chosen by the system to search for smoke should be negligible.

However, this also assumes that the largest skin-coloured contour outside of the face region is the smoker's hand. This assumption fails if significant portions of the smoker's skin are uncovered (e.g. they are wearing a sleeveless garment which leaves their arms bare). If this is the case, the area chosen by the system to represent the hand region may be too large, or in the wrong section of the frame.

The system also assumes that any smoke present near an individual's face or hand is due to smoking. The smoke may be from other sources (e.g. they are standing near an open fire, or holding a smoke-emitting object, such as a recently-fired gun), but the system is unable to distinguish the difference.

The system is also limited to short clips (ideally, less than ten seconds in length). As such, it is assumed that any depiction of smoking will occur over a significant number of frames. While this assumption may be valid for short clips, it is unsuitable for longer footage (such as feature-length films).

4 Implementation

This chapter gives a full and detailed description of each component of the system, the analysis techniques used and how they function together.

As detailed in the previous section, the system analyses each frame of video input in three stages:

- Search for faces in the frame
- Find a hand corresponding to each identified face
- Search for smoke in the areas around each face and hand

If the system finds smoke near an identified face or hand, it concludes that this frame of footage depicts smoking. If a sufficiently high proportion of frames are classified by the system as depicting smoking, the source video is deemed to depict smoking.

4.1 Face Detection

Cascade classifiers based on Haar-like features are commonly used to find objects in an image. This method of detection was first proposed by Viola and Jones [Viola, 2001].

A cascade classifier requires a large volume of training data, with positive and negative samples of the object it is intended to detect. From this training set, a large number of features indicative of the object to be recognised are determined. Each feature is simply the difference between the sum of pixels in a number of rectangular regions (between two and four). Adaboost, a machine learning algorithm, is used to select the most useful features, which are then used in the final classifier. Individually, the selected features are weak classifiers, unable to accurately detect the object being searched for. Combined, they form a strong classifier, with far greater levels of accuracy.

The classifier is divided into a number of stages, each of which contains a number of features. To search for an object in a subsection of an image, the classifier evaluates each stage in turn. If a subimage fails a stage, it is discarded and the other stages are not evaluated. A subimage is only deemed to contain an object if it passes all stages.

For the purposes of this project, face classifiers provided by the OpenCV library were used: one made for detecting frontward faces and another made for detecting faces in profile. The classifiers search for faces in each frame of test footage, returning the position and size of each identified face in the frame.

A key consideration is the number of neighbours each potential face must have before it is classified as a face region. Setting this parameter too low results in many false positives: the classifier will include many face-like objects in the frame, along with the true faces. Setting this parameter too high results in too many false negatives: the classifier excludes many true faces.

An additional step implemented for this system is to track faces from frame to frame, and only include them in further analysis if they persist for three or more consecutive frames. This helps to eliminate a small number of false positives included by the face classifier. To do this, the faces found in each frame are compared with the faces found in the previous two frames. The overlap between two faces is found by performing an AND operation between the two face regions. If a face in the current frame has greater than 90% overlap with a face from each of the preceding frames, it is retained for further analysis. Otherwise, it is discarded.

The final output of this stage is positions and sizes for each potential face in the frame, which can be used to form a bounding box around each one.

4.2 Hand Detection

Having identified the areas in the frame corresponding to faces, the system now searches for a hand to match each face.

4.2.1 Back-Projection

For each face region, the system takes a rectangular subsection from the centre, one quarter of the size of the whole region. This subimage is used as a sample of skin colour of the person this face belongs to. The system finds other areas in the frame of a similar colour using back-projection, a method first proposed by Swain and Ballard [Swain, 1990]. The methods used in this project for back projection are all provided by the OpenCV library.

First, the colour histogram of the skin sample is calculated. The range of colours in the sample is divided into a number of bins. Pixels are then sorted into these bins, with the choice of bin depending on the colour of the pixel. This gives a colour histogram for the skin sample: counts for how many pixels occur in each colour range. The colour histogram is then normalised, so the maximum value is 1. This allows the histogram to represent probabilities: the higher the histogram value, the more likely it is to correspond to a skin pixel.

The skin sample histogram is then back-projected onto the current frame of footage. This sorts the pixel values from the current frame into the bins defined by the skin sample histogram. Effectively, this assigns a probability value to each pixel in the current frame. The higher a pixel's probability value, the more likely it is to represent skin (based on the prepared skin sample). The output of the back-projection is a greyscale representation of the current frame of footage, where the brightness of each pixel is proportional to how likely it is to represent skin (Figure 4.1).

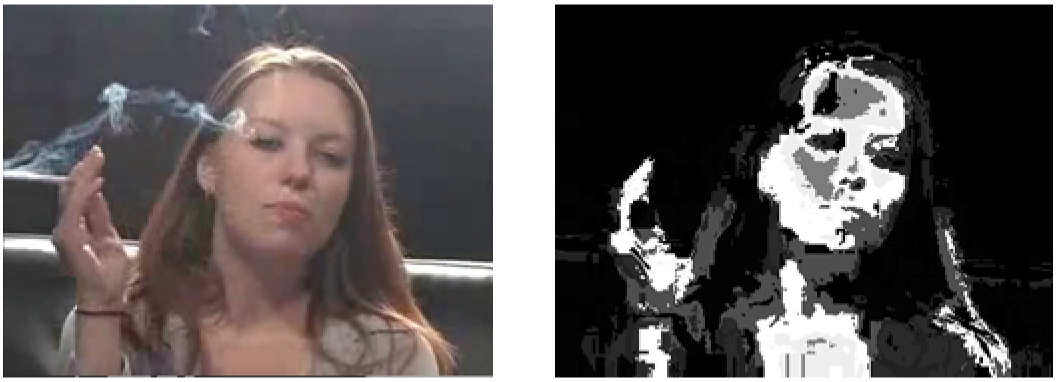


Figure 4.1: Input image (left) and result of back-projection (right).
Input image source: [HMDB].

Pixels with a low probability of representing skin are then discarded by applying a minimum probability threshold to the back-projection image. Pixels with a value exceeding the threshold are coloured white, while those below the threshold are coloured black; thus giving a binary image (Figure 4.2).

Figure 4.2: Back-projection image with minimum threshold applied.

4.2.2 Colour Analysis

The accuracy of the back-projection stage is highly dependent on the skin sample used. This in turn depends on the accuracy of the face detection stage, detailed previously. If the area judged by the face classifier to be a face is off-centre, the sample of skin taken from this area may include incorrect, non-skin colours. This results in large portions of the frame being mistakenly classified as skin (Figure 4.3, Figure 4.4).

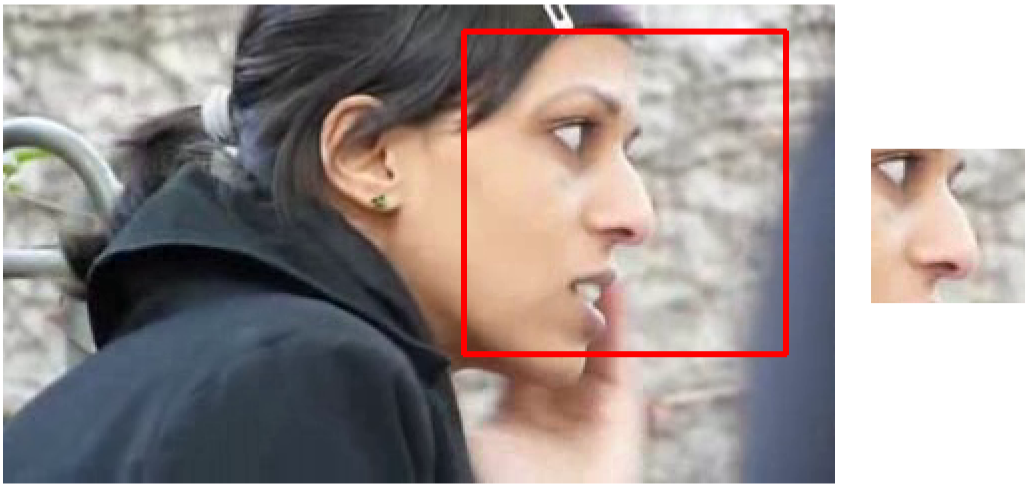


Figure 4.3: Erroneous face detection (left) causes the taken skin sample (right) to include a portion of the background. Input image source: [HMDB].

Figure 4.4: Back-projection misidentifies large sections of the wall in the background as skin.

In order to verify that the pixels selected by the back projection process are actually skin-coloured, an additional colour analysis method, proposed by Kolkur et al. [Kolkur, 2017], is used. The colour of each potential skin pixel in the Red Green Blue (RGB), Hue Saturation Value (HSV) and Luminance Chrominance (YCbCr) spaces are measured, to check that the colour falls within sensible ranges for skin colour.

Kolkur et al. state that for any pixel corresponding to skin, either (1) or (2) should hold true:

$$\begin{aligned}&(0.0 \leqslant H \leqslant 50.0) \text{ AND } (0.23 \leqslant S \leqslant 0.68) \text{ AND}\\&(R > 95) \text{ AND } (G > 40) \text{ AND } (B > 20) \text{ AND}\\&(R > G) \text{ AND } (R > B) \text{ AND } (|R - G| > 15) \text{ AND } (A > 15)\end{aligned} \tag{1}$$

$$\begin{aligned}&(R > 95) \text{ AND } (G > 40) \text{ AND } (B > 20) \text{ AND}\\&(R > G) \text{ AND } (R > B) \text{ AND } (|R - G| > 15) \text{ AND } (A > 15) \text{ AND}\\&(Cr > 135) \text{ AND } (Cb > 85) \text{ AND } (Y > 80) \text{ AND}\\&(Cr \leqslant (1.5862 * Cb) + 20) \text{ AND } (Cr >= (0.3448 * Cb) + 76.2069) \text{ AND}\\&(Cr \geqslant (-4.5652 * Cb) + 234.5652) \text{ AND}\\&(Cr \leqslant (-1.15 * Cb) + 301.75) \text{ AND } (Cr \leqslant (-2.2857 * Cb) + 432.85)\end{aligned} \tag{2}$$

H = Hue; S= Saturation; R = Red; B = Blue; G = Green; Cr, Cb = Chrominance components; Y = Luminance.

To verify this, the RGB source image is copied to the HSV and YCbCr colour spaces. A new binary image is created: pixels are marked as white if they satisfy either of the conditions, black otherwise (Figure 4.5). However, due to a technical limitation of the OpenCV library, the alpha component of each pixel is not retained for video input. Hence, the condition that (A > 15) is not checked for in this stage, and is assumed to be true for every pixel.

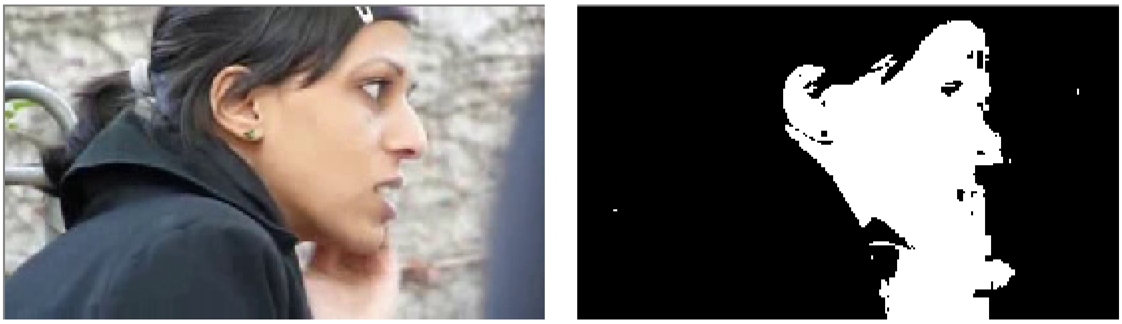


Figure 4.5: Input image (left) and output of colour analysis stage (right). Input image source: [HMDB].

The output of the back projection stage and the output of the colour analysis stage are combined by performing a bitwise AND operation (Figure 4.6). The result is that only the pixels that were classified as skin by both stages are retained as skin pixels. This mitigates the impact of false positives introduced by either stage.

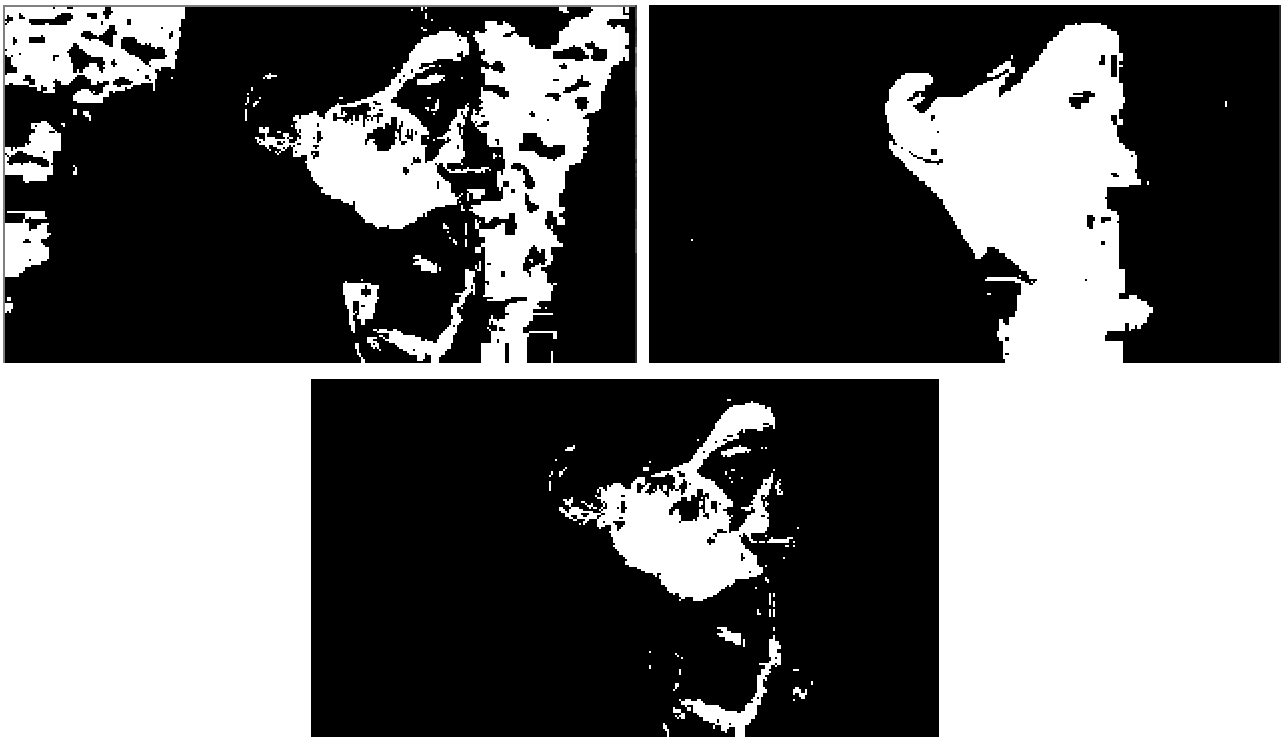


Figure 4.6: Skin pixels from back-projection (top-left) and colour analysis (top-right) are combined to form final skin areas (bottom).

4.2.3 Connected Components Analysis

Having identified areas of skin in the frame, the system uses this information to select an area corresponding to a hand. The system masks out the face region, and selects the largest remaining cluster of skin pixels as the hand corresponding to that face.

To find the largest cluster of skin pixels, connected components analysis is used. This uses the OpenCV implementation of an algorithm by Suzuki and Abe [Suzuki, 1985]. This iterates through every pixel in the image, searching for the boundary of a connected region. Intuitively, the algorithm evaluates pixels from left to right and top to bottom. If a white pixel is found, and its previously-evaluated neighbours are black pixels, then this pixel represents a boundary (Figure 4.7).

The algorithm groups together all of the pixels along this border as a single contour, enclosing a single connected shape within the image. This process is repeated for every contour within the image, giving the full set of connected areas (Figure 4.8).

In order to identify the contour corresponding to the hand, the system masks out the identified face region before finding the contours. The largest remaining contour is taken to be the hand (Figure 4.9).

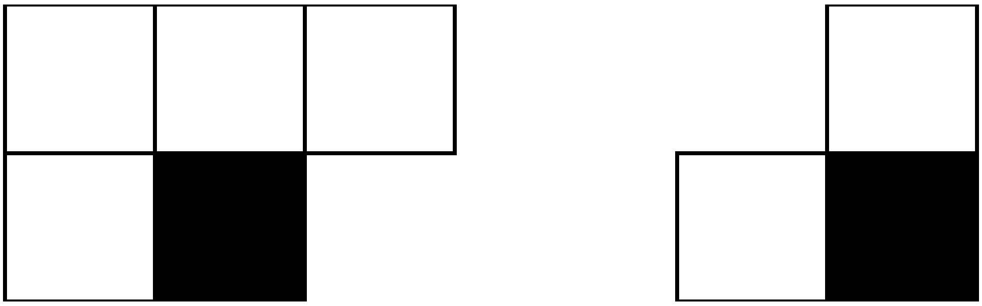


Figure 4.7: Illustration of connected components analysis. The black pixel signifies a border, as its previously evaluated neighbours are all white. The neighbours considered for this can include all neighbours (left) or only neighbours along the horizontal and vertical axes (right).

Figure 4.8: Input image (left) and connected skin areas (right).
Input image source: [HMDB].

4.3 Smoke Detection

Having identified face and hand regions in the frame, the system searches for smoke in these areas. Three methods are used to evaluate the presence of smoke:

- Motion detection
- Colour analysis
- Disorder analysis

4.3.1 Motion Detection

Smoke constantly diffuses in the air, making motion a key indicator of its presence. While motion does not uniquely distinguish smoke from other objects, motion detection can quickly mark potential areas of smoke for more rigorous analysis [Kaabi, 2017].

This project uses OpenCV's implementation of the Gaussian Mixture Model (GMM) [Stauffer, 1999] to carry out motion analysis. Every pixel in the background of the video is

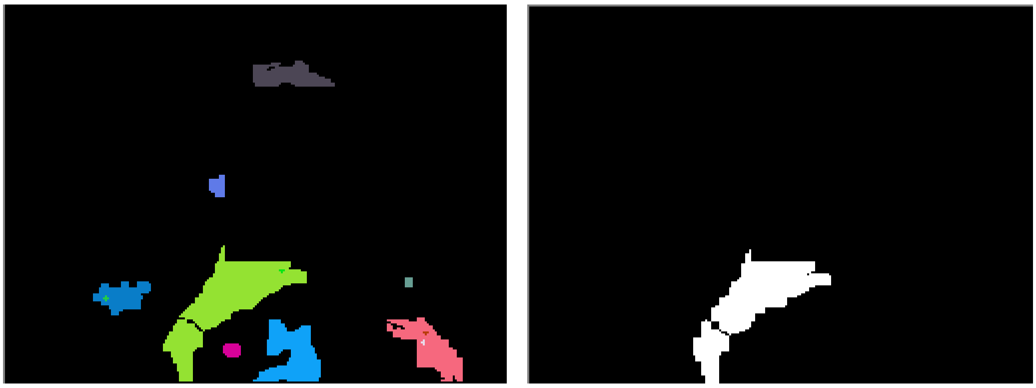


Figure 4.9: Connected areas with face region masked out (left) and chosen hand (right).

modelled by a number of Gaussian distributions. Every frame, each pixel is matched to the closest of the Gaussian distributions modelling that point. If none of the distributions are sufficiently close, a new distribution is created for that point. Each pixel is then designated as foreground or background based on the distribution that it is currently matched to.

This procedure is applied to each frame, with the GMM background model updating over time. The output is a binary image, with all pixels perceived to be moving marked in white (Figure 4.10).

Figure 4.10: Input image (left) and areas of movement (right).
Input image source: [HMDB].

4.3.2 Colour Analysis

The colour of smoke varies, depending on its temperature, but generally ranges from bluish-white to grey. These ranges of colour are another key indicator of smoke. Moving pixels that fall outside of the common colour ranges for smoke can be safely discarded.

To verify that pixels in each frame are smoke-coloured, this project adapts an algorithm by

F. Yuan [Yuan, 2008]. This algorithm is based on three conditions:

$$|C_{max} - C_{min}| < T_{grey} \tag{3}$$

$$C_{Max} = B \text{ AND } |C_{Max} - C_{Min}| < T_{\text{blue}} \tag{4}$$

$$T_{\text{min intensity}} < I < T_{\text{max intensity}} \tag{5}$$

R, G, B = RGB colour components
$C_{min} = min(R, G, B)$; $C_{max} = max(R, G, B)$
$I = \frac{R+G+B}{3}$
$T_{grey/min\ intensity/max\ intensity/blue}$ = *Empirical thresholds*

(3) reflects the grey colour of smoke. The RGB components of any pixel of this colour will be within a small range of each other. This maximum range is specified by T_{grey}.

(4) accounts for the possibility that smoke may be closer to white tinged with blue, rather than grey. If a smoke pixel's predominant component is blue, there will be a larger range between its maximum and minimum RGB values. This larger range is specified by T_{blue}.

(5) reflects the relatively low intensity/saturation of smoke. The intensity of a smoke pixel should fall within a particular range, specified by $T_{min\ intensity}$ and $T_{max\ intensity}$.

For a pixel to be classified as a smoke pixel, it must satisfy (5) and either (3) or (2).

The input image passed to this stage of analysis is the current frame, with everything outside of the designated search areas (face and hand regions) masked out. The output of this region is combined with the output of the movement analysis stage in a bitwise AND operation, making the final smoke pixels those that are both moving and within the appropriate ranges of colour (Figure 4.11).

As demonstrated by Figure 4.11, the final search areas are expanded to be slightly larger than the identified head and hand areas. This is to allow for the possibility of smoke appearing slightly outside of the face and hand regions (Figure 4.12).

4.3.3 Disorder Analysis

Disorder analysis is the final, dynamic stage of smoke detection. Using an algorithm by Chen et al. [Chen, 2004], this step helps to distinguish between actual smoke and moving, smoke-coloured objects.

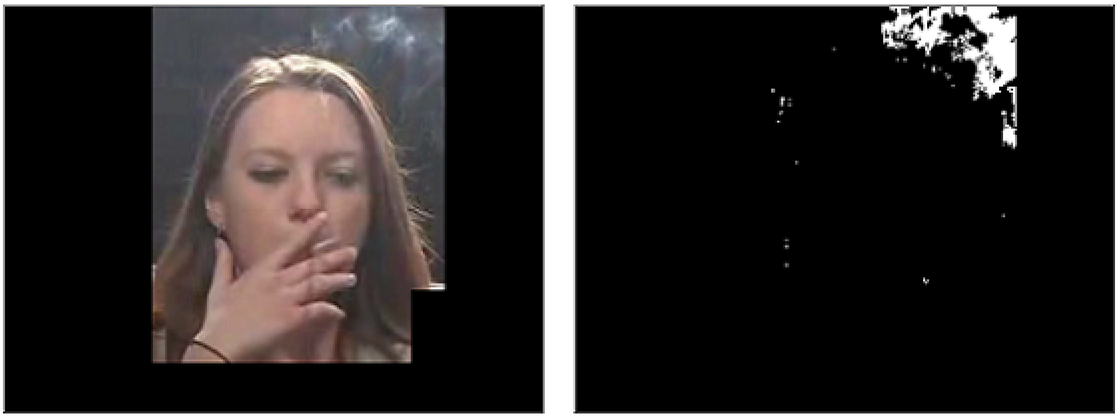


Figure 4.11: Input image with areas to search for smoke (left) and smoke pixels found in these areas (right). Input image source: [HMDB].

Figure 4.12: Image illustrating the need for search areas to be slightly larger than the face and hand. Some of the smoke in this image falls just outside of the face and hand regions. Image source: [HMDB].

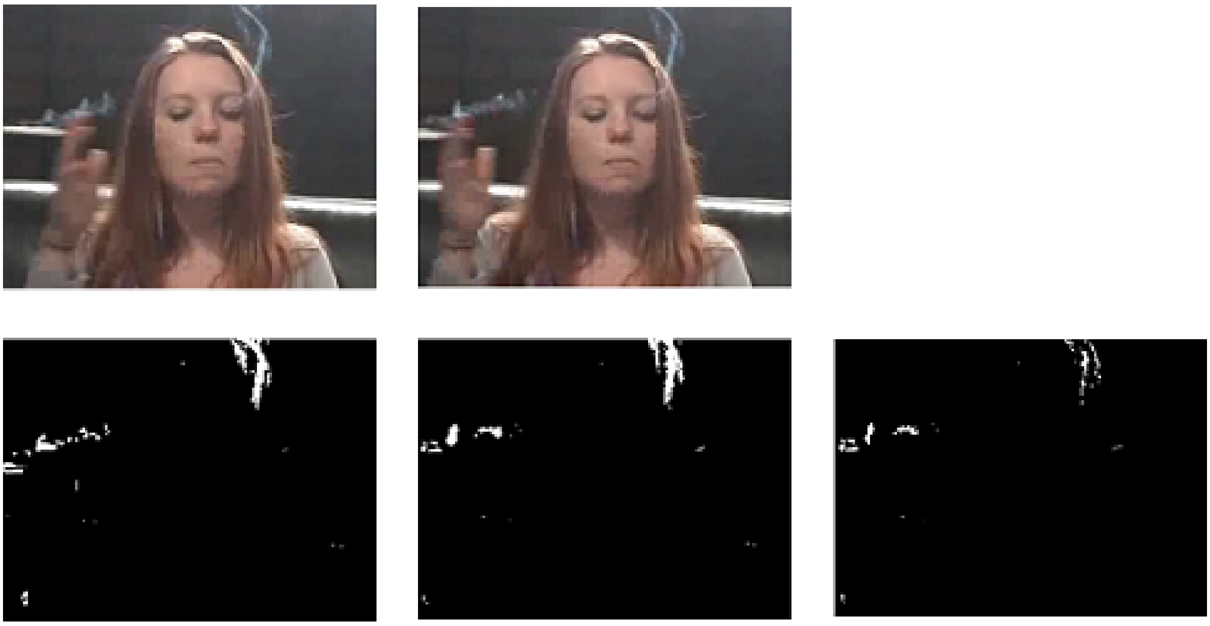
Due to the constant, chaotic movement of smoke, any smoke region should shift and grow continuously between frames. The difference in a smoke region's shape and size between consecutive frames indicates the disorder of the region. Large changes in disorder between consecutive frames are a sign of smoke.

The disorder between two consecutive frames is given by subtracting one's smoke pixels from the other. The quantitative measure of disorder is the number of smoke pixels the images differ by (Figure 4.13). In other words, this is the number of smoke pixels that are in one frame, but not the other:

$$SD_t = S_t(x, y) - S_{t\text{-}1}(x, y) \tag{6}$$

SD_t = Disorder in current frame
$S_t(x, y)$ = Smoke image obtained for current frame

$S_{t-1}(x, y)$ = *Smoke image obtained for previous frame*



(a) Previous frame (b) Current frame (c) Disorder between frames

Figure 4.13: The disorder between the previous and current frames (c) is obtained by subtracting the previous frame's smoke image (a) from the current frame's smoke image (b). Input images source: [HMDB].

To determine if a frame is a smoke frame, the level of disorder between it and the frame before and after it must be obtained. The frame is classified as a smoking frame if (7) holds true:

$$\frac{|SD_{t+1} - SD_t|}{SD_t} \geqslant T_{SD} \tag{7}$$

SD_{t+1} = *Disorder between current frame and next frame*
SD_t = *Disorder between current frame and previous frame*
T_{SD} = *An empirical threshold*

4.4 Final Classification

After the system has classified each frame in the video as either smoking or non-smoking, the system must make its final decision for whether or not the source video contains an instance of smoking.

The designation of a small number of frames as smoking frames could be due to false alarms, so the system will only classify the source video as smoking if a relatively high proportion of frames were designated as smoking frames. Since the videos the system is designed to handle are rather short (generally less than 10 seconds), it is assumed that any instance of smoking will occur over a significant number of frames. Thus, if the proportion of smoking frames exceeds a certain threshold, the video is classified as depicting smoking.

5 Evaluation

This chapter gives an overview of the performance of the system and some of the issues encountered.

5.1 Source Datasets

To evaluate the system performance, this project used a selection of videos from the Bilkent [Bilkent] and HMDB [HMDB] datasets. The selected videos are intended to be a representative sample of modern film footage: many (though not all) are from modern films. For a full list of the videos used, see Appendix A1.

5.2 Issues Encountered

The key issue encountered was that, even within the relatively small sample set used, there is a very large amount of variation: differences in lighting, backgrounds, the colour of smoke, etc. This makes it very difficult to properly calibrate the system for optimal results. Optimising the system to accurately handle one particular video runs the risk of making the system unable to properly process others.

Additionally, the nature of cinematography is such that a lot of footage in modern films is not optimised for computer vision systems. A scene may be very dark, the people within may be turned away from the camera, the background may be quite visually complex, and so on. This contributes to the difficulty with optimising a system of this nature (i.e. one intended for films): the range of inputs is both very broad and potentially difficult to analyse.

It was found through experimentation that, in general, the system can be optimised for high detection rates of positive samples, or high discrimination rates for negative samples. Relaxing certain parameters ensures that most instances of smoking will be detected, but many false positives will be mistaken for smoke as well. Placing greater restrictions ensures these false positives are eliminated, but this causes many positive samples to be missed.

5.2.1 Face Detection

The face classifier tends to miss faces when not all of the facial features are visible. This can be due to a face being turned away from the camera (Figure 5.1) or partially obscured by hands, hair or very tight framing (Figure 5.2). The face classifier also struggles with poor lighting: brightly lit faces are easier to detect.

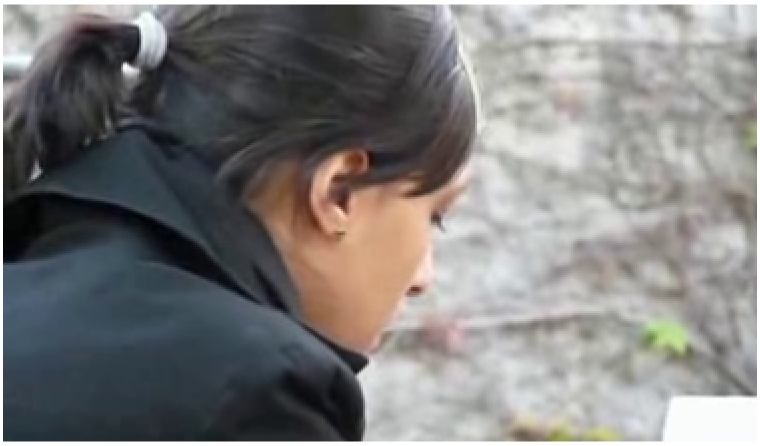


Figure 5.1: Face is turned away, obscuring relevant features from the face classifier. Image source: [HMDB].

Figure 5.2: Face is partially obscured by hand, and the top of the head is cut off by the edge of the frame. Image source: [HMDB].

The face classifier can be adjusted to include more faces by lowering the minimum number of neighbours each face region requires. While this allows a great number of faces that would otherwise have been missed to be included, it introduces many false positives.

5.2.2 Hand Detection

The success of the back-projection stage depends heavily on the accuracy of the face detector. If the identified face region is incorrect, the back-projection will be as well. In general, provided the face detection is accurate, back-projection works reasonably well. That said, uneven lighting and shadows can affect its accuracy. In addition, the skin colour sample can sometimes match background objects as well (Figure 5.3).

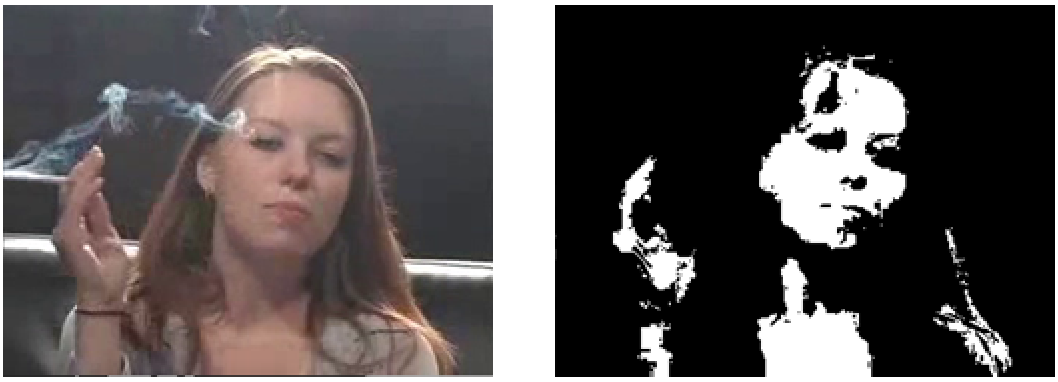


Figure 5.3: Input image (left) and thresholded output of back-projection stage (right). The back-projection misses certain areas of skin that are in shadow, and incorrectly includes some hair. Input image source: [HMDB].

The colour analysis stage generally works well, but suffers from the same issues as the back-projection stage. In particular, the colour analysis is greatly effected by lighting. Brightly-lit areas of skin are easily detected, but shadowed areas of skin are often mistakenly excluded.

As mentioned in Section 3.2, the primary issue with the algorithm for selecting a hand region is that it assumes both that a hand is actually present, and is the largest contour outside of the face region. Ideally, if no hand is present in the frame, any identified skin areas will be of negligible size. However, this is frequently not the case, due to the introduction of false positives by the back-projection and colour analysis stages. Additionally, if the person on-screen has other areas of skin visible, the hand is not guaranteed to be the largest connected skin region. While this method of hand detection does work, it is prone to failures.

5.2.3 Smoke Detection

In general, the motion detection stage works extremely well. The GMM implementation is able to detect movement with extremely high accuracy, so virtually all smoke pixels are retained by this stage. However, as mentioned in Section 3.2, the use of GMM limits the project to footage captured using a stationary camera, as GMM is unable to handle changes in the background. This issue also occurs whenever there is a camera cut (i.e. the view switches to another camera, in a different position). As everything in the frame changes, GMM marks everything as moving for a small number of frames before the background model properly adjusts. Given that camera cuts don't occur very frequently, and the issue rectifies itself quite quickly, this is a relatively minor issue: the false positives it introduces are negligible by comparison.

The colour analysis stage is where the majority of failures are introduced. From empirical observation, the range of colours that constitute smoke is quite broad, and varies significantly from video to video. Increasing the thresholds used for this stage to allow for all of the possible variations in colour results in very large portions of the frame being misclassified as smoke. Areas of skin are particularly prone to being misidentified in this manner, if the thresholds are made too large. Restricting the threshold ranges to eliminate these false positives runs the risk of omitting smoke of certain colours entirely. This stage is, by far, the most prone to errors. The range of positive colours is too large to accommodate without including various similarly-coloured objects in the scene.

The disorder analysis stage suffers due to issues from the previous stages of analysis. Since the detection of smoke pixels isn't entirely reliable (due to false positives and the area being searched changing over time), it is difficult to set an appropriate minimum threshold for disorder. Hence, the disorder analysis stage is limited in how effectively it can distinguish between actual areas of smoke and smoke aliases.

5.3 Results

As mentioned, it is possible to optimise the system for either high positive detection rates or high negative discrimination rates. To demonstrate this, five runs of the system were performed with different values for the following parameters:

- Minimum neighbours needed for detected face regions to be retained
- Colour thresholds for smoke: $T_{\text{grey/min intensity/max intensity/blue}}$
- Minimum disorder threshold for smoke: T_{SD}
- Minimum proportion of smoking frames needed for a video to be classed as depicting smoking

The full set of results is given in Table 5.1. See Figure 5.4 for an accompanying graph. (For a more detailed breakdown of these results, see Appendix A2.)

As these results illustrate, it is possible to correctly classify either many positive or negative samples, at the expense of misclassifying a large proportion of the other sample set.

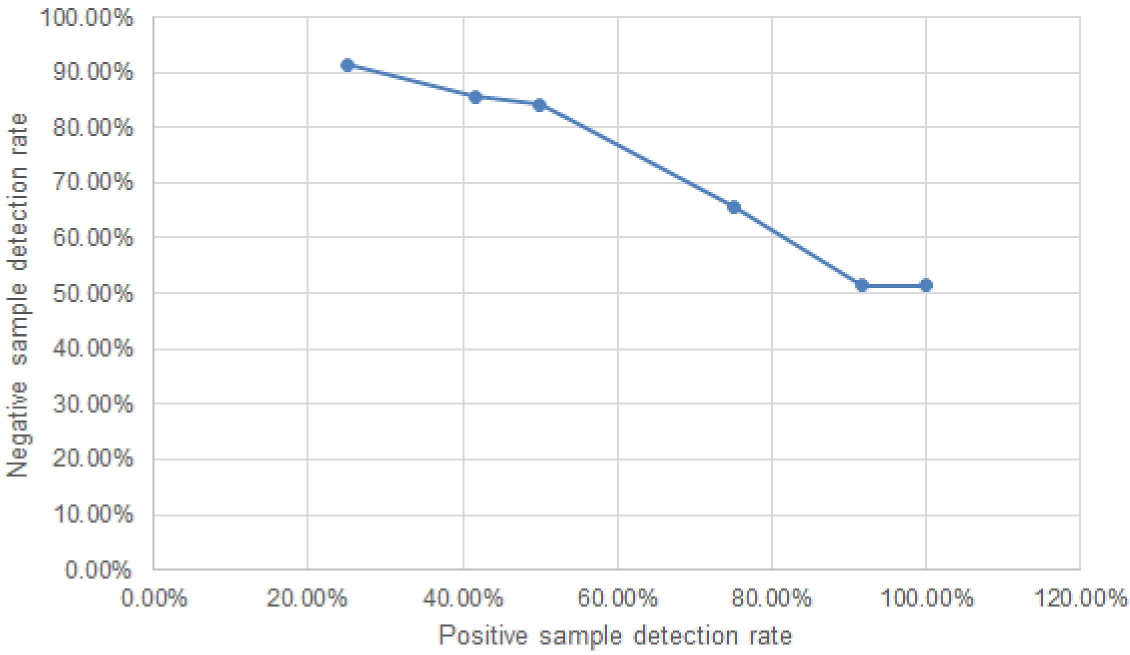


Figure 5.4: Graph of negative vs positive sample detection rates over all five runs.

Table 5.1: System results over six runs, with different parameters each time.

Test	True Positives	False Positives	True Negatives	False Negatives
1	3	3	32	9
2	5	5	30	7
3	6	6	32	6
4	9	12	23	3
5	11	17	18	1
6	12	17	18	0

6 Conclusion

The aim of this project was to develop an automated system that can determine whether smoking occurs in a video sequence.

The system was designed to search for smoke in the area around people's faces and hands in each frame, and to use this as an indicator of smoking.

The final results were mixed, in part due to the variety and complexity present in typical film footage. Accommodating the wide range of possible positive samples will introduce a significant number of false positives to the system. Thus, the system's parameters can be altered to give it either a higher rate of positive or negative detection.

The final chapter should give a short summary of the key methods, results and findings in your project. You should also briefly identify what, if any, future work might be executed to resolve unanswered questions or to advance the study beyond the scope that you identified in Chapter 1.

6.1 Future Work

There is room for improvement in the given results, which could be a potential direction for future research. As many of the errors in results are introduced in the smoke detection stage, the application of more advanced smoke detection methods may serve to mitigate some of these.

As mentioned previously, a major limitation of the system is that its use of GMM to identify movement limits it to footage shot with a stationary camera. Converting the system to use a motion detection algorithm that can handle changes in the background would make the system usable with far more footage.

Additionally, extending the system to work on longer sequences of footage (such as feature-length films) would require changes to the final classification step. In longer videos, smoking may still occur, but over a relatively small number of frames. To accommodate this, it may be wise for the system to split long videos into short sections, and check for smoking in each section on an individual basis.

Bibliography

[Bilkent] E. Cetin. "Sample Fire and Smoke Video Clips". In: URL: http://signal.ee.bilkent.edu.tr/VisiFire/Demo/SmokeClips/.

[Brovko, 2013] N. Brovko, R. Bogush, and Sergey Ablameyko. "Smoke detection algorithm for intelligent video surveillance system." In: *The Computer Science Journal of Moldova* 21.1 (2013), pp. 142–156. URL: http://dblp.uni-trier.de/db/journals/csjm/csjm21.html#BrovkoBA13.

[Calderara, 2008] Simone Calderara, Paolo Piccinini, and Rita Cucchiara. "Smoke Detection in Video Surveillance: A MoG Model in the Wavelet Domain". In: *ICVS*. Vol. 5008. 2008, pp. 119–128.

[Chen, 2004] Thou-Ho Chen, Ping-Hsueh Wu, and Yung-Chuen Chiou. "An early fire-detection method based on image processing". In: *Image Processing, 2004. ICIP '04. 2004 International Conference on*. Vol. 3. Oct. 2004, 1707–1710 Vol. 3. DOI: 10.1109/ICIP.2004.1421401.

[Cheng, 2001] H.D. Cheng et al. "Color image segmentation: advances and prospects". In: *Pattern Recognition* 34.12 (2001), pp. 2259–2281. ISSN: 0031-3203. DOI: https://doi.org/10.1016/S0031-3203(00)00149-7. URL: http://www.sciencedirect.com/science/article/pii/S0031320300001497.

[HMDB] H. Kuehne et al. "HMDB: a large video database for human motion recognition". In: *Proceedings of the International Conference on Computer Vision (ICCV)*. 2011. URL: http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/.

[Hsieh, 2014] Jun-Wei Hsieh et al. "Handheld object detection and its related event analysis using ratio histogram and mixture of HMMs". In: 25 (Aug. 2014), pp. 1399–1415.

[Hsu, 2002] Rein-Lien Hsu, M. Abdel-Mottaleb, and A. K. Jain. "Face detection in color images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.5 (May 2002), pp. 696–706. ISSN: 0162-8828. DOI: 10.1109/34.1000242.

[Kaabi, 2017] R. Kaabi et al. "Video smoke detection review: State of the art of smoke detection in visible and IR range". In: *2017 International Conference on Smart, Monitored and Controlled Cities (SM2C)*. Feb. 2017, pp. 81–86. DOI: 10.1109/SM2C.2017.8071823.

[Kim, 2005] Kyungnam Kim et al. "Real-time foreground–background segmentation using codebook model". In: *Real-Time Imaging* 11.3 (2005). Special Issue on Video Object Processing, pp. 172–185. ISSN: 1077-2014. DOI: https://doi.org/10.1016/j.rti.2004.12.004. URL: http://www.sciencedirect.com/science/article/pii/S1077201405000057.

[Kolkur, 2017] S. Kolkur et al. "Human Skin Detection Using RGB, HSV and YCbCr Color Models". In: *ArXiv e-prints* (Aug. 2017). arXiv: 1708.02694 [cs.CV].

[Stauffer, 1999] C. Stauffer and W. E. L. Grimson. "Adaptive background mixture models for real-time tracking". In: *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*. Vol. 2. 1999, 252 Vol. 2. DOI: 10.1109/CVPR.1999.784637.

[Surgeon General, 2012] U.S. Department of Health and Human Services. *Preventing Tobacco Use Among Youth and Young Adults: A Report of the Surgeon General*. Tech. rep. U.S. Department of Health et al., 2012.

[Suzuki, 1985] Satoshi Suzuki and Keiichi Abe. "Topological structural analysis of digitized binary images by border following". In: *Computer Vision, Graphics, and Image Processing* 30 (1985), pp. 32–46.

[Swain, 1990] M. J. Swain and D. H. Ballard. "Indexing via color histograms". In: *[1990] Proceedings Third International Conference on Computer Vision*. Dec. 1990, pp. 390–393. DOI: 10.1109/ICCV.1990.139558.

[Toreyin, 2006] B. U. Toreyin, Y. Dedeoglu, and A. E. Cetin. "Contour based smoke detection in video using wavelets". In: *2006 14th European Signal Processing Conference*. Sept. 2006, pp. 1–5.

[Tynan, 2017] Michael A. Tynan et al. "Tobacco Use in Top-Grossing Movies — United States, 2010–2016". In: 66 (July 2017), pp. 681–686.

[Viola, 2001] Paul Viola and Michael Jones. "Robust Real-time Object Detection". In: *International Journal of Computer Vision*. 2001.

[WHO, 2015] World Health Organization. *Smoke-free movies: from evidence to action, 3rd ed.* Tech. rep. 2015. URL: `http://www.who.int/iris/handle/10665/190165`.

[Wu, 2010] P. Wu et al. "Human Smoking Event Detection Using Visual Interaction Clues". In: *2010 20th International Conference on Pattern Recognition*. Aug. 2010, pp. 4344–4347. DOI: `10.1109/ICPR.2010.1056`.

[Wu, 2011] W. C. Wu and C. Y. Chen. "Detection System of Smoking Behavior Based on Face Analysis". In: *2011 Fifth International Conference on Genetic and Evolutionary Computing*. Aug. 2011, pp. 184–187. DOI: `10.1109/ICGEC.2011.51`.

[Yuan, 2008] Feiniu Yuan. "A fast accumulative motion orientation model based on integral image for video smoke detection". In: 29 (May 2008), pp. 925–932.

A1 Videos Used for Testing

All videos used in testing are from either the Bilkent [Bilkent] or HMDB [HMDB] datasets. The videos are listed below, with their sources and original names.

Smoking videos: [HMDB]

1. smoke/A_Beautiful_Mind_2_smoke_u_cm_np1_fr_goo_0.avi
2. smoke/Buckle_Up_smoke_h_cm_np1_ri_goo_0.avi
3. smoke/girl_smoking_a_cigarette_smoke_h_nm_np1_fr_med_2.avi
4. smoke/american_history_x_smoke_h_nm_np1_fr_goo_26.avi
5. smoke/glory_smoke_h_nm_np1_fr_goo_41.avi
6. smoke/The_Fugitive_2_smoke_h_nm_np1_fr_med_23.avi
7. smoke/Veoh_Alpha_Dog_1_smoke_h_nm_np1_fr_goo_4.avi
8. smoke/Veoh_Alpha_Dog_2_smoke_h_cm_np1_fr_med_4.avi
9. smoke/smoking_2_smoke_h_cm_np1_ri_med_0.avi
10. smoke/smoking_3_smoke_h_cm_np1_fr_med_1.avi
11. smoke/smoking_smoke_h_nm_np1_fr_med_0.avi
12. smoke/raucher_smoke_u_nm_np1_fr_med_2.avi

Non-Smoking videos (smoke, fire, cars): [Bilkent]

1. CarLights1.avi
2. CarLights2.avi
3. sEmptyR1.avi
4. sEmptyR2.avi

5. sBehindtheFence.avi
6. sWasteBasket.avi
7. ShorterIsyamNight.avi

Non-Smoking videos (people near smoke):

1. sWindow.avi [Bilkent]
2. shoot_gun/Pirates_5_shoot_gun_h_nm_np1_le_goo_2.avi [HMDB]
3. Pirates_5_shoot_gun_u_nm_np1_fr_goo_4.avi [HMDB]
4. The_Matrix_5_shoot_gun_u_cm_np1_fr_goo_2.avi [HMDB]

Non-Smoking videos (people, no smoke): [HMDB]

1. walk/AboutABoy_walk_f_nm_np1_le_med_6.avi
2. walk/21_walk_h_cm_np1_fr_med_10.avi
3. walk/20060723sfjffprofessionalhelp_walk_u_nm_np2_le_med_0.avi
4. talk/American_History_X_talk_h_nm_np1_fr_goo_20.avi
5. talk/Italian_Job_2_talk_h_nm_np1_fr_med_6.avi
6. talk/Italian_Job_3_talk_h_nm_np1_ri_med_0.avi
7. talk/Pirates_1_talk_h_nm_np1_fr_goo_2.avi
8. walk/21_walk_u_cm_np1_fr_med_11.avi
9. walk/AllThePresidentMen_walk_u_nm_np1_le_med_0.avi
10. walk/AmericanGangster_walk_f_nm_np1_ba_med_30.avi
11. walk/AmericanGangster_walk_f_nm_np1_fr_bad_41.avi
12. walk/AmericanGangster_walk_f_nm_np1_fr_med_45.avi
13. drink/310ToYuma_drink_h_nm_np1_fr_goo_2.avi
14. drink/Hitch_Part_1_drink_h_nm_np1_le_goo_5.avi
15. eat/IamLegendII_eat_h_nm_np1_le_goo_8.avi

16. eat/Finding_Forrester_3_eat_h_nm_np1_fr_goo_14.avi
17. sit/AMADEUS_sit_u_nm_np1_fr_med_4.avi
18. sit/Sixthsense_sit_u_cm_np1_fr_med_3.avi
19. talk/Faith_Rewarded_talk_u_nm_np1_fr_goo_8.avi
20. talk/Italian_Job_1_talk_h_nm_np1_fr_goo_2.avi
21. talk/Pirates_6_talk_h_nm_np1_fr_med_1.avi
22. talk/American_History_X_talk_u_nm_np1_fr_goo_21.avi
23. clap/boom-snap-clap_clap_u_nm_np1_fr_med_0.avi
24. clap/Kurt_Kr_mer_-_Klatschen_im_Flugzeug_clap_f_nm_np1_fr_med_0.avi

A2 Extended Results

As mentioned in section 5.3, six runs of the system were performed to obtain six sets of results. Each run had slightly different values for:

1. The minimum number of neighbours a face region needs to have before it is recognised by the face classifier
2. The four thresholds used in colour analysis for smoke: T_{grey}, $T_{\text{min intensity}}$, $T_{\text{max intensity}}$ and T_{blue}
3. The minimum disorder threshold used in the disorder analysis stage for smoke, T_{SD}
4. The minimum proportion of frames that must be classified as smoking frames for the system to label a video as containing a depiction of smoking.

The parameters used in each run are listed in table A2.1. An list of the results from each run is included in tables A2.2 and A2.3.

Table A2.1: System parameters for each run.

Parameter	**1**	**2**	**3**	**4**	**5**	**6**
Minimum neighbours for face detection	5	3	3	3	3	3
T_{grey}	40	40	20	40	40	40
$T_{\text{min intensity}}$	40	40	80	40	40	40
$T_{\text{max intensity}}$	150	150	130	150	150	150
T_{blue}	75	75	75	75	75	75
T_{SD}	0	0	0	0	0.25	0
Minimum proportion	0.375	0.5	0.25	0.375	0.25	0.25

Table A2.2: Extended results for all six runs.

Video	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Truth
A_Beautiful...avi	F	F	F	T	F	T	T
Buckle-Up...avi	F	T	T	T	T	T	T
girl_smoking...avi	T	T	T	T	T	T	T
american_history...avi	F	F	F	T	T	T	T
glory_smoke...avi	F	F	F	T	T	T	T
the_fugitive...avi	F	T	T	T	T	T	T
veoh_...1...avi	F	F	F	F	T	T	T
veoh_...2...avi	T	T	T	T	T	T	T
smoking_2...avi	F	F	T	T	T	T	T
smoking_3...avi	T	T	T	T	T	T	T
smoking_smoke...avi	F	F	F	T	T	T	T
raucher_smoke...avi	F	F	F	F	T	T	T
CarLights1.avi	F	F	F	F	T	T	F
CarLights2.avi	F	F	F	F	T	T	F
sEmptyR1.avi	F	F	F	F	T	T	F
sEmptyR2.avi	F	F	F	F	F	F	F
sBehindtheFence.avi	F	F	F	F	F	F	F
sWasteBasket.avi	F	F	F	F	F	F	F
ShorterIsyamNight.avi	F	F	F	F	F	F	F
sWindow.avi	F	F	F	F	F	F	F
Pirates_5...2.avi	F	F	F	F	F	F	F
Pirates_5...4.avi	F	F	F	F	F	F	F
The_Matrix...avi	F	F	F	T	T	T	F
AboutABoy_walk...avi	F	F	F	F	F	F	F
21_walk...avi	F	F	F	F	F	F	F
20060723sfjffprofessional...avi	F	F	F	F	T	T	F
American_...talk...avi	F	T	T	T	T	T	F
Italian_Job_2...avi	F	F	F	F	T	T	F
Italian_Job_3...avi	T	T	T	T	T	T	F
Pirates_1...2.avi	F	F	F	F	F	F	F
21_walk...avi	F	F	F	T	T	T	F
AllThePresidentMen....avi	F	F	F	F	F	F	F
AmericanGangster...30.avi	F	T	F	T	T	T	F
AmericanGangster...41.avi	F	F	F	F	F	F	F
AmericanGangster...45.avi	F	F	F	T	T	T	F
310ToYuma_...avi	F	F	F	T	T	T	F
Hitch...avi	F	F	T	T	T	T	F
IamLegendII...avi	F	F	T	T	T	T	F
Finding_Forrester...avi	T	T	F	T	T	T	F

Table A2.3: Extended results for all six runs.

Video	**Run 1**	**Run 2**	**Run 3**	**Run 4**	**Run 5**	**Run 6**	**Truth**
AMADEUS...avi	F	F	F	F	F	F	F
Sixthsense...avi	F	F	F	F	F	F	F
Faith_Rewarded...avi	F	F	F	F	F	F	F
Italian_Job_1...avi	T	T	T	T	T	T	F
Pirates_6...1.avi	F	F	F	F	F	F	F
American...talk...21.avi	F	F	F	T	T	F	F
boom-snap-clap...avi	F	F	F	F	F	F	F
Kurt_Kr...avi	F	F	F	F	F	F	F