TAMPERE UNIVERSITY OF TECHNOLOGY

UGUR KART

Image Classification in Fashion Domain

Master of Science Thesis

Examiners and language approved in the Faculty Council of CEE on 5 March 2014

Examiners: Prof. Moncef Gabbouj and Prof. Serkan Kiranyaz

PREFACE

This Thesis has been carried out in the Department of Signal Processing, Tampere University of Technology as a part of EIT-SMax project.

First of all, I would like to express my gratitude to my supervisors Academy Professor Moncef Gabbouj and Professor Serkan Kiranyaz for providing me the opportunity to work in this project and their helpful guidance. Additionally, my sincere thanks to Dr. Esin Guldogan for her supervision in my previous projects. I also wish to thank Ezgi Can Ozan and Stefan Uhlmann for our fruitful discussions. For their friendly and helpful working environment, I would like to thank all MUVIS Team members as well.

Finally I want to thank my dear parents Mehmet and Perihan for their endless support and their love, my brothers and sister-in-law; Kenan, Fatih and Jülide for always being there when I needed them.

Tampere, May 2014

Ugur Kart

Insinöörinkatu 60 D, 373

33720 Tampere, FINLAND

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY

Department of Signal Processing

Kart, Ugur: Image Classification in Fashion Domain

Master of Science Thesis, 68 pages

Examiners: Academy Prof. Moncef Gabbouj and Prof. Serkan Kiranyaz

Funding: European Institute of Innovation and Technology (EIT), Project S-MAX – Social Multimedia Annotations & eXperiences (2012: Event Centric Multimedia Content)

May 2014

Keywords: Image classification, fashion, machine learning, pattern recognition, feature extraction.

Thanks to the wide spread usage of mobile devices with imaging capabilities, the number of digital images has been increasing exponentially. Every year, billions of images are taken worldwide. However, as the number increases, managing and arranging the pictures manually also becomes infeasible. Image classification is an elegant solution for this problem where one can define categories according to his/her needs thus, the images can be automatically separated and put into the categories. The flexible nature of image classification can make it possible to use it in different domains. In this thesis, it is shown that image classification can be also used in a specific field of fashion domain where the categories become more abstract. Three different approaches for solving the four categories image classification problem are proposed where the categories are *"Informal Woman, Formal Woman, Informal Man and Formal Man"*. Formality of a cloth is defined as the Western type of evening dresses where the dark colors with homogenous color distributions dominate the formal cloths and lighter colors with heterogeneous distributions.

In order to bridge the gap between the pixels and the meaningful concepts such as color, shape and texture, four different low-level feature descriptors are used. For the shape, Histogram of Oriented Gradients is used whereas Local Binary Patterns is adopted for the texture description. The color distributions are described by using Color Structure Descriptor and Color Layout Descriptor. As the building block for the proposed classification topology, Support Vector Machines are used.

The proposed approaches differ from each other in the way the features are extracted. Features are extracted globally in the Global Approach. In the second and third approaches, the images are divided into nine non-overlapping grids and the features are extracted separately from each of the individual grids. After the extraction process, they are concatenated into a single vector to represent the whole image. In the Seven Grids Feature Extraction Approach, the grids next to the head are excluded in order to further reduce the noise present.

Once the features are extracted and concatenated, they are fed into a two-level classification scheme in which the male-female separation is followed by the formality check for the cloth.

The experimental results show that Seven Grids Feature Extraction Approach outperforms the other two approaches yet the Global Approach shows close results.

Table of Contents

List of Tables

List of Figures

Abbreviations

AMT	Amazon Mechanical Turk
API	Application Programming Interface
BSD	Berkeley Software Distribution
CCD	Charge-Coupled Device
CLD	Color Layout Descriptor
CPU	Central Processing Unit
CSD	Color Structure Descriptor
DCT	Discrete Cosine Transform
GLBP	Gradient Local Binary Patterns
GNU	GNU's Not Unix
GPL	General Public License
GPU	Graphics Process Unit
HOG	Histogram of Oriented Gradients
JPEG	Joint Photographic Experts
JVM	Java Virtual Machine
LBP	Local Binary Patterns
LBP-TOP	Volume Local Binary Patterns
LibSVM	A Library for Support Vector Machines
MPEG	Moving Picture Experts Group
OpenCL	Open Computing Language
OpenCV	Open Source Computer Vision
RAM	Random Access Memory
SIFT	Scale-Invariant Feature Transform
SVM	Support Vector Machine
Weka	Waikato Environment for Knowledge Analysis

1 Introduction

The famous quotation from *Arthur Brisbane* says; *"Use a picture. It's worth a thousand words."* [1] It is indeed true because of the fact that the visual perception is one of the most important tools which humans possess in order to observe the world we live in. Thanks to the millions of years of evolution, we have developed a perfect way to obtain environmental information that scientists still try to imitate [2]. The reason why the visual information is so crucial stems from the fact that one can get a precise description of what s/he is surrounded with. Using this advantage, we have been able to protect ourselves from predators, gather food, and find shelter. Consequently, it has become our primary source of information.

Throughout the human history, there have been many attempts to project and store the visual information onto different mediums by numerous cultures around the world such as Chinese, Ancient Greeks, Arabs, English, Italians [3]. *Camera Obscura (Latin; "darkened chamber")* can be considered as the most primitive camera system in which one would use a simple device in order to project what is observed by it onto a surface. The basic structure is a box which has a pinhole for letting the light from outside to inside as it can be seen in Figure 1.1. Even though the resulting image is rotated 180 degrees, the color and the perspective are perfectly preserved. This projection can be easily traced onto a medium such as paper, thus a highly detailed representation of this view can be created.

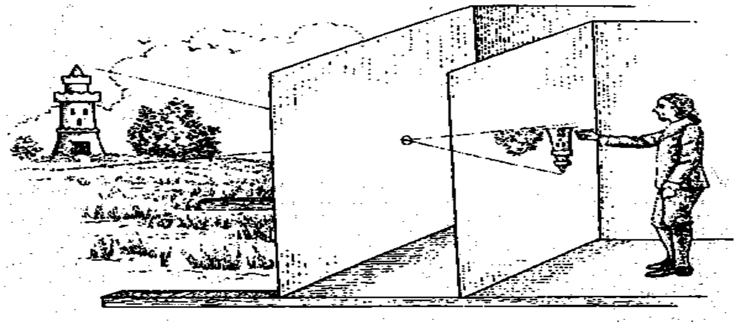


Figure 1.1 - Camera Obscura

Courtesy: http://people.wcsu.edu/mccarneyh/acad/camera_obscura.gif

The person who first had the idea of capturing images by using light-sensitive material was the Englishman *Thomas Wedgewood* in 1800s [4]. However, he was not able to protect his works from the effects of the sunlight which made the pictures fade eventually.

The world's first surviving picture was taken by the French inventor *Joseph Nicéphore Niépce* in 1826 or 1827 by using a polished pewter plate [5] [6]. He used bitumen as the light-sensitive material and more than eight hours of exposure time which is extremely long comparing with the modern cameras. This procedure brings hardened parts proportionally with the amount of sunlight they were exposed to. After that, the softer parts can be wiped off using suitable chemicals in order to reveal the actual picture. The result of this lengthy process is given in Figure 1.2.

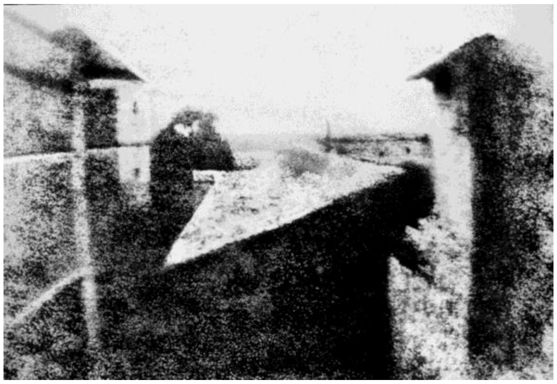


Figure 1.2 - The world's first surviving picture

Courtesy: http://sugarquotes.weebly.com/uploads/5/2/1/4/5214858/4999091_orig.jpg

In 1861, *Thomas Sutton* [7] achieved another leap by creating the first color photograph in which he used three color filters. In spite of the fact that his demonstration was the first successful one as shown in Figure 1.3, the commercial application of color photography was introduced in 1907 by using the ideas of *Louis Ducos du Hauron* [8]. The difference between *Sutton's* and *du Hauron's* methods is; *Sutton* took three black-and-white pictures using three different filters which are red, green and blue. On the other hand, *du Hauron* used a mosaic of color filters, which have small filter elements. This provided blending of the colors in the eye.

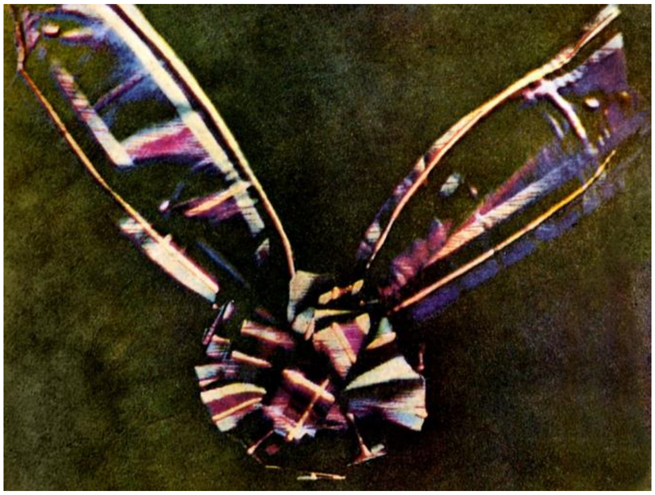


Figure 1.3 - The world's first color picture taken by Thomas Sutton in 1861

Courtesy: http://images.nationalgeographic.com/wpf/media-live/photos/000/013/cache/color-tartan-ribbon_1376_990x742.jpg

1.1 History of Digital Imaging

The very first application of the digital image processing dates back to 1920s when Bartlane cable transmission system was used for transmitting images between London and New York [9] It provided 56 times faster data transmission rate across the Atlantic and was able to quantize pictures in 5 gray scale levels, however, this changed in 1929 when the number of levels increased three fold to 15 which made possible to transmit better quality pictures as explained with more details in [10],

The invention of the *Charge-Coupled Device (CCD)* by *Willard Boyle* and *George E. Smith* in 1969 [11] was one of the major breakthroughs in digital imaging technology where the idea is measuring the amount of light on a specific part of a sensor.

The basics of its operation are the following; first the picture which is intended to be captured is projected onto the surface of the photoactive region, i.e. the capacitor array. The charges accumulated on these capacitors are transferred into a charge amplifier and converted into voltage. Finally, this potential is converted into the digital domain adopting a sampling process and stored in the memory. A summary of this process is explained in Figure 1.4.

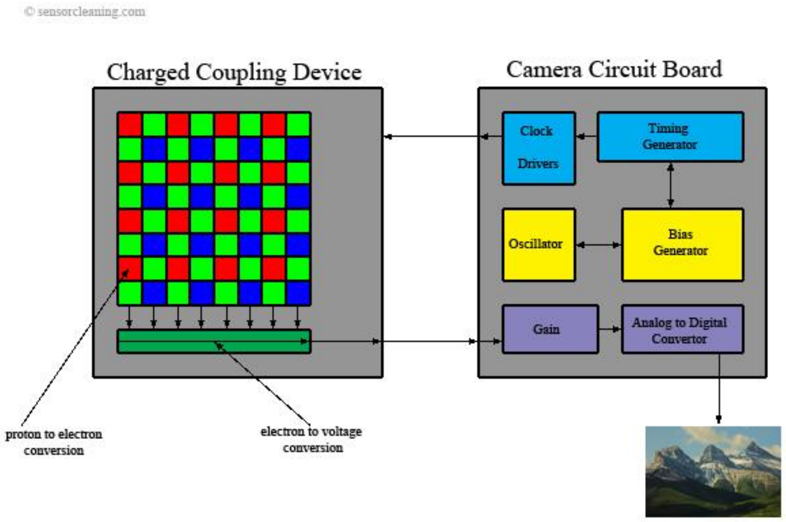


Figure 1.4 - Basic working principle of Charged Coupling Device

Courtesy: http://www.sensorcleaning.com/pics/CCD_sensor_diagram.jpg

1.2 Image Classification: An Overview

Thanks to the wide-spread usage of the devices with picture taking properties, the amount of digital images has grown exponentially [12]. However, mostly these images are not arranged according to any attribute. Instead, they are usually bulks of bytes in the storage units. This problem indicates the necessity of the classification algorithms in order to separate pictures according to pre-defined properties for better accessibility and utilizing the information in them.

In the dictionary, the term classification is defined as "*the act of putting people or things into groups based on ways that they are alike*" [13] whereas this definition can be expanded for image classification as the process in which a set of pictures are separated based on certain visual attributes. These attributes can be low–level cues such as basic shapes or high–level semantics such as the visual object and scene categories. Moreover, the number of common properties and classes can be diverse. However, as the content of the images become semantically more abstract and/or the number of classes increases, the methods to make the classification become more complex.

For example, one can simply use a color histogram feature (e.g., Red-Green-Blue *(RGB))* [14] in order to classify the pictures with red busses and the pictures with blue busses. Basically, the color histograms can be extracted and the peaks in the bins corresponding to the color red are found. Thus, those pictures are labeled as the red busses and the rest as the blue busses. Unfortunately, when a more complicated problem such as the classification of the four-legged animal images of four-legged animals is considered, the solution is not trivial anymore because the information included is semantically very high-level. To illustrate this situation, a good example would be the four-class classification problem of a set of pictures consisting of images containing dogs, cats, wolves and foxes. Even though they are easy to categorize for human eyes, it stands as a great challenge for a computer because the human beings have superior cognitive understanding of the environment meanwhile it is a hard task to teach the semantic connections in the images and the differences between the classes to a computer as computers interpret their inputs as ordinary numbers.

To summarize; image classification offers practical and automatic solutions for the generic problem of utilizing the information in the digital images. Despite the fact that it is very useful, there are still many limitations which should be overcome because the gap between the human understanding and the computers' is yet to be filled.

In the next section, the applications of image classification different domains are presented in order to give an idea about what have been achieved so far in different application domains. In section 1.2.2, examples of image classification in fashion domain are presented as merging fashion and image classification is the focus point of this thesis.

1.2.1 **Image Classification Applications**

An image content can vary drastically. Naturally this implies that researchers can apply image classification algorithms for different application scenarios. Even though there are numerous efforts in this field, a few examples are provided in this section to visualize the broad spectrum of image classification applications.

To begin with, one of the commercially successful applications of image classification is handwritten digit recognition by *Yann LeCun* exploiting *Artificial Neural Networks (ANNs)* [15] where the researchers are inspired from the brain’s working mechanism. *ANNs* are based on the small processing units called “artificial neurons”, which are biologic neuron-like small processing units. With the appropriate topologies and parameters learned from the training data, they are proven to be powerful. The goal of his work was to recognize people’s handwritings and convert them into digital form so that the digitized characters can be used for transactions which is illustrated in Figure 1.5. *LeCun’s* method has been used for reading more than 10% of the checks in the United States in the second half of 90s, which means over 25 million people used image classification in their daily lives.

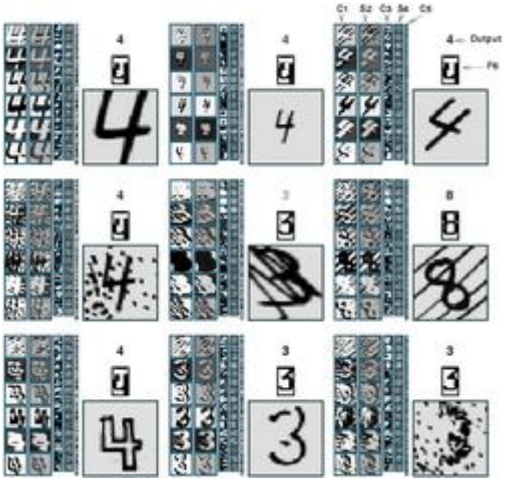


Figure 1.5 - Handwritten digit recognition

Courtesy: http://yann.lecun.com/ex/images/invar.png

Another application of image classification is face detection where the classification problem has a binary nature; the presence of face(s) in a picture. Once the faces are detected, it is also possible to recognize them from the pictures existing in a database. Since the face is unique for each person, it is used in police records, passports and driving licenses as a source of biometric information [16]. Additionally, it is also adopted by the public services for crowd surveillance, electronic mug shots book, and bank/store security [17].

An interesting stream of works related with face detection is emotion detection in still images and videos using facial expressions such as anger, happiness, sadness etc. Since the early 2000s, researchers have been trying to classify pictures and videos into pre-defined categories using diverse approaches. These algorithms generally consist of three main steps; face detection, feature extraction, and classification. For the first step, any well-working face detection algorithm such as Viola and Jones face detector's OpenCV implementation [18] can be used whereas for the feature extraction, there are three widely-adopted approaches exist in the literature. These can be divided as geometry-based, appearance-based and the combinations of these two. After the features are extracted, finally the classification phase is applied according to the authors' choices [19].

As a good example for the usage of image classification with different modalities, the authors in [20] propose classifying brain images obtained from *Computed Tomography (CT)* scans using image classification. The categories in this context are defined as "*inflammatory, tumor and stroke"*. In spite of the fact that the physicians are the primary decision makers, the authors claim that the *Computer Aided Diagnosis (CAD)* tools can be used as a second opinion to help the physicians. Similar to the works mentioned above, they extract features and classify them.

To conclude, image classification can be adopted to be used in different parts of people's lives to automatize the works which would be infeasible to do manually. Public security, automatic transactions, medicine are a handful of examples for its capabilities. In this theses, an application in fashion domain is demonstrated where the images are classified according to male vs. female and formal vs. informal cloth categories.

1.2.2 Image Classification and Fashion

Intuitively, any picture which includes human beings will most likely include clothes as well. Thus, applying image classification algorithms in fashion domain is meaningful as it can be helpful commercially such as arranging the collections of a design firm or recommending people specific types of clothes.

The general problems of image classification which are in-class variations and the gap between computer and human understanding are inherited and boosted in this specific domain because the description of the visual properties of the clothes in terms of color, texture, and shape is challenging. Besides these, the clothes can be folded, occluded or they can take the shape of the body that they are worn so that the geometric transformations should be taken into account.

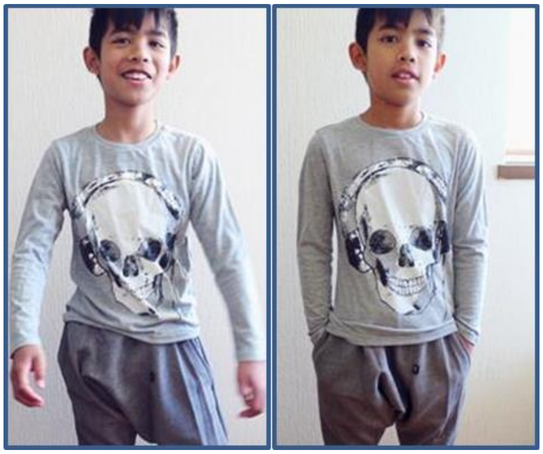


Figure 1.6 - Example for in-class variations in fashion domain

As it can be seen in Figure 1.6, even though these two pictures were taken in the same room with the same clothes and from the same person, content description of the clothes can vary because of the body pose and movement which cause changes in the cloth layout, i.e., the arms are wide open on the left hand side whereas the hands are in the pockets on the right hand side.

It is also a good example for exhibiting an illumination difference, i.e., the picture on the left hand side is slightly brighter comparing with the one on the right. This creates a problem for the classification process because different illumination will result variations of the colors in the picture. Consequently the values in three color planes (Red-Green-Blue) will be different as well as the extracted features.

The early attempts for adopting image classification in fashion domain focused on retrievals using color histograms, color sketch, shapes and textures, e.g., [21] where the authors proposed a fashion-centered system in order to handle large number of images. In [22], Cheng et. al proposed a method that can search clothes from a user's inventory. They extracted information about the physical appearance of the clothes and then learn these features according to the predefined categories. Once the system is created, the user can query an image keyword and occasion type so that the system can recommend the matching items. An interesting application was proposed by Chao et. al in [23],where they created a system called *Smart Mirror* which is a fashion recommendation system. Adopting a face detector, they extracted a part of the upper body using the relative position of the face. Following this, low-level color, texture and shape descriptors are obtained from the region cropped before. Finally they computed the (dis-) similarity distances between the query image and the images in the dataset. Thus the system recommends the closest matches. Qingqing et. al proposed an intelligent personalized fashion recommendation system in [24] by taking style, favorite color, and skin color into account. An automatic search system for clothes was developed by Wang et. al in [25]. Low-level dominant color is used for visual codebook creation and then certain high-level features are used for improving the search efficiency. Another interesting work proposed by Song et. al [26] predicts the occupation by using human clothing and scene contexts. The authors used four different parts of the human body which are *head, central upper body, left shoulder* and *right shoulder* by adopting face and human detection algorithms. Following this, four key points were located and four image patches are extracted, respectively. From these four patches, low-level descriptors were extracted for shape, texture and color representations.

In [27], Yang et. al proposed a real-time clothing recognition system for surveillance videos. For this purpose, they first took advantage of a face detection algorithm. After that, a rectangular region which is 5 times the face width and 9 times the face height is cropped. Note that this region includes both the body and the face. Finally, using color and shape descriptors they created the description of the clothes. Yamaguchi et. al gave an example for how to parse clothing in fashion photographs effectively [28]. The authors obtained super-pixels [29] as the first step. Then they estimated pose configuration and predicted the clothes. Chen et. al developed a fully automated system for populating a list for nameable attributes such as “No collar”, “Tank top”, “No sleeve” by adopting pose-adaptive low-level feature extraction from torso and arm regions [30]. Liu et. al tackled a different scenario in which a user captures a human photo on street and tries to retrieve related clothing pictures from an online shop [31]. In

[32], Bossard et. al proposed a method for recognizing and classifying 15 different cloth types under unrestricted conditions using multi-class learner based on a Random Forest.

The number of works for adopting image processing and classification techniques in the field of fashion are numerous yet according to the best of our knowledge, there is no study in the literature that addresses high-level sematic classification into four major categories; *Woman Informal, Woman Formal, Man Informal, and Man Formal.* In this context, the definition of *“Formal”* is the type of clothing item that one can wear in formal events such as weddings, cocktails etc. and the rest is defined as *“Informal”*. Examples for these categories are illustrated in Figure 1.7.

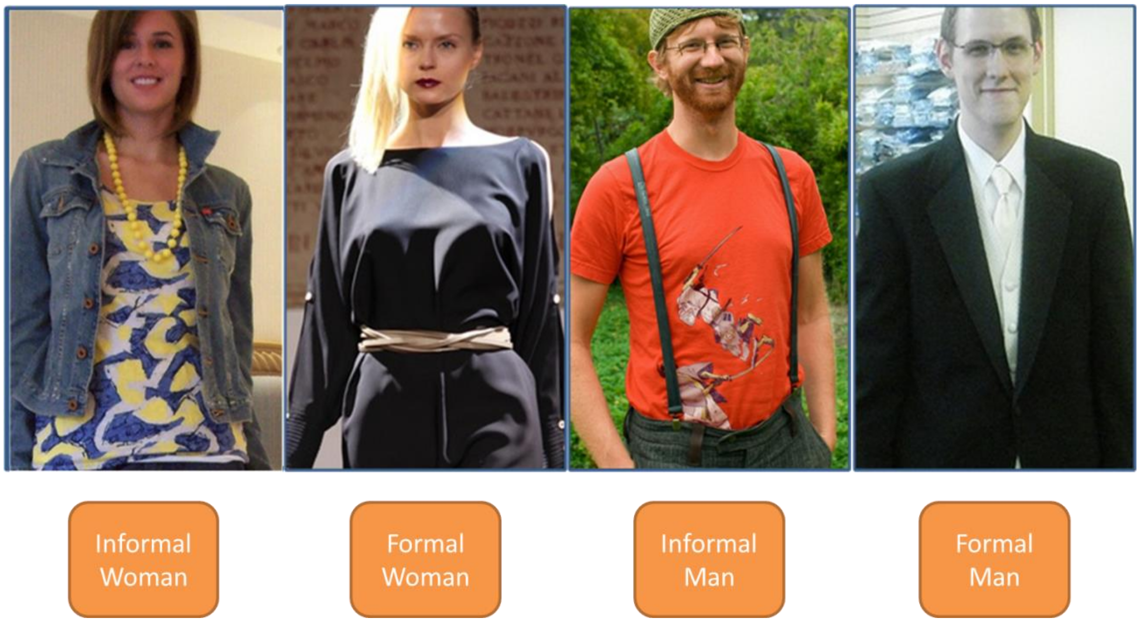


Figure 1.7 - Samples from each category

The common approach of the aforementioned studies is to extract low-level feature descriptors and then use them for further analysis and classification. Additionally, it is also important to choose the feature extraction schematic carefully as it can affect the effectiveness of the features drastically. For example in [26], the authors focused on part based models; however, in our work, we do not adopt such an approach because our problem is defined for clearly visible frontal pictures. Instead, we have used a region of interest *(ROI)* based approach as in [23]. For the sake of high level semantics, the attribute based approaches are also promising but our problem was being lack of sufficiently detailed labeled data. This fact made infeasible to use it.

In order to fill the gap between the low-level features and high-level semantics, a two level network classification topology is adopted, i.e. *Late Fusion* [33]. The first level of our system deals with the low-level features and gathers the basic information. The second level obtains the information from the classifiers in the first level and fuses them to further improve it. This choice is based on the assumption that each low-level feature is expected to get the information individually about shape, color, and texture. Then these features are needed to be unified by another classifier to translate them into high-level semantics. An overview of our approach is presented in Figure 1.8.

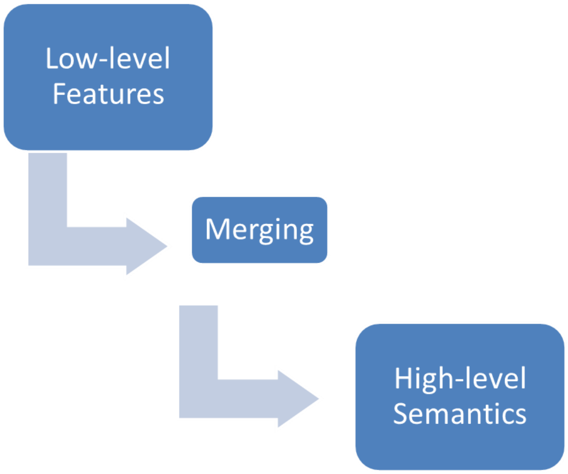


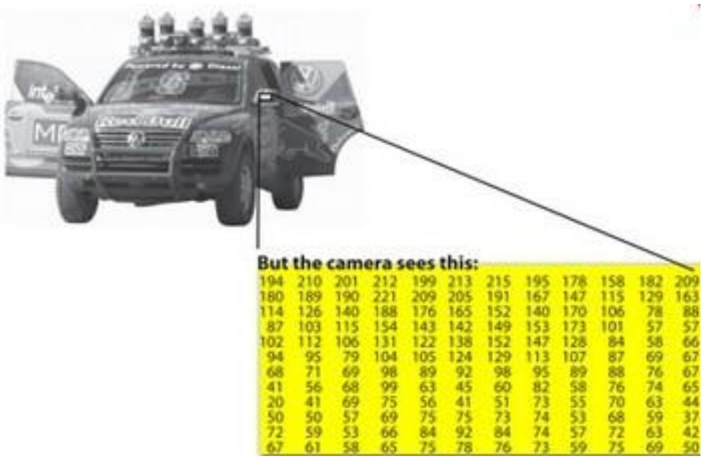
Figure 1.8 - Overview of the proposed method

2 Preliminaries

In this section, the tools which have been utilized for the feature extraction and the classification processes are introduced in sections 2.1 and 2.2, respectively. Moreover, brief descriptions of the software packages that are used during the implementation stage are provided in section 2.3.

2.1 Low-Level Descriptors

As it is illustrated in Figure 2.1, the digital images consist of discrete numbers whereas they mean shapes, colors and textures for the humans. In addition, human beings do not only recognize these shapes, colors and textures separately, but they can also comprehend the abstract relationships among them.



But the camera sees this:

194	210	201	212	199	213	215	195	178	158	182	209
180	189	190	221	209	205	191	167	147	115	129	163
114	126	140	188	176	165	152	140	170	106	78	88
87	103	115	154	143	142	149	153	173	101	57	57
102	112	106	131	122	138	152	147	128	84	58	66
94	95	79	104	105	124	129	113	107	87	69	67
68	71	69	98	89	92	98	95	89	88	76	67
41	56	68	99	63	45	60	82	58	76	74	65
20	41	69	75	56	41	51	73	55	70	63	44
50	50	57	69	75	75	73	74	53	68	59	37
72	59	53	66	84	92	84	74	57	72	63	42
67	61	58	65	75	78	76	73	59	75	69	50

Figure 2.1 - Digital image example

Courtesy:

http://docs.opencv.org/doc/tutorials/core/mat_the_basic_image_container/mat_the_basic_image_container.html

Considering a picture taken by a 5 megapixel camera, the resolution of the image we obtain is 2560 x 1920 and thus, the total number of pixels is equal to 4915200. In addition to this, the values of the pixels are extremely vulnerable against environmental changes such as illumination. Therefore, it is necessary to generate a compact description of the low-level concepts in the image. This process is done by using so-called *low-level features*. These mathematical operators can translate the raw pixels in an image into basic shapes, textures and colors as shown in Figure 2.2. Consequently, they create a basis for creating a high-level semantic relation among them.

In this work, among several visual descriptor alternatives the following shape, texture, and color descriptors are extracted: *Histogram of Oriented Gradients (HOG)*, on *Local Binary Patterns (LBP)* and *Color Structure Descriptor (CSD)*, *Color Layout Descriptor (CLD)*. Detailed information is provided in the following sections.

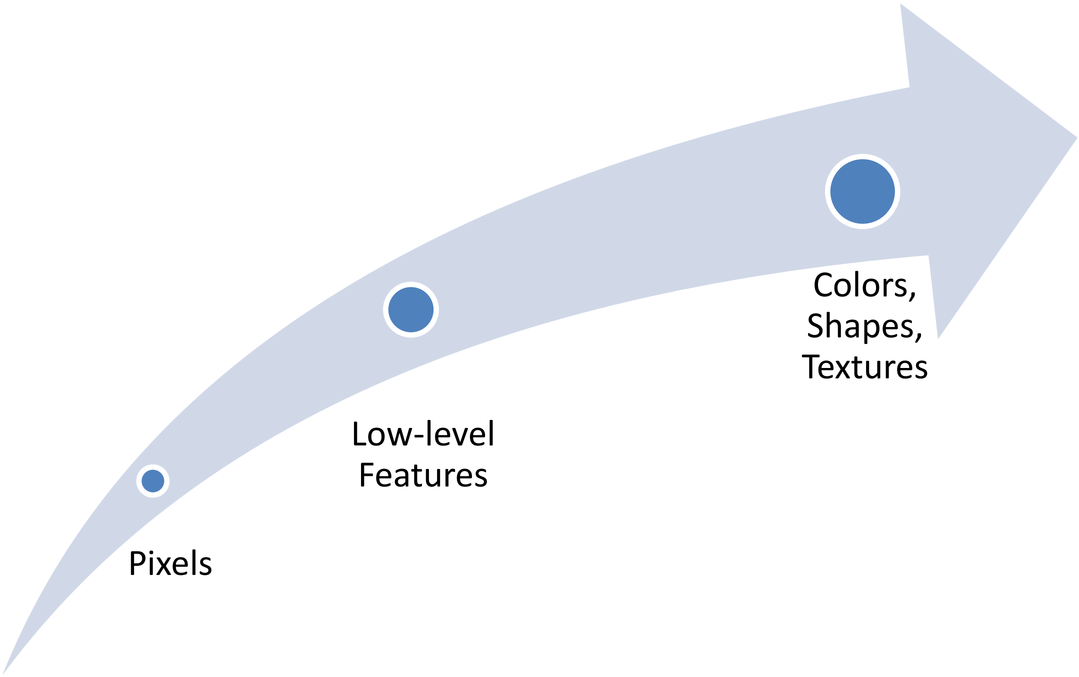


Figure 2.2 - The importance of the low-level features

2.1.1 Histogram of Oriented Gradients (HOG)

Since its introduction in *CVPR 2005* by Dalal et. al [34], Histogram of Oriented Gradients *(HOG)* has become one of the most widely used feature descriptors in the field of computer vision and image processing especially for object detection. The theory behind is counting how often the gradient orientations occur in pre-defined windows on an image. Instead of computing the histograms in salient points in an image as it is done in Scale-Invariant Feature Transform *(SIFT)* [35], *HOG* extracts them densely and uniformly. It retains the local changes while also providing the global shape information. As the silhouettes of the people differ according to the cloth they wear in most cases, it is therefore, chosen as the shape descriptor for the proposed approach.

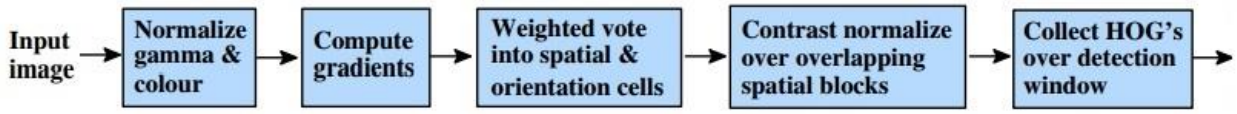

Input image
Normalize gamma & colour
Compute gradients
Weighted vote into spatial & orientation cells
Contrast normalize over overlapping spatial blocks
Collect HOG's over detection window

Figure 2.3 - Computation pipeline for Histogram of Oriented Gradients [34]

In spite of the fact that the authors' pipeline for the computation of *HOG* includes *Normalization of Gamma & Color* as it can be seen in Figure 2.3, they proposed to start with the gradient computation because it is reported that normalizations had only a small effect on performance. Moreover, their default detector does not include any gamma correction.

They tried many different masks such as cubic-corrected, 3x3 Sobel masks, 2x2 diagonal ones yet simple 1-D [-1,0,1] with smoothing scale σ = 0 gave the best results. The authors also mention that using larger masks reduces the performance.

The steps for computing the descriptor values are given below.

Step.1: Given the filter kernels;

$$D_x = [-1\ 0\ 1] \text{ and } D_y = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \quad \textbf{(1)}$$

$$D_x = [-1\ 0\ 1] \text{ and } D_y = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \tag{2}$$

The image *I* is convolved with the filter kernels and the following derivatives are calculated;

$$I_x = I * D_x \tag{3}$$

$$I_y = I * D_y \tag{4}$$

Where the magnitude of the gradient is;

$$|G| = \sqrt{I_x^2 + I_y^2} \tag{5}$$

and the orientation of the gradient;

$$\theta = \arctan \frac{I_Y}{I_X} \tag{6}$$

Step.2: The second step of the calculation process is “Spatial / Orientation Binning” in which a weighted vote is cast by every pixel in the cells and the orientation bins are populated with them. The bins are created with equal spaces between 0° - 180° or 0° - 360° which are called *unsigned gradient* and *signed gradient* respectively. In this context, the votes are the magnitudes because the authors claimed that they gave the best results.

Step.3: Following the binning process, the blocks are normalized in order to take the changes in illumination and contrast into account. For this purpose, four alternative normalization formulas are provided;

Let v is the unnormalized descriptor vector and $||v||_k$ is its *k-norm* for *k*=1,2 where $\boldsymbol{\varepsilon}$ is a small constant.

a) *L2-norm*

$$\mathbf{v} = \frac{\mathbf{v}}{\sqrt{||\mathbf{v}||_2^2 + \boldsymbol{\varepsilon}^2}} \qquad \textbf{(7)}$$

b) *L2-Hys;* the authors call this as *Lowe-style clipped L2-norm*

L2-norm is followed by limiting the maximum value of **v** to 0.2 and renormalizing it as in [35].

c) *L1-norm*

$$\mathbf{v} = \frac{\mathbf{v}}{||\mathbf{v}||_1 + \boldsymbol{\varepsilon}} \qquad \textbf{(8)}$$

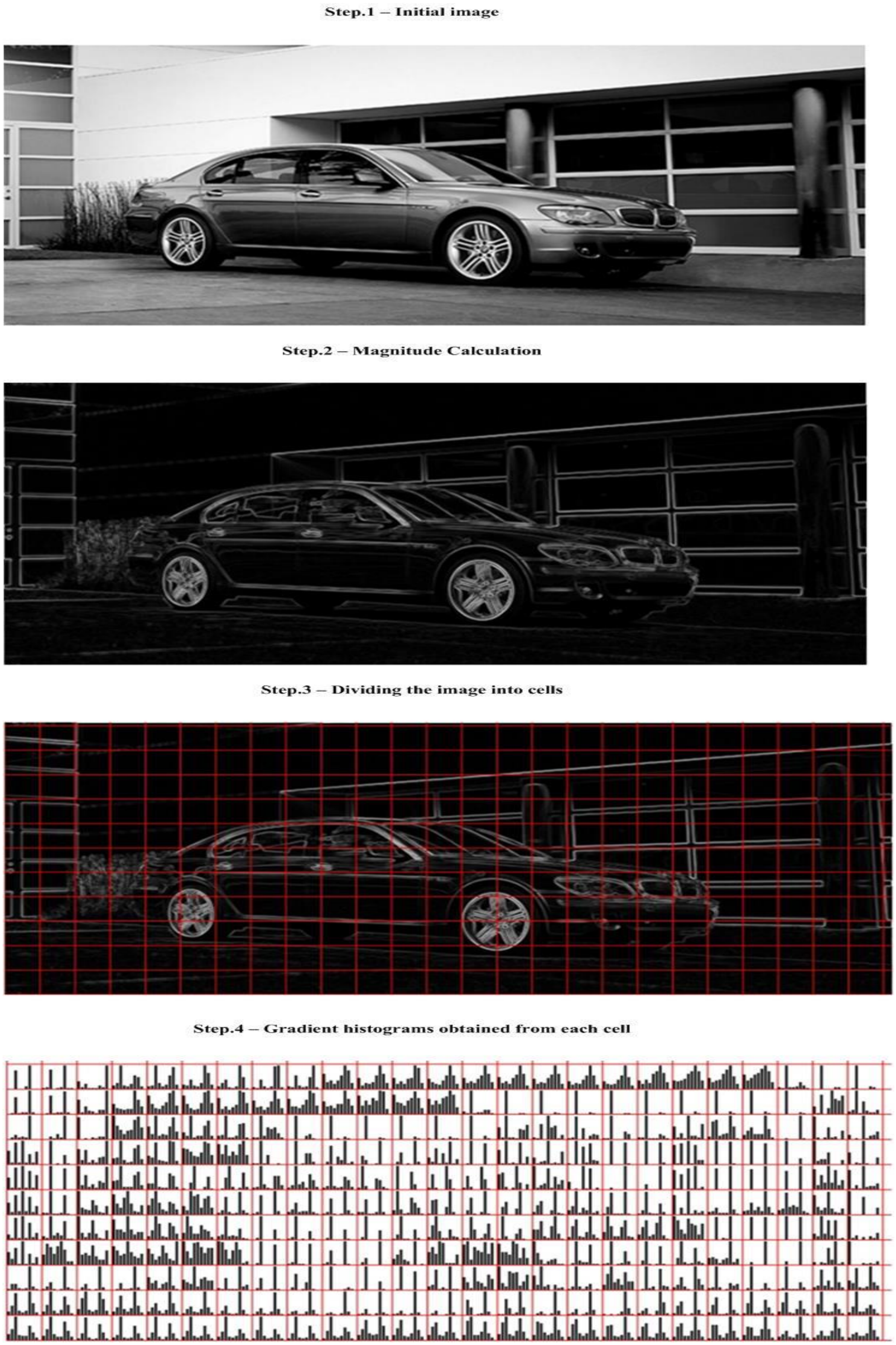
d) *L1-sqrt*

$$\mathbf{v} = \sqrt{\frac{\mathbf{v}}{||\mathbf{v}||_1 + \boldsymbol{\varepsilon}}} \qquad \textbf{(9)}$$

According to the Dalal et. al’s experiments, simple L1-norm degrades the performance whereas L2-Hys, L2-norm and L1-sqrt perform similarly. However, discarding the normalization process results in a 27% drop in the performance.

Instead of applying a global normalization process, Dalal et. al adopted a schematic where a group of spatially connected cells are merged into one bigger block and then the contrast normalization is applied locally. They do not only evaluate rectangular approach but also a circular topology which are called *R-HOG* and *C-HOG* respectively.

For *R-HOG,* Dalal et. al’s choice was using square η *x* η grids which create μ *x* μ cells with β orientation bins. On the other hand, *C-HOG* has four parameters which are given as the numbers of angular and radial bins, the radius of the central bin in pixels and the expansion factor for subsequent radii. Whole process is illustrated in Figure 2.4.



Step.1 – Initial image

Step.2 – Magnitude Calculation

Step.3 – Dividing the image into cells

Step.4 – Gradient histograms obtained from each cell

Figure 2.4 - HOG calculation illustrated

Courtesy: http://users.utcluj.ro/~igiosan/Resources/PRS/L5/lab_05e.pdf

2.1.2 Local Binary Patterns (LBP)

Proposed by Ojala et. al. in 1994 [36], *LBP* is a simple yet powerful texture descriptor. Even though many variants of it have been developed [37], [38], [39] the basic idea behind all is to compare a pixel value in a picture with its surrounding pixels. As a sample LBP computation is illustrated in Figure 2.5, in a certain window, if the compared pixel's gray level value is greater or equal than the center pixel, it is assigned as 1. Otherwise, it is labeled as 0. After the comparison, the assigned values are multiplied with the powers of two in a pre-defined order (i.e. clockwise or counter-clockwise) and summed. This process is applied to whole image and using the values obtained, a histogram is created as the representative feature vector.

The discriminative power and the simple computation made *LBP* one of the most commonly used feature descriptors in image analysis, retrieval and classification. For example Xianchuan et. al used LBP in medical image retrieval [40] whereas Ahonen et. al used it for face recognition application in [41]. In [42], Jiang et. al used *LBP* for human detection scenario by developing a novel variant which is called *Gradient Local Binary Patterns (GLBP)*. Zhao et. al uses *Volume Local Binary Patterns (VLBP)* with three orthogonal planes *(LBP-TOP)* in [43].

The value of the LBP code of a pixel (x_c, y_c) is given by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \qquad s(x) = \begin{cases} 1, if\ x \geq 0; \\ 0, otherwise. \end{cases}$$

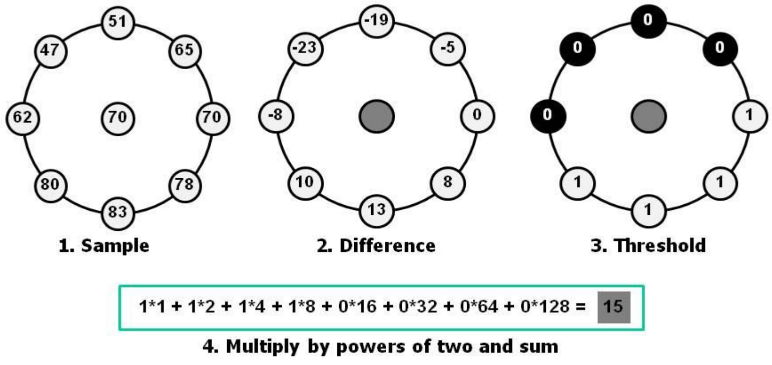


Figure 2.5 - Local Binary Patterns calculation

Courtesy: http://images.scholarpedia.org/w/images/thumb/7/77/LBP.jpg/400px-LBP.jpg

In this work, an extension of *LBP* called *Uniform LBP* is adopted [44] as it is aimed to achieve immunity against rotation and grayscale variance. To calculate *Uniform LBP*, first a circular neighborhood *(P,R)* is defined where *P* is the number of sample points and *R* is the radius of the neighborhood. The sample points are chosen as $(x_p,y_p) = (x + R\cos(2\pi p/P), y - R\sin(2\pi p/P))$ and in the case when the points do not overlap with the integer coordinates, a bilinear interpolation approach is used. The important difference between this variant and the ordinary *LBP* is using so-called *Uniform Patterns.* Ojala et. al noticed in their experiments that some of the binary patterns can be found more frequently in the images. Thus, they defined *Uniformity* of a binary pattern if it has at most two bitwise transitions from 0 to 1 or 1 to 0.

For a *(8,R)* neighborhood, there are 58 uniform patterns and the rest is labeled as non-uniform. Consequently, the number of bins is calculated as 59. The uniform patterns are illustrated in Figure 2.6.

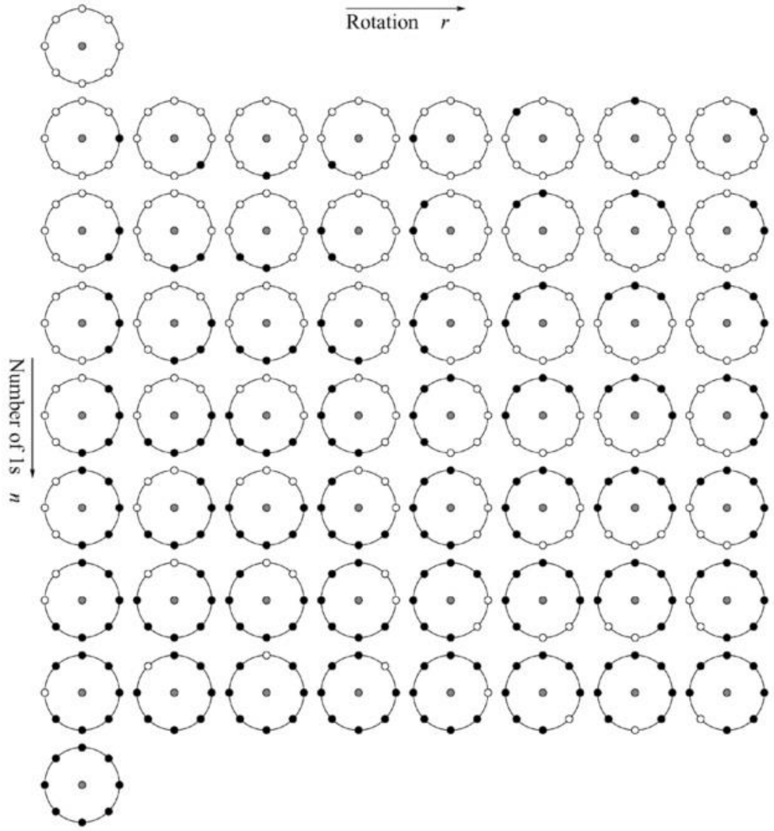


Figure 2.6 - 58 uniform patterns [44]

2.1.3 Color Structure Descriptor (CSD)

Considering that the formal clothing is expected to have a more homogenous color distribution which can be clearly seen in Figure 1.7, color descriptors for spatial color distribution are also used for classification in this study. In the case of ordinary color histograms, each pixel in the picture is assigned to the bin that its color matches. This approach gives a general idea about the statistics of the color quantities according to the quantized color bins of that image. However, it cannot describe how the color is distributed spatially in the image. In order to solve this problem, as a part of *MPEG-7 Visual Standard* [45], *CSD* is adopted which provides information about how the colors are spread in the image. As it can be seen in Figure 2.7, the numbers of colored pixels are same in both pictures however their distributions in the space are different. By using an ordinary color histogram, the feature descriptors for both images would be identical whereas adopting *CSD* gives different results as it takes spatial information into account.

The working principle of *CSD* is the following; the structuring element scans the whole image in a sliding window manner and checks if a specific color exists in the current window. If it is the case, the counter which corresponds to that color is incremented. Note that the number of colored pixels overlapping with the structuring elements does not make any difference. The counter will be incremented by one regardless the number of colored pixels in the window. A summary of this process is illustrated in Figure 2.8. This is the major difference between the ordinary color histograms and *CSD.*

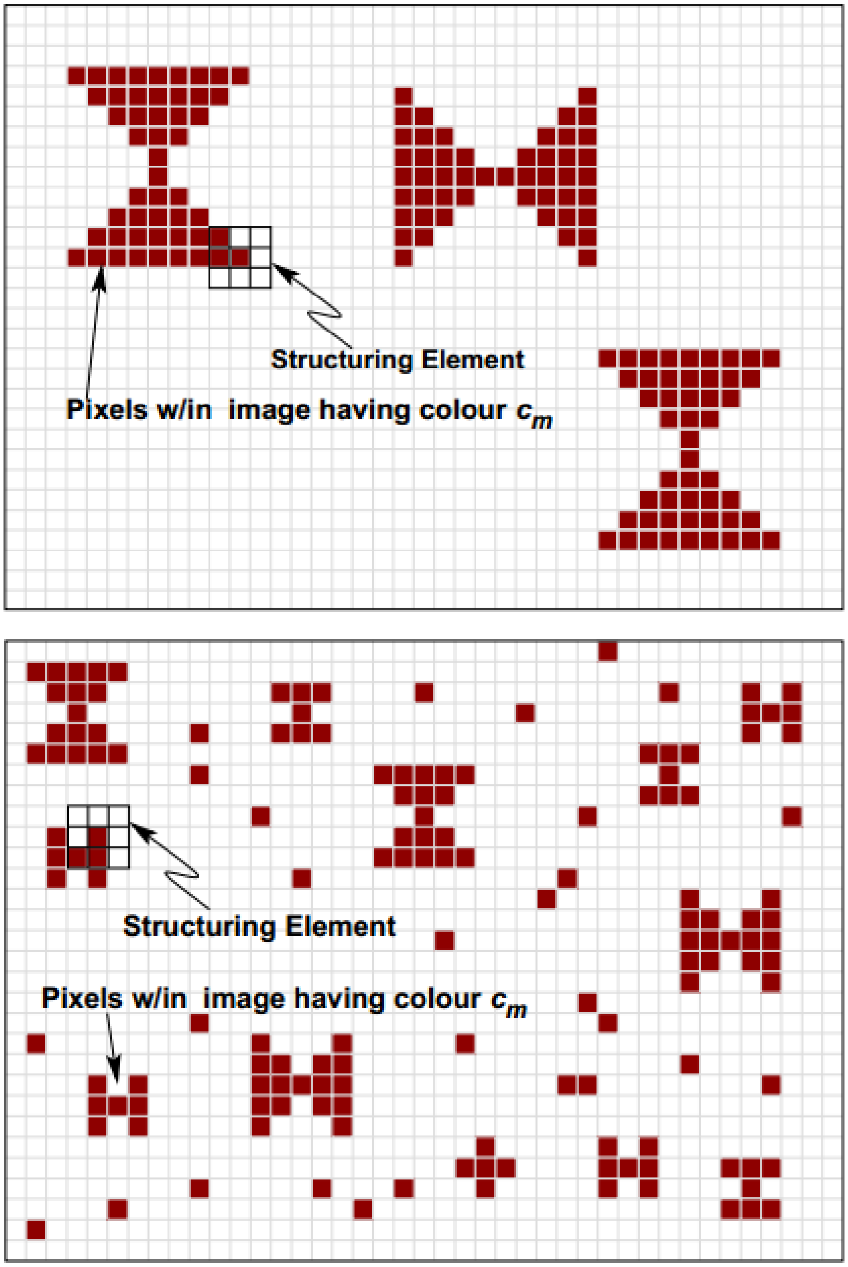
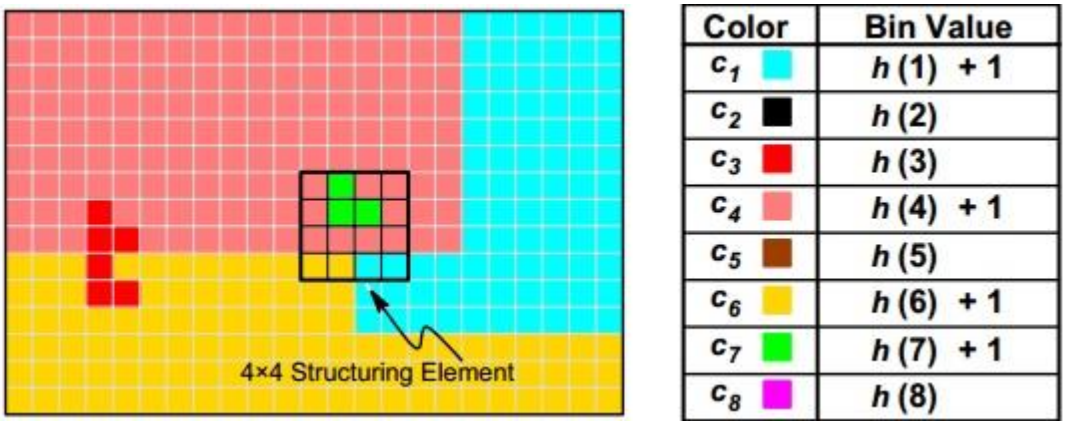

Structuring Element
Pixels w/in image having colour c_m
Structuring Element
Pixels w/in image having colour c_m

Figure 2.7 - Densely structured and sparsely structured distributions [46]



Color	Bin Value
c_1 🟦	h (1) + 1
c_2 ⬛	h (2)
c_3 🟥	h (3)
c_4 🟥	h (4) + 1
c_5 🟫	h (5)
c_6 🟨	h (6) + 1
c_7 🟩	h (7) + 1
c_8 🟪	h (8)

Figure 2.8 - Color Structure Descriptor calculation [47]

2.1.4 Color Layout Descriptor (CLD)

The second color-based feature descriptor adopted is *Color Layout Descriptor* [48]. It is also designed to extract the spatial distribution information of the major colors present in an image. For this purpose, the image is first partitioned into *64* blocks in order to secure the resolution invariance. Following this, a representative color is chosen from each block using any method such as averaging. Then each color plane is transformed using *Discrete Cosine Transform (DCT)* [49] . Finally, the obtained coefficients are zigzag scanned because of the fact that the resulting coefficients are sparsely distributed usually with many zeroes. The main process chain for the computation of *CLD* is given in Figure 2.9 and the detailed illustration is provided in Figure 2.10.

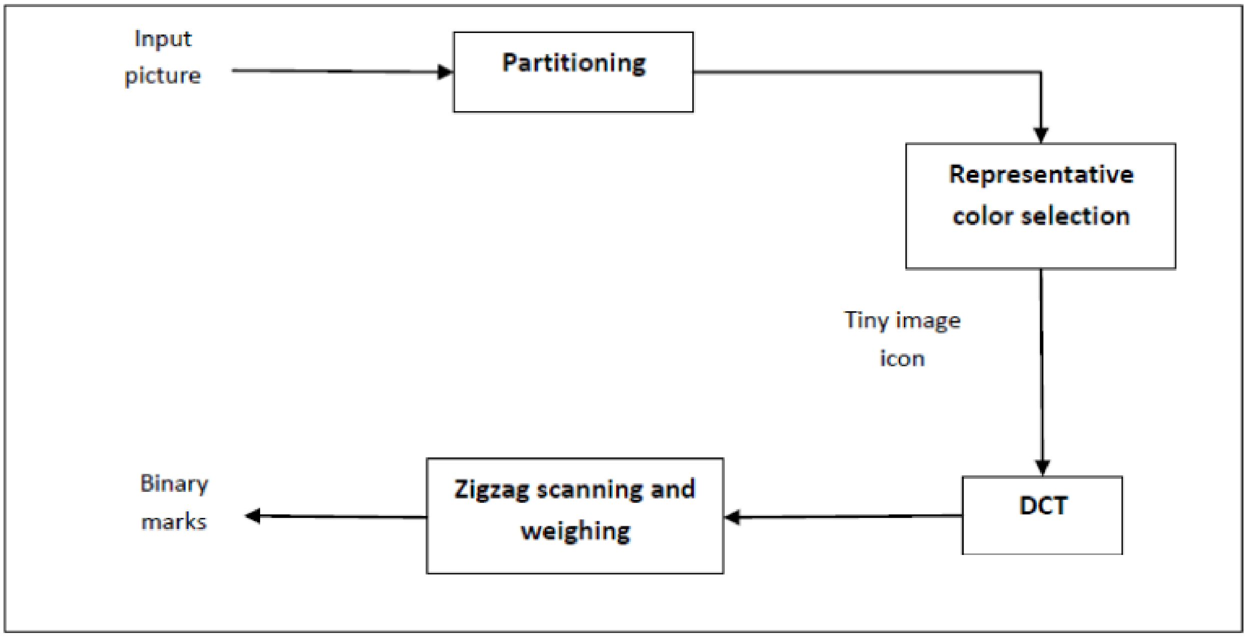
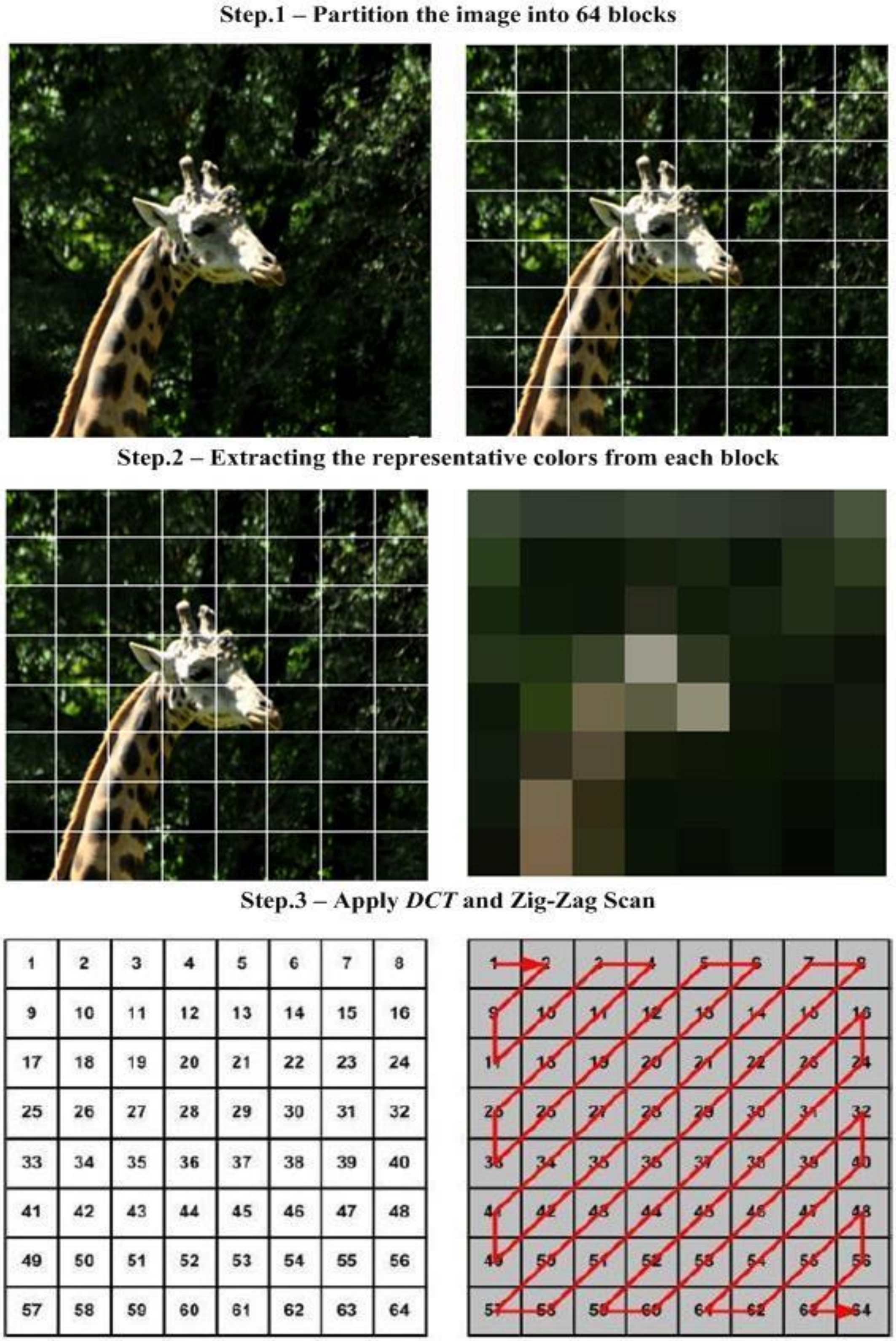


Figure 2.9 - CLD overview [50]



Step.1 – Partition the image into 64 blocks

Step.2 – Extracting the representative colors from each block

Step.3 – Apply *DCT* and Zig-Zag Scan

1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56
57	58	59	60	61	62	63	64

Figure 2.10 - CLD calculation illustrated

Courtesy: http://en.wikipedia.org/wiki/Color_layout_descriptor

2.2 Support Vector Machines

Introduced in 1995 by *Cortes* and *Vapnik* [51], Support Vector Machines are one of the most widely used classifiers in different fields thanks to their powerful classification abilities. It is proven as a superior classifier in the literature [26], [27], [30], [52], [53] and hence it is well-implemented and optimized thanks to the open source libraries such as *LibSVM* [54].

Assume a dataset in which the samples belong to two different categories and denoted as {(**x**, **y**)}. In this exemplification, **x** represents the samples, **y** represents the label vectors (+1 or -1) and **w** represents the weight vector.

$$f\,(\mathrm{x}) = \mathbf{w}^{\mathbf{T}}\mathbf{x} + b = 0 \qquad \textbf{(10)}$$

If b = 0, the points which hold the equation (10) build a *hyperplane* and according to b, it will create an offset from the origin for the *hyperplane*. The form of the *hyperplane* depends on the dimension of the feature space; i.e. in a two dimensional Cartesian Space, this will be a line whereas it will have the shape of a two dimensional plane in the case of three dimensional space. The end result of this approach will be a simple linear classifier which divides the sample space into two parts. Aforementioned sample space and the separating plane can be observed in Figure 2.11.

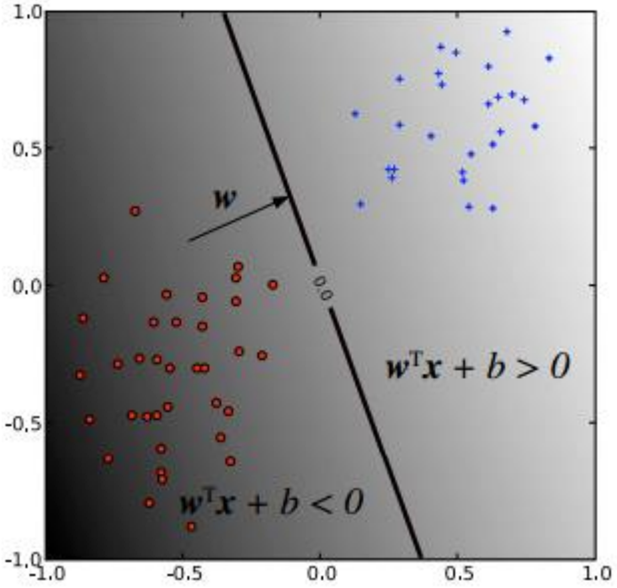


Figure 2.11 - Linear classifiers [55]

As it is explained in [55] in detail, it would be naive to expect linear classifiers to have strong classification properties however they are advantageous in terms of computational complexity especially when the dataset is considerably big. Thus the idea of using linear classifiers to create non-linear classifiers arose.

The first approach is extremely straightforward; mapping the input space X to F by using a non-linear function $\phi : X \rightarrow F$ and hence the discriminant function becomes;

$$f\,(\mathrm{x}) = \mathrm{w}^{\mathrm{T}}\, \phi(\mathbf{x}) + b \qquad \textbf{(11)}$$

An example for direct mapping would be the following [55];

$$\phi(\mathrm{x}) = (x_1{}^2, \sqrt{2}\, x_1\, x_2, x_2{}^2)^{\mathrm{T}} \qquad \textbf{(12)}$$

and then

$$\mathrm{w}^{\mathrm{T}}\, \phi(\mathbf{x}) \;= w_1 x_1{}^2 + \; \sqrt{2} w_2 x_1 x_2 + w_3 x_2{}^2 \qquad \textbf{(13)}$$

It can be seen that the problem has a quadratic nature which requires a quadratic increase in terms of memory and computation time. Infeasibility of this methodology for the large amount of data and huge dimensions albeit the small amount of data with low dimensional equations can be solved thanks to the powerful computers of our era. In order to solve this problem, the method what is called *Kernel Trick* is applied which is illustrated in Figure 2.12. Let’s assume that vector **w** is defined as a linear combination of the input samples (training samples);

$$\mathrm{w} = \textstyle\sum_{i=1}^{n} a_i\, x_i \qquad \textbf{(14)}$$

When the equation (14) is put into equation (10);

$$f\,(\mathrm{x}) = \textstyle\sum_{i=1}^{n} a_i \mathrm{x}_i^T \mathrm{x} + b \qquad \textbf{(15)}$$

The translation of equation (15) into the feature space F implies;

$$f\,(\mathrm{x}) = \textstyle\sum_{i=1}^{n} a_i\; \phi(\mathrm{x})^{\mathrm{T}}\, \phi(\mathrm{x}) + b \qquad \textbf{(16)}$$

At this point, if the kernel function $k\,(x,x^{'})$ is defined as;

$$k\,(x,x^{'}) = \phi(x)^{T}\,\phi(x^{'}) \qquad \textbf{(17)}$$

The discriminant function can be simply obtained as the following;

$$f\,(x) = \textstyle\sum_{i=1}^{n} a_i\; k\,(x,x_i) + b \qquad \textbf{(18)}$$

The equation (18) proves that without explicitly mapping of samples into the feature space *F*, the mapping can be computed.

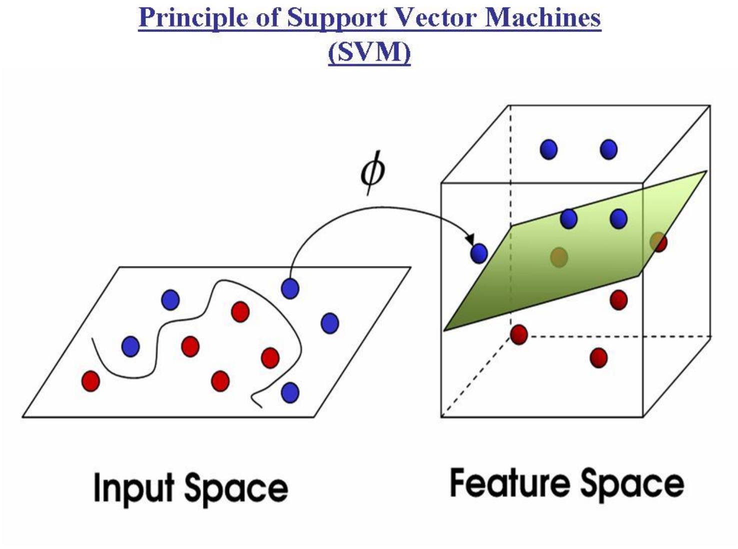


Figure 2.12 - Kernel transformation in Support Vector Machines

Courtesy: http://www.nectarineimp.com/wp-content/uploads/2013/08/machine-learning-svm.jpg

For a given dataset *D*, the margin of its *hyperplane* f can be defined as;

$$m_D\,(\,f\,) = \frac{1}{2}\,\hat{w}^{T}\,(x_{+} - x_{-}) \qquad \textbf{(19)}$$

where $\hat{w}$ is a unit vector in the direction of w, x_{+} is the closest sample to the *hyperplane f* for positive examples and x_{-} is the vice versa.

In the scenario that both the x_{+} and x_{-} have the same distance to the decision boundary, the following equations are obtained;

$$f\,(\mathrm{x}_{+}) = \mathrm{w}^{\mathrm{T}}\mathrm{x} + \ + b = a \tag{20}$$
$$f\,(\mathrm{x}_{-}) = \mathrm{w}^{\mathrm{T}}\mathrm{x} + \ + b = a$$
$$\text{where } a > 0.$$

If the equation (20) is put into equation (19)

$$m_D\,(\,f\,) = \frac{1}{||w||} \tag{21}$$

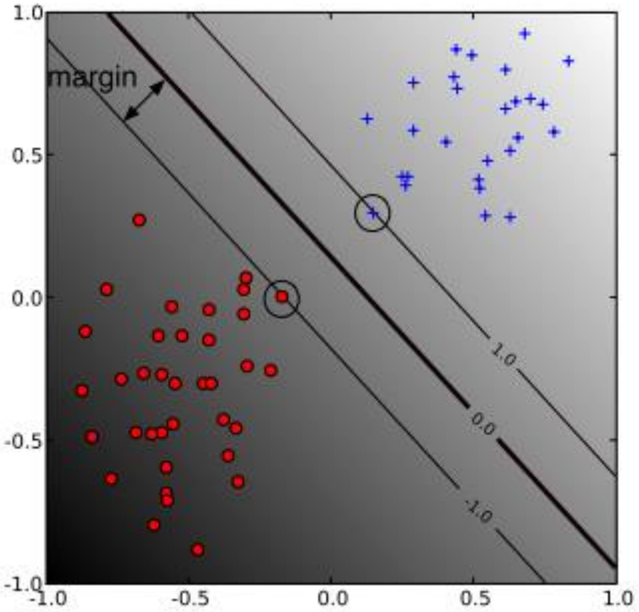


Figure 2.13 - Demonstration of margin [55]

As it can be seen in Figure 2.13, if the maximum margin can be found, the errors during the test phase can be minimized because this margin will create the largest possible separation between the two data sets.

Expectedly, it is not possible to find a margin which can separate the classes perfectly. This led people to invent *Soft-Margin SVM.* The idea here is to allow some of the training examples to be misclassified by using constraints where the misclassified examples are called as *Support Vectors*. Two samples which are encircled in Figure 2.13 are a simple example of this case.

It can be inferred from the equation (21) that minimizing $\frac{1}{||w||}$ will generate a maximum geometric margin thus if $||\mathrm{w}||^2$ can be minimized, it will also do the same job. Now the problem has become the following;

$$\text{minimize } \frac{1}{2}\,||\mathrm{w}||^2$$
$$\text{subject to: } y_i\,(\mathrm{w}^{\mathrm{T}}x_{\mathrm{i}} + b) \geq 1 \;\; i = 1,\ldots, n \qquad \textbf{(22)}$$

In this situation, it is known that the decision boundary will classify all the samples correctly however this will also lead to a smaller margin which is undesirable as it is explained above. With a small change in the equation (22), it can be allowed for classifier to have some errors whereas it achieves a much larger geometrical margin and a better separation. Thus, so-called *slack variables* are introduced;

$$y_i(\mathrm{w}^{\mathrm{T}}x_{\mathrm{i}} + b) \geq 1 - \xi_{\mathrm{i}} \;\; i = 1,\ldots, n$$

Since it is necessary to limit the number of misclassified examples, the errors must be penalized. Consequently, the problem becomes;

$$\text{minimize } \frac{1}{2}\,||\mathrm{w}||^2 + C\sum_{i=1}^{n} \xi\mathrm{i},\; C \geq 0$$
$$\text{subject to: } y_i\,(\mathrm{w}^{\mathrm{T}}x_{\mathrm{i}} + b) \geq 1 - \xi_{\mathrm{i}}\,,\; \xi_{\mathrm{i}} \geq 0 \qquad \textbf{(23)}$$

As it can be deduced from equation (23), the parameter (soft-margin constant) *C* affects the orientation and the width of the *hyperplane* which have a huge effect on the success of the classifier. If a big value for *C* is used, there will be less support vectors to be used and a narrow margin will be obtained whereas it will be the vice versa in the case of small *C*. This situation is illustrated in Figure 2.14.

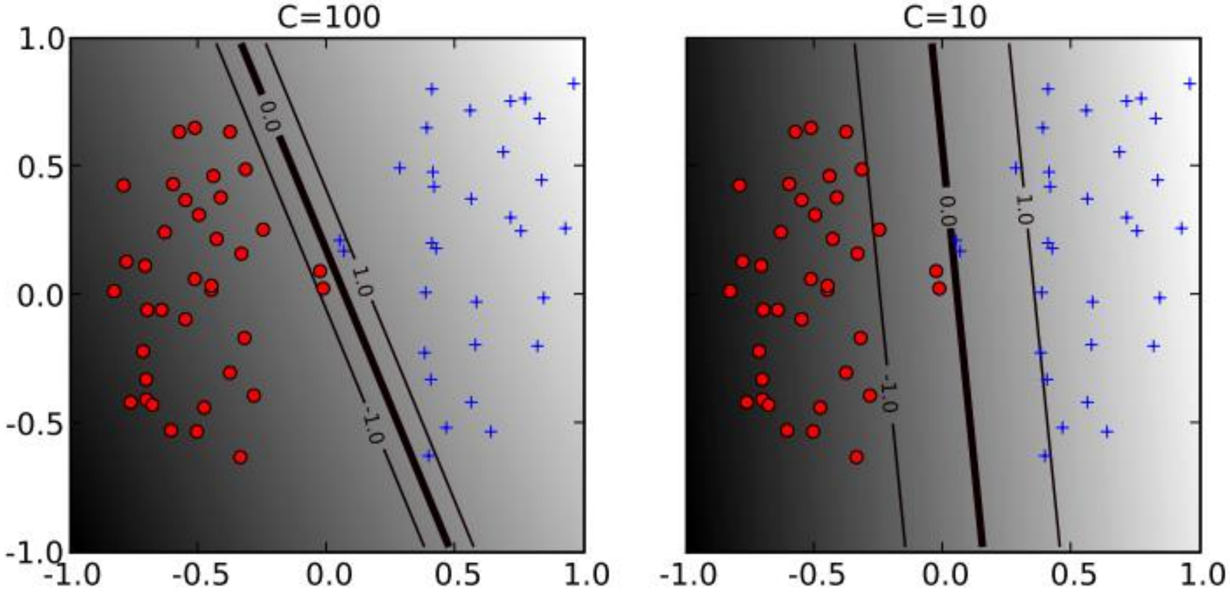


Figure 2.14 - The effects of parameter *C* [55]

For the Gaussian Kernel *k* (x,x') = exp (- γ ||x-x'||^2), it is also important to tune γ. In Figure 2.15, it can be observed that γ is inversely proportional with the number of support vectors. Despite the fact that choosing a big γ will result a good classification rate on training set, it will overfit which will cause problems during the test scenarios.

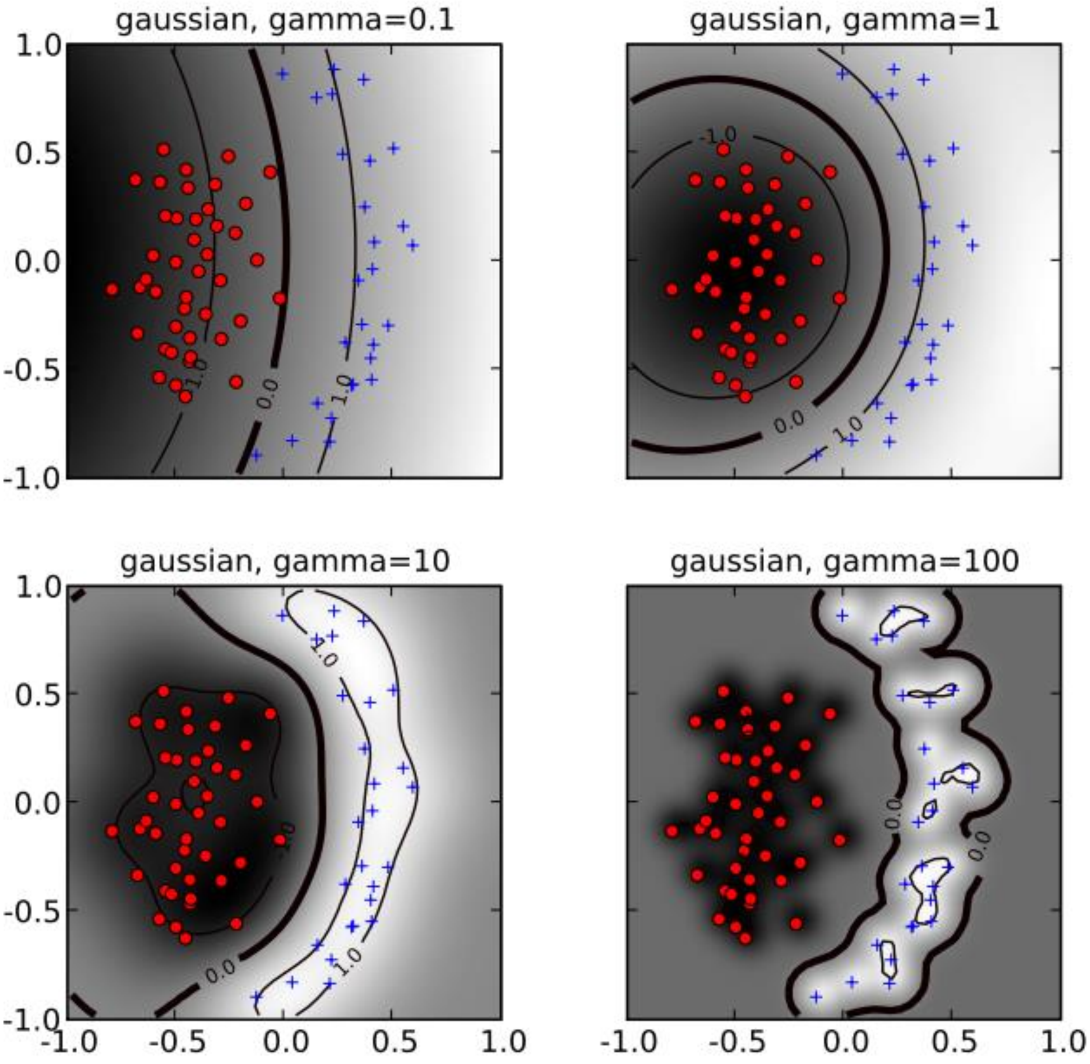


Figure 2.15 - Effect of parameter γ [55]

2.3 Software Libraries

In order to implement the proposed approach, a software application in C++ is developed using mainly two open source libraries: Open Source Computer Vision *(OpenCV)* [56] and Waikato Environment for Knowledge Analysis *(Weka*) [57]. *OpenCV*'s image storing, processing and analyzing capabilities are exploited. Additionally, *Weka* is used in order to explore different classifiers. In the following subsections, each software package will be presented.

2.3.1 OpenCV

OpenCV is a commonly used open source library in image and video analysis, processing and computer vision fields. Starting as an Intel Research initiative in 1999, it has become an essential tool for many researchers, students thanks to its open source nature with *BSD* license. Even though it is written in C++, there is also full support for the interfaces for Python, Java, MATLAB/OCTAVE. As a cross platform library, it is possible to run it on Windows, Linux and OS X. On the other hand, there has been a great effort to use its properties on mobile platforms such as iOS, Android, Blackberry.

Currently there are more than 2500 optimized algorithms for computer vision and machine learning including [58];

- Detecting and recognizing faces
- Identifying objects
- Classifying human actions in videos
- Tracking camera movements
- Tracking moving objects
- Extracting 3D models of objects
- Producing 3D point clouds from stereo cameras
- Stitching images together to produce a high resolution image of an entire scene
- Finding similar images from an image database,
- Removing red eyes from images taken using flash
- Following eye movements
- Recognizing scenery and establishing markers to overlay it with augmented reality

The reason why it is used in this work is the simple interface that it provides for image processing and analyzing.

2.3.2 **Weka**

Weka is a tool and a library for machine learning and data mining algorithms [57]. As an open source software with Gnu's Not Unix *(GNU*) General Public License *(GPL)*, it provides a broad spectrum of algorithms which are for data pre-processing, classification, regression, clustering, association rules, and visualization. Even though it was first developed in C, it is migrated to Java since 1997. Thanks to the platform independence of Java, it can be used on any platform which can run Java Virtual Machine *(JVM)*. Additionally, it also has a graphical user interface which creates a steep learning curve for the beginners.

3 The Proposed Approaches

The common main points for all three proposed approaches are; finding the region where the person's face and body are located, extracting the features and finally classifying the features using *SVMs.*

In this work, *HOG, LBP*, *CSD* and *CLD* are chosen as the features to be extracted because it is proven that they provide superior performance for extracting information from the pictures with clothing information [23], [26], [27], [32]. For evaluating the effects of the different feature extraction approaches, three alternative extraction methods which are grouped under two main sections are proposed;

- Global Feature Extraction Approach
- Grid Based Feature Extraction Approaches

Additionally*, Grid Based Feature Extraction Approaches* are also divided into two sub-sections;

- Nine Grids Feature Extraction Approach
- Seven Grids Feature Extraction Approach

3.1 Feature Extraction Methods

The initial step for all the methods proposed is running a face detection algorithm on each image in order to obtain the coordinates and the size of the face present in the image. According to the width and the height measures of the face, a bounding rectangle box is created which covers both upper-body and the face. The reason behind this cropping process is discarding the irrelevant information and consequently reducing the amount of noise which would be extracted by the feature descriptors. Since the pictures in the dataset are mostly frontal and with straight body poses, it is a safe assumption that the body will be found as a rough rectangle under the face. The final step is to extract the four features according to one of the feature extraction schematics proposed and write the extracted feature vectors into text files. This process is shown in Figure 3.1.

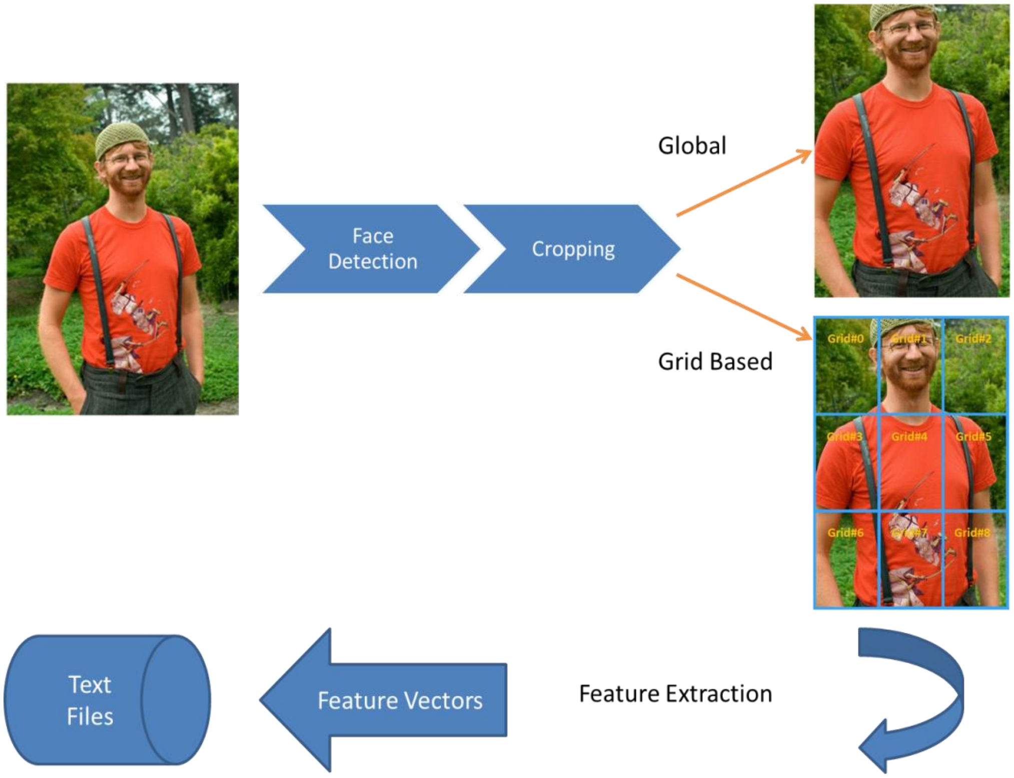


Figure 3.1 - Overview of the proposed feature extraction method(s)

3.1.1 Global Feature Extraction Approach

As the first feature extraction method, a global extraction schematic is adopted. Each of the four feature descriptors is applied globally and their outputs are stored. Thus, the feature vector sizes for this approach are the following;

- *HOG:* 3780 dimensions (105 blocks x 36 bins)
- *LBP:* 59 dimensions (8,R uniform patterns)
- *CSD:* 32 dimensions (32 bins histogram)
- *CLD:* 58 dimensions (28 bins histogram for luminance, 15+15 bins histogram for the color components)

The advantage of global approach is being able to generate a condensed description of a given image [59].

3.1.2 Grid Based Feature Extraction Approaches

The second group of feature extraction methods is grid based approaches. Even though the global approaches are expected to work well in the sense that they can obtain the general appearance, shape and the texture information in the image, they also tend to miss the local information. Additionally, they are also vulnerable against possible occlusions [59]. Thus grid based approaches in which the pictures are divided into 3x3 grids as shown in Figure 3.2 are also applied. Following this, the features are extracted from each grid separately. Finally they are concatenated in the end of the process. The number of grids, 9, is chosen for avoiding the possible noise-like local variances while retaining the information from the processed grid.

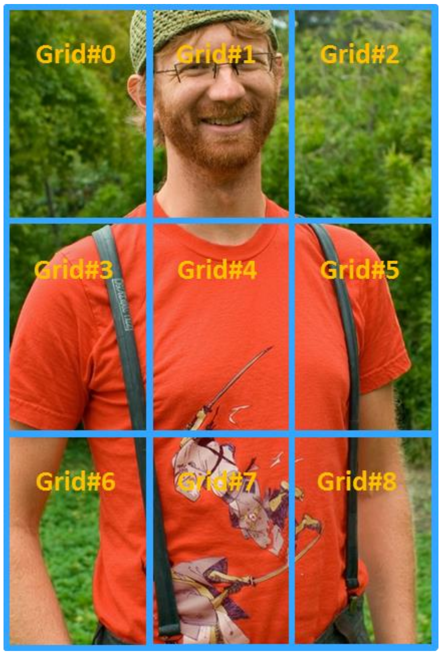


Figure 3.2 - Grid based approach

3.1.2.1 Nine Grids Feature Extraction Approach

In this scenario, *LBP*, *CSD* and *CLD* are extracted separately from each grid and then concatenated for creating a single feature vector as it is illustrated in Figure 3.3. *HOG* is intentionally excluded from this process because of the fact that *HOG's* nature already implies a window based feature extraction which makes the explicit grid based feature extraction redundant. Instead, the feature vectors obtained from globally extracted *HOG* are divided into nine according to the part of the image they are extracted and each sub-vector is treated as in *LBP/CSD/CLD*.

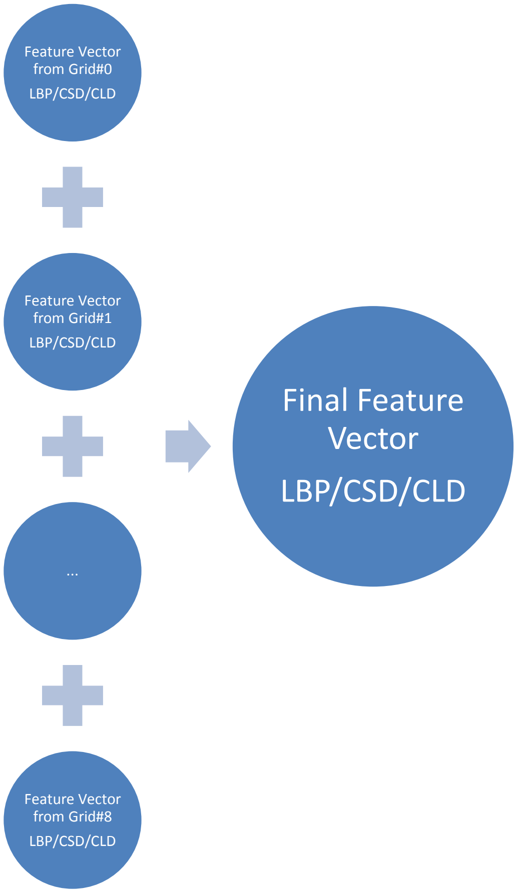

Feature Vector from Grid#0
LBP/CSD/CLD
Feature Vector from Grid#1
LBP/CSD/CLD
...
Feature Vector from Grid#8
LBP/CSD/CLD
Final Feature Vector
LBP/CSD/CLD

Figure 3.3 - Grid based feature extraction example for *LBP, CSD, CLD*

The sizes of the final feature vectors are the following;

- *HOG:* 9 x 420 = 3780 dimensions
- *LBP:* 9 x 59 = 531 dimensions
- *CSD:* 9 x 32 = 288 dimensions
- *CLD:* 9 x 58 = 522 dimensions

3.1.2.2 Seven Grids Feature Extraction Approach

Note that in Figure 3.4, *Grid#0* and *Grid#2* do not contain any valuable information about the cloth item or the person. This visual observation led to consider an alternative approach in order to try to further reduce the noise present in the images. Consequently, the second method is proposed in which *Grid#0* and *Grid#2* are excluded.

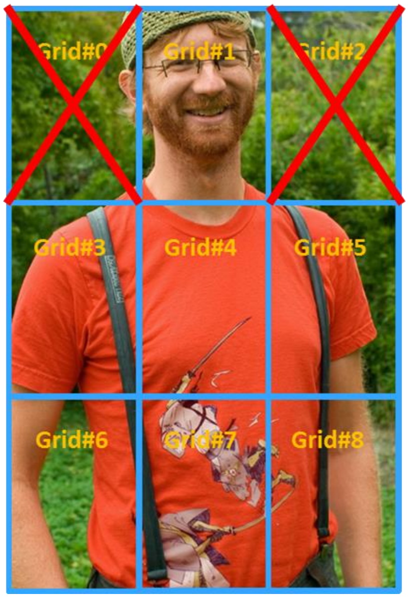


Figure 3.4 - Exclusion of the redundant grids

In this approach, the features which are calculated for nine grids approach are used again. However, instead of concatenating all of the nine grids, only the features extracted from *Grid#1*, *Grid#3, Grid#4, Grid#5, Grid#6, Grid#7, Grid#8* are concatenated to form the final feature vector.

The sizes of the final feature vectors are the following;

- *HOG:* 7 x 420 = 2940 dimensions
- *LBP:* 7 x 59 = 413 dimensions
- *CSD:* 7 x 32 = 224 dimensions
- *CLD:* 7 x 58 = 406 dimensions

3.2 The Classification Framework

A supervised classification framework is adopted in this work in order to correctly classify the pictures into the corresponding categories; *Woman Informal, Woman Formal, Man Informal, and Man Formal*. Using Support Vector Machines *(SVM)* [51] as the basic building block, a two level topology is used.

3.2.1 General Topology

The adopted classification method in this work consists of a binary-tree structure. First of all, the pictures are classified according to the person's gender, i.e. *Woman-Man.* Then according to the result of the first stage, the second phase which is the decision of *Informal-Formal* is evaluated, and the final result is obtained.

Let *I* be the image that is going to be classified and its features are I_{HOG}, $I_{LBP,}$ I_{CSD} and I_{CLD}. Initially, these features are used for deciding whether the person in the image *I* is a female or male. If it is a female, I_{HOG}, $I_{LBP,}$ I_{CSD} and I_{CLD} are sent to *Woman Informal – Formal* classifier. For the male's case, this classifier is replaced with *Man Informal – Formal*. The pipeline of the mentioned process is shown in Figure 3.5.

In this thesis, a *late fusion* scheme [33] is adopted for fusing the information from different sources, i.e. features. According to this scheme, firstly, each feature is fed as an input to a separate classifier as it is illustrated in Figure 3.6 and then their results are used for feeding a second level classifier which makes the ultimate decision. Let us consider the female - male separation as an example; each of the SVMs in Figure 3.6 gives a class label as the result of its own classification. By concatenating the four class label outputs, a new vector V is formed which represents the decisions for the image I in terms of shape, texture and color. Using V, a second level of classification is applied thus the final decision is made.

The reason why this approach is chosen stems from the nature of the cloths and the human beings. Intuitively, they consist of the combinations of shapes, textures and colors. Moreover, the human genders have usually different silhouettes which can be separated by using the shape as a feature.

In order to exploit these facts, we let the classifier learn the relationships between each of the individual features (shape, texture, color) and the concepts (either genders or the formality of the cloth) in the first level of the classification network. Once these relationships are established, the second level of the network uses this information created in the first phase as a basis to further explore the associations among the combinations of the low-level visual cues and the high level concepts.

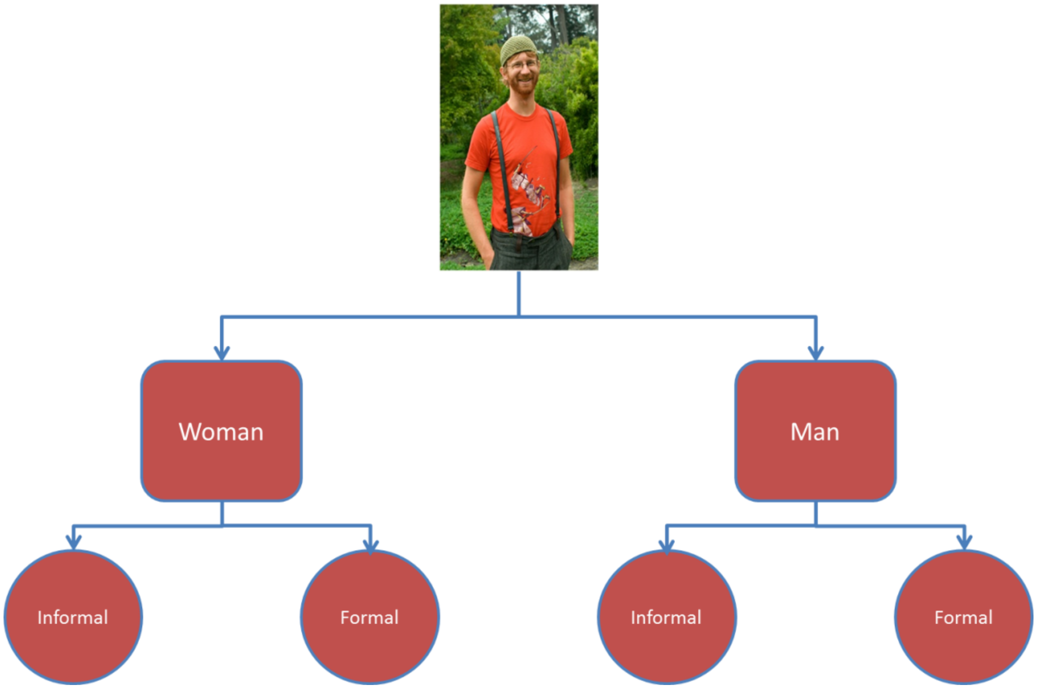


Figure 3.5 - Classification pipeline

Note that each of the red blocks in Figure 3.5 includes all the components shown in Figure 3.6. Thus, for each of the binary classifications, i.e. *Woman, Man, Informal*, *Formal*, four classifiers for four different features are used as the first step. Hence, a merging classifier to make the decision is also adopted in the second step.

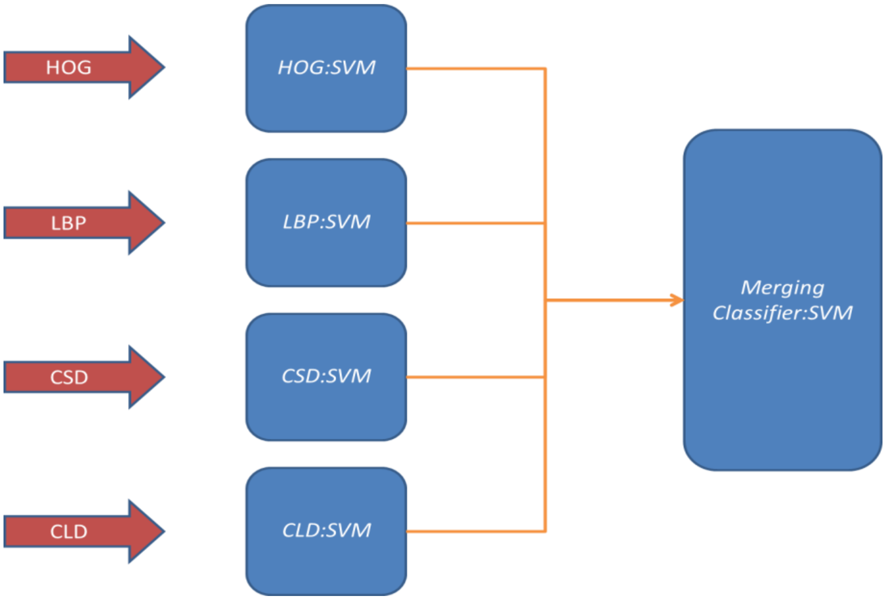


Figure 3.6 - Classification topology for binary classifications (Late Fusion)

4 Experimental Results

For evaluating the proposed algorithms, detailed experiments are conducted. In the classification experiments we used a PC with an Intel Core i7-3720QM 2.60 GHz Central Processing Unit *(CPU)*, NVIDIA Quadro K2000M *GPU* and 16 GB Random Access Memory *(RAM)* using Windows 7 64-bit Enterprise. The features are extracted according to the algorithms proposed in Section 3.1 and then they are given as the inputs to the classification algorithm mentioned in Section 3.2.

In this section, first the information about the dataset used in this work is provided in Section 4.1. Following this, the metrics to measure the performance of the proposed approaches are explained in Section 4.2. Section 4.3 gives an insight about the optimization process to tune the *SVM* parameters. Finally, the results of the experiments are shown in Section 4.4.

4.1 Dataset

In this work, the dataset which was proposed by Loni et. al [60] is used. It consists of 4810 JPEG images and the metadata associated included with them (if exists). Loni et. al collected the images from Flickr [61] by querying the list obtained from Wikipedia page *Index of Fashion Article* [62] which has 470 topics. Following this procedure, they used Flickr API for downloading the pictures up to top 1000 results. If the metadata such as internal id, title, owner, description, date taken, tags, comments etc. was present, it was also fetched.

For the annotation task, Amazon Mechanical Turk *(AMT)* [63] is used. *AMT* is a system which creates an environment where the individuals or the companies can meet with people to assign problems that cannot be solved by computers. In spite of the fact that the computers with artificial intelligence have been developed, there is still a long way before they can cope with the human beings in many tasks which require high-level semantic understanding. To solve this problem, *AMT* offers a practical system which is hiring people for carefully prepared small tasks and in return the people are paid according to the job they do. The individuals or the companies are called *The Requesters* whereas the people who do the job called *Turkers*. *The Requesters* create tasks which are called as *Human Intelligence Tasks (HIT)*.

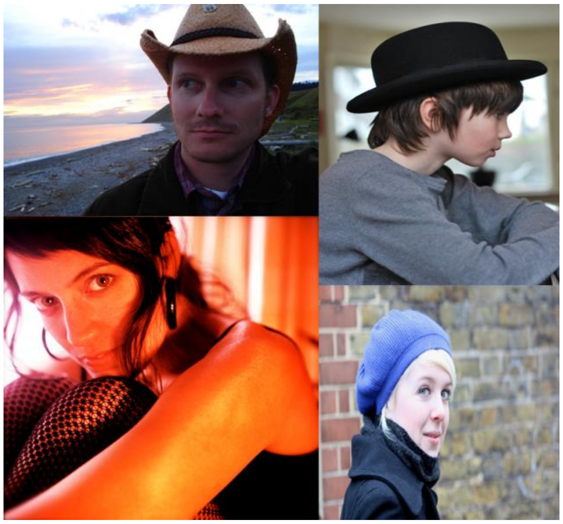


Figure 4.1 - Examples of filtered-out pictures with people

Since the dataset is not specifically created for the problem defined in this work, a filtering pre-process is necessary as there are many pictures which are out of our problem's scope.

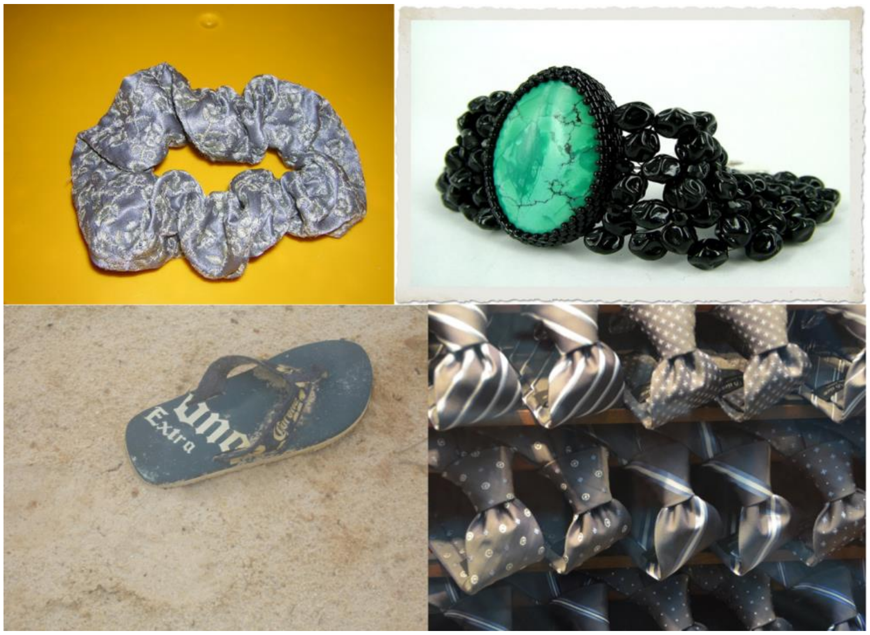


Figure 4.2 – Examples of filtered-out pictures without people

To illustrate this problem, Figure 4.1 and Figure 4.2 are presented. the pictures in Figure 4.1, include people; however, it is impossible extract the clothing information. Figure 4.2 contains some sample pictures which do not contain any person.

The filtering process is the following; firstly, the pictures annotated as the categories *Woman Informal, Woman Formal, Man Informal, Man Formal* are extracted. Then by visual inspection, the ones with occlusions and extreme rotations are excluded. In the end, a dataset with the following statistics is created;

- Woman Informal: 240 pictures
- Woman Formal: 180 pictures
- Man Informal: 191 pictures
- Man Formal: 142 Pictures

To avoid any biases during the training stage, the number of training samples from each category is chosen equally. On the other hand, it is also undesirable to have only a few samples left for the test process either because the maximum difference in terms of number of samples is 98 among the categories. Thus 75 samples from each category are randomly labeled as the training set and the rest are defined as the test set. The final test sets for each category are;

- Woman Informal (Test Set): 165 pictures
- Woman Formal (Test Set): 105 pictures
- Man Informal (Test Set): 116 pictures
- Man Formal (Test Set): 67 pictures

4.2 Performance Metrics

For measuring the success rate of the system, three different performance metrics which are *Precision, Recall, F – Measure and Error Rate is* used. Let us first define the standard evaluation metrics: *True Positive, True Negative, False Positive* and *False Negative* [64];

- *A true positive test result is one that detects the condition when the condition is present*
- *A true negative test result is one that does not detect the condition when the condition is absent*
- *A false positive test result is one that detects the condition when the condition is absent*
- *A false negative test result is one that does not detect the condition when the condition is present*

A summary of these definitions is given in Table 4.1.

Table 4.1 - Basic statistical definitions

		Condition	
		Present	Absent
Test	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Using this notation, precision is defined as;

$$Precision\text{:}\ \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives + Number\ of\ False\ Positives}$$

The interpretation of precision implies that it is a measure of correctly retrieved results among all the results retrieved.

$$Recall\text{:}\ \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives + Number\ of\ False\ Negatives}$$

On the other hand recall provides a measure for the proportion between relevant retrieved results among all the relevant samples.

$$F - Measure\text{:}\ 2 * \frac{Precision * Recall}{Precision + Recall}$$

F – Measure is the harmonic mean of the two metrics mentioned above.

$$Error\ Rate = \frac{Number\ of\ Correctly\ Classified\ Samples}{Number\ of\ Samples\ in\ the\ Test\ Dataset}$$

4.3 SVM Parameter Optimization

Radial Basis Function (RBF) kernel is chosen as the kernel for all the SVMs used in this work because *RBF* kernel can be used as a linear and sigmoid kernels by adopting certain parameters [65]. As it is explained in Section 3.1 and Section3.2.1, three different feature extraction methods are proposed in this thesis. In addition to that, each proposed method's classifier network is tuned in itself. Tuning process of each *SVM* is done separately by using Weka's *Grid Search* facility in which two parameters (cost and gamma) are optimized by applying 2-fold cross validation in a logarithmically scaled grid (base 2) using root mean squared error on training set.

As the authors of *LibSVM* suggest [66] , the search space is defined for cost as [-5.0, 15.0] and for gamma [-15.0, 3.0]. Thus in base 10, the search space is [2^{-5}, 2^{15}] and [2^{-15}, 2^{3}].

Optimal *cost* and *gamma* parameters found for *SVMs* are given below in Table 4.2, Table 4.3 Table 4.4;

Table 4.2 - SVM parameters for the classifiers used with Global Feature Extraction Approach

Global Approach	Cost HOG	Gamma HOG	Cost LBP	Gamma LBP	Cost CSD	Gamma CSD	Cost CLD	Gamma CLD
Woman - Man	2.00	0.03	16384.00	2.00	32.00	0.50	16.00	0.50
Woman Informal - Formal	4.00	0.016	2048.00	8.00	8.00	2.00	0.50	8.00
Man Informal - Formal	4.00	0.016	32768.00	0.25	2.00	1.00	0.50	8.00

It can be observed from Table 4.2 that except *LBP*, the cost parameter for the features in global approach are small enough to indicate a good generalization whereas vice versa is true for *LBP*. Bigger cost means that the number of support vectors which can be used to create the separating *hyperplane* are only a few as it is explained in detail in Section 2.2. This results with a narrow *hyperplane* which is expected to have a worse test performance. For gamma, *LBP* again fails to obtain a small number for *LBP* so it can be stated that the generalization ability of *LBP* classifiers in the global approach does not work well enough.

For *CLD*, even though the cost is reasonable, gamma parameter is on the border of the search space thus, it can be expected that it might overfit to the training data.

Table 4.3 - SVM parameters for Nine Grids Feature Extraction Approach

Nine Grids Approach	Cost HOG	Gamma HOG	Cost LBP	Gamma LBP	Cost CSD	Gamma CSD	Cost CLD	Gamma CLD
Woman - Man	2.00	0.03	16.00	2.00	2.00	0.125	16.00	0.25
Woman Informal - Formal	4.00	0.016	8.00	8.00	2.00	0.063	8.00	0.25
Man Informal - Formal	4.00	0.016	64.00	0.50	4.00	0.063	1.00	1.00

In Table 4.3, it is seen that costs for all four features are at a reasonable scale. However, gamma for *LBP's Woman Informal – Formal* case is on the border of the search space.

Table 4.4 - SVM parameters for Seven Grids Feature Extraction Approach

Seven Grids Approach	Cost HOG	Gamma HOG	Cost LBP	Gamma LBP	Cost CSD	Gamma CSD	Cost CLD	Gamma CLD
Woman - Man	4.00	0.03	128.00	0.25	2.00	0.063	8.00	1.00
Woman Informal - Formal	128.00	9.7656e-04	16.00	8.00	8.00	0.063	4.00	0.125
Man Informal - Formal	4.00	0.015	8.00	2.00	64.00	0.015	1.00	1.00

For *Seven Grids Feature Extraction Approach,* all the parameters have proper values thus overfitting is not expected.

4.4 **Results**

In this section, the results of the three different approaches, *Global Approach, Nine Grids Feature Extraction Approach and Seven Grids Feature Extraction Approach,* are discussed. First of all, the classification results according to each feature (*HOG, LBP, CSD, CLD*) are presented. Finally the merged results are given in terms of error rate. Since there is not any work in the literature that can be taken as the baseline according to the best of our knowledge, uniform distribution, 75.00% error rate (success rate) in four categories is decided to be used to compare this work.

4.4.1 Global Feature Extraction Approach

The results for each binary classification as it is proposed in Figure 3.5 **Error! Reference source not found.**are given in below;

Table 4.5 - The classification results for Global Feature Extraction Approach (Woman – Man)

Woman – Man	**Precision**	**Recall**	**F – Measure**
HOG	79.00 %	78.10 %	78.30 %
LBP	63.90 %	61.10 %	61.50 %
CSD	60.10 %	59.60 %	59.80 %
CLD	69.30 %	69.10 %	69.20 %

Table 4.6 - The classification results for Global Feature Extraction Approach (Woman Informal – Formal)

Woman Informal – Formal	**Precision**	**Recall**	**F – Measure**
HOG	61.90 %	58.90 %	59.40 %
LBP	68.30 %	65.20 %	65.60 %
CSD	55.30 %	53.30 %	53.90 %
CLD	65.70 %	66.30 %	65.80 %

Table 4.7 - The classification results for Global Feature Extraction Approach (Man Informal - Formal)

Man Informal – Formal	**Precision**	**Recall**	**F – Measure**
HOG	73.10 %	72.10 %	72.10 %
LBP	65.00 %	65.00 %	65.00 %
CSD	68.30 %	67.80 %	68.00 %
CLD	73.70 %	74.30 %	73.60 %

As it can be observed in Table 4.5, *HOG* is the most successful feature descriptor in *Woman – Man* classification with ~10% difference compared with the closest descriptor *CLD*. This makes sense because of the fact that our database consists of mostly frontal pictures with at least upper-body included. Thus *HOG* was expected to extract the information about the shape; in this case the silhouettes of the bodies on the pictures. Another distinguishable feature which was expected to be extracted by *HOG* is the difference between the hair styles of women and men. Even though a generalization such as *"women are supposed to have long hair and men are supposed to have short hair"* is not always valid, it is still a fair assumption which is expected to hold in most of the cases.

Table 4.6 shows that *LBP* can separate *Woman Informal – Formal* case the best in *Global Feature Extraction Approach*.

For *Man Informa*l – *Formal* problem, *HOG* and *CLD* show similar performances as it can be seen in Table 4.7. This can be explained with the expectation which is formal clothing of men are suits with homogenous color distribution whereas this distribution is much less consistent in informal case. On the other hand, the distinctive shape of formal clothing is also expected to be obtained by *HOG*. It can be also noticed that *CSD* is more successful in *Man Informal – Formal* category comparing with other two binary classifications which supports our claim about the difference between the color distributions of formal and informal clothes.

The final results of the individual features are given in Table 4.8. Despite the fact that merging the features does not provide a considerable increase, an almost 30 percent increase is achieved comparing with the baseline.

Table 4.8 - Final results per feature for the Global Feature Extraction

Final Results for Global Approach for Four Classes	Error Rate
HOG	47.46 %
LBP	60.71 %
CSD	64.90 %
CLD	52.10 %
HOG + LBP + CSD + CLD	46.58 %

4.4.2 Nine Grids Feature Extraction Approach

Table 4.9 - The classification results for Nine Grids Feature Extraction Approach (Woman – Man)

Woman – Man	Precision	Recall	F – Measure
HOG	79.00 %	78.10 %	78.30 %
LBP	71.30 %	70.40 %	70.60 %
CSD	63.40 %	60.70 %	61.0 %
CLD	72.90 %	72.20 %	72.40 %

Table 4.10 - The classification results for Nine Grids Feature Extraction Approach (Woman Informal – Formal)

Woman Informal – Formal	Precision	Recall	F – Measure
HOG	61.90 %	58.90 %	59.40 %
LBP	64.50 %	62.20 %	62.70 %
CSD	61.70 %	57.80 %	58.20 %
CLD	64.00 %	59.60 %	60.00 %

Table 4.11 - The classification results for Nine Grids Feature Extraction Approach (Man Informal – Formal)

Man Informal – Formal	Precision	Recall	F – Measure
HOG	73.10 %	72.10 %	72.40 %
LBP	69.80 %	70.50 %	69.90 %
CSD	70.60 %	71.00 %	70.80 %
CLD	75.00 %	75.40 %	75.10 %

In this case, it can be seen that *HOG* is again the most successful feature in *Woman – Man* classification problem as it is shown in Table 4.9. What is worth for noting is that the results of *LBP*, *CSD* and *CLD* improved when they are extracted in a grid manner. This can be explained with the fact that they obtain more enough information in the global case. When the sizes of the feature vectors (*N* vs. *9*N*) are considered, it is possible to obtain more details using the latter method because the resolution of feature extraction increases nine fold. Thus the features are able to capture more information than the global scenario. However, it should be also noted that it is possible to introduce more noise in this scheme.

On the other hand, in *Woman Informal – Formal* case, it is observed that the global features show superior performance (Table 4.10). *LBP* and *CLD* provide worse results in this approach comparing with the global approach whereas *CSD* increased more than 6 percent. However, it is also seen up to 4 percent increase when it comes to *Man Informal – Formal* problem. The reasons of the fluctuations between the grid based approaches and the global approach will be discussed in detail at the end of the results section.

Table 4.12 - Final results per feature for Nine Grids Approach

Final Results per Feature for Nine Grids Approach for Four Classes	Error Rate
HOG	47.46 %
LBP	54.30 %
CSD	58.50 %
CLD	50.77 %
HOG + LBP + CSD + CLD	50.55 %

Remarkably, the final results for merging the four features give worse results than the global approach even though the individual features performed better (Table 4.12). Overall, it can be concluded that the individual features' success rates are not necessarily to be correlated with the results of the merged features.

4.4.3 Seven Grids Feature Extraction Approach

Table 4.13 - The classification results for Seven Grids Feature Extraction Approach (Woman – Man)

Woman – Man	Precision	Recall	F – Measure
HOG	75.70 %	75.10 %	75.20 %
LBP	72.30 %	71.50 %	71.70 %
CSD	64.10 %	62.20 %	62.40 %
CLD	72.50 %	72.40 %	72.50 %

Table 4.14 - The classification results for Seven Grids Feature Extraction Approach (Woman Informal – Formal)

Woman Informal – Formal	Precision	Recall	F – Measure
HOG	62.90 %	60.00 %	60.50 %
LBP	67.40 %	64.40 %	64.90 %
CSD	59.00 %	56.30 %	56.80 %
CLD	59.20 %	57.80 %	58.20 %

Table 4.15 - The classification results for Seven Grids Feature Extraction Approach Seven Grids Approach (Man Informal – Formal)

Man Informal – Formal	Precision	Recall	F – Measure
HOG	72.00 %	72.10 %	72.10 %
LBP	68.80 %	68.90 %	68.80 %
CSD	68.80 %	67.80 %	68.10 %
CLD	77.40 %	77.60 %	77.50 %

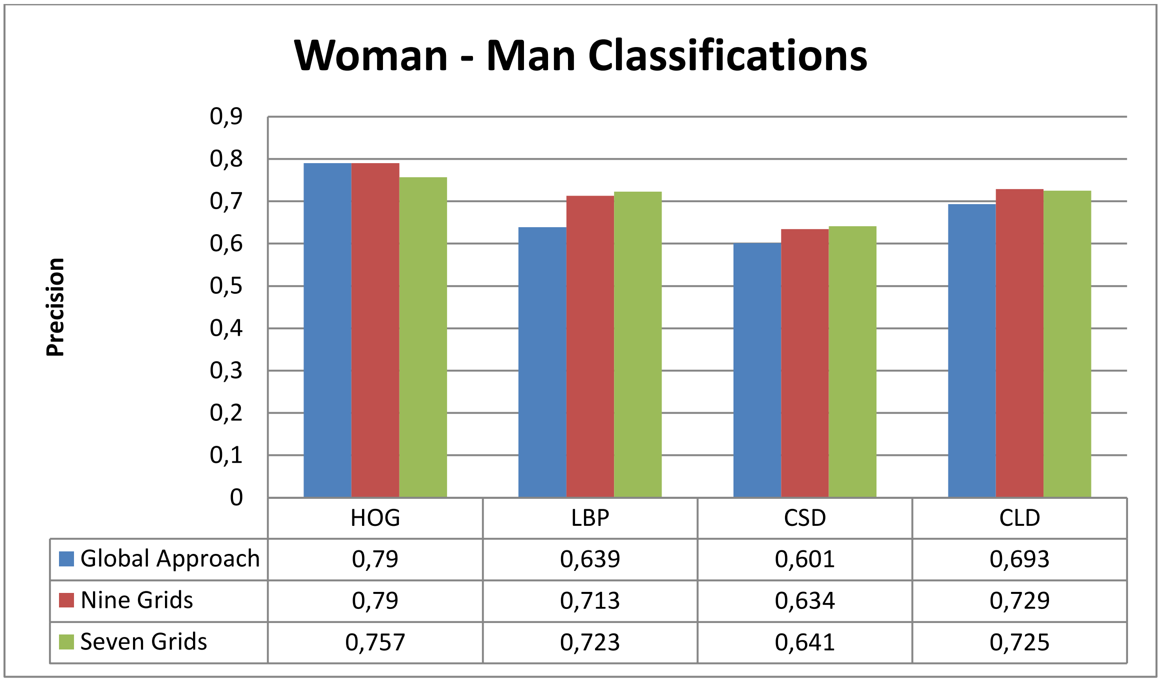
The results of this approach are given in Table 4.13, Table 4.14, Table 4.15, Table 4.16. As it can be observed, *HOG* provides superior results than the other three features for *Woman –Man* problem where the only feature which obtained under 70.00 % precision is *CSD.* When the *Woman Informal – Formal* problem is considered, it can be seen that *LBP* is clearly the best among the four with almost 5 percent better than the closest one. Finally for *Man Informal – Formal* case, a color descriptor, *CLD,* is the most successful feature with 77.40 % precision.

Table 4.16 - Final results per feature for Seven Grids Feature Extraction Approach

Final Results per Feature for Seven Grids Approach for Four Classes	**Error Rate**
HOG	51.43 %
LBP	54.53 %
CSD	60.26 %
CLD	52.98 %
HOG + LBP + CSD + CLD	45.25 %

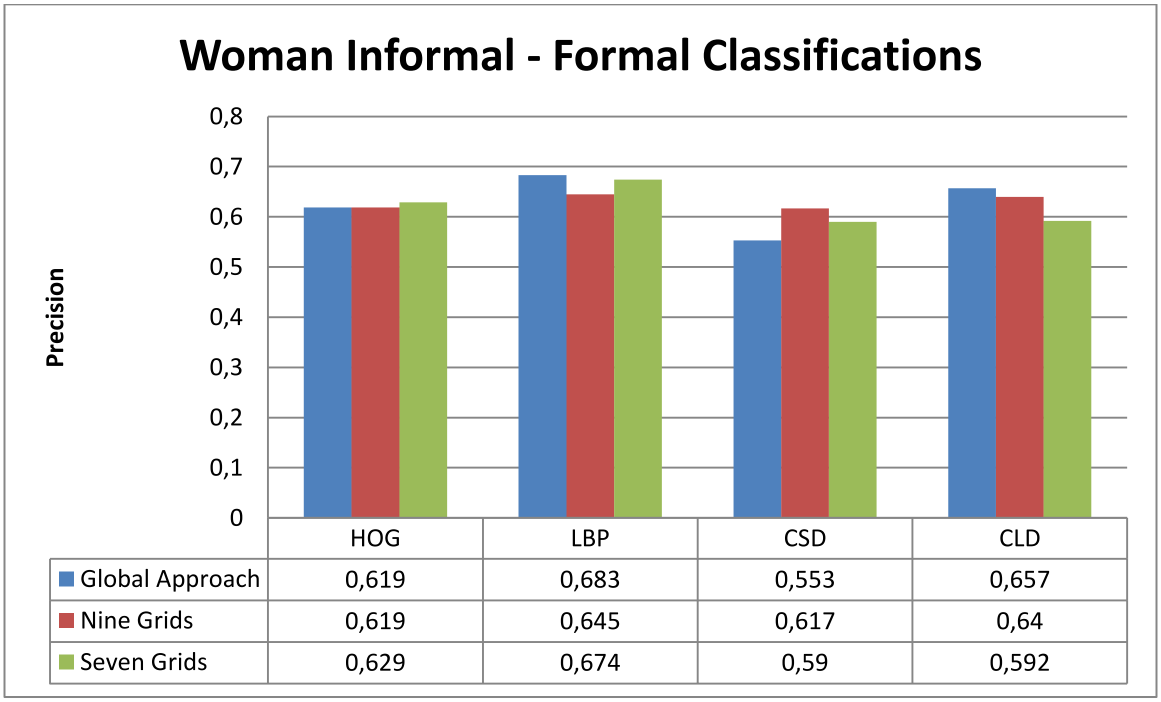
Final results for four categories show interesting results because even though individually there is not a big difference between the features, it gives the best result among three approaches proposed. It is 1 percent and 5 percent better than global and nine grids feature extraction approaches, respectively. This is possibly due to the fact that the individual features are able to fetch the information which complements each other.

In order to further visualize and compare the change in the individual classification results of the features, Figure 4.3, Figure 4.4 and Figure 4.5 are provided below. Note that these are different illustrations using the same data given above.



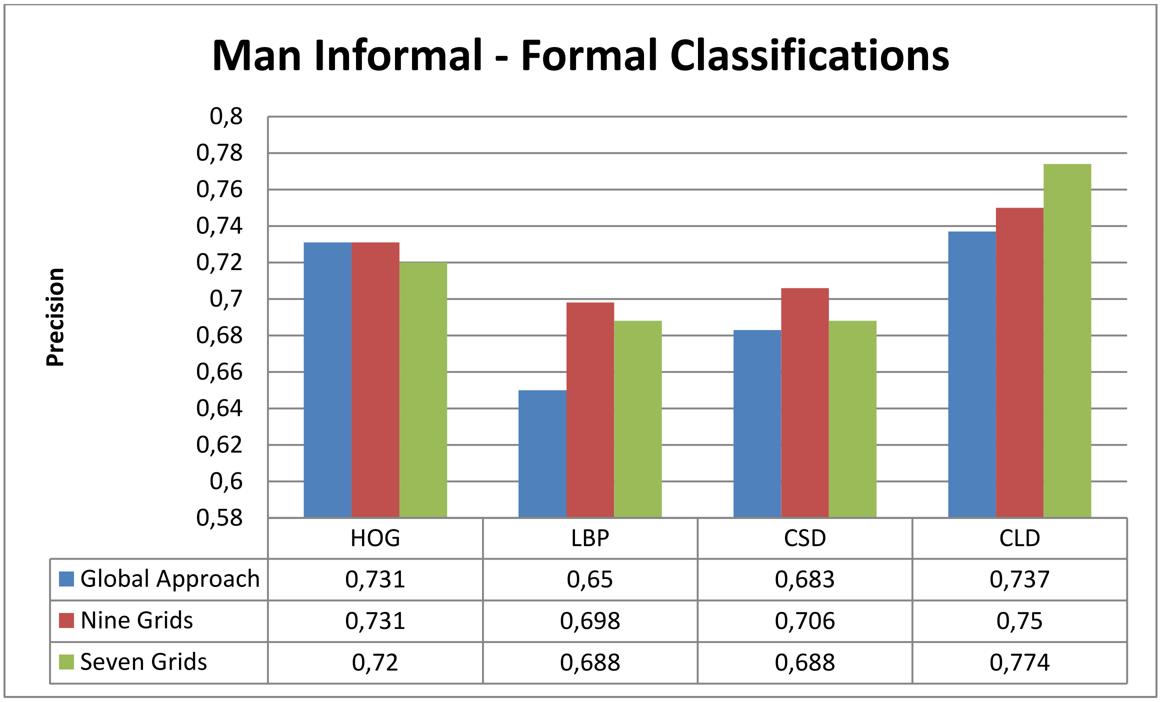
	HOG	LBP	CSD	CLD
🟦 Global Approach	0,79	0,639	0,601	0,693
🟥 Nine Grids	0,79	0,713	0,634	0,729
🟩 Seven Grids	0,757	0,723	0,641	0,725

Figure 4.3 - Comparison of the features in three approaches (Woman - Man)



	HOG	LBP	CSD	CLD
🟦 Global Approach	0,619	0,683	0,553	0,657
🟥 Nine Grids	0,619	0,645	0,617	0,64
🟩 Seven Grids	0,629	0,674	0,59	0,592

Figure 4.4 - Comparison of the features in three approaches (Woman Informal - Formal)



	HOG	LBP	CSD	CLD
🟦 Global Approach	0,731	0,65	0,683	0,737
🟥 Nine Grids	0,731	0,698	0,706	0,75
🟩 Seven Grids	0,72	0,688	0,688	0,774

Figure 4.5 - Comparison of the features in three approaches (Men Informal - Formal)

When the three approaches are analyzed, it can be observed that the grid based approaches generally tend to express the important information in the pictures better except *Woman Informal – Formal* case. One possible explanation for that would be in *Woman – Man* and *Man Informal – Formal* cases, the separation between the classes are clear and consequently more the information extracted, better the classification results. The classes are labelled homogenously and hence they are easier to learn as the amount of data provided increases. However, in *Woman Informal – Formal* case, local approaches brings more noise-like features because the amount of inconsistent local changes are more likely to occur in this problem.

5 Conclusions and The Future Work

In this thesis, three different algorithms are proposed for solving the four categories classification problem where the categories are defined as *Woman Informal, Woman Formal, Man Informal* and *Man Formal*. To evaluate the proposed approaches, experiments are conducted on *Fashion10000* dataset which is created by collecting the user taken images taken under unrestricted environments from Flickr. For the *Woman vs. Man* classification, the results reveal that *HOG* undoubtedly outperforms every other feature descriptor which is not surprising because the silhouettes of male and female bodies intuitively differ. Thus, adopting shape as a visual clue provides better results than texture and color. In the *Woman Informal vs. Woman Formal* case, the texture descriptor *LBP* gives consistently superior results even in the global feature extraction scheme. This result was expected because with a visual inspection, it can be seen that the variance between woman cloth styles can be distinguished by capturing the texture information rather than the shape. On the other hand, *Man Informal vs. Man Formal* case shows that *HOG* and *CLD* perform the best in all three approaches which can be interpreted as shape and color is vital for separating the formality of the man's clothing. This result overlaps with our visual observations as the common clothing item for man's formal scenario is suit with usually homogenously distributed dark colors.

When all the binary classifications, i.e. *Woman vs. Man, Woman Informal vs. Woman Formal* and *Man Informal vs. Man Formal* are compared, what is noticeable is *Woman Informal vs. Formal* classification results are worse than the other two binary classifications. This can be easily related with the observation that the semantic separation line between woman formal clothes and woman informal clothes is fuzzier. For example, in the man's scenario, one can tell whether a man is wearing formal clothes or informal clothes without hesitation in most of the cases because traditionally man's formal clothes are associated with suits. However, woman clothes tend to have more variety which affects the classification results negatively.

It should be noted that the images are taken without any restrictions. Thus, even though the dataset mostly consist of frontal poses, the pictures are not aligned. Additionally, the illumination conditions differ among the pictures. These factors can severely affect the effectiveness of the features extracted.

Among the three approaches proposed, the best results obtained for the four category classification by using *Seven Grids Feature Extraction Approach*. One reason for this result is due to the fact that this approach minimizes the amount of noise included in the images. In addition to that, since it is a grid based approach, it can also obtain the local information.

To summarize, our experiments showed that the shape is the most crucial information for separating women and men whereas the texture and the color are more suitable for understanding the formality of the clothes. In order to obtain better results, there are different possible approaches can be applied as the future work. Having more training samples is one of the important factors because if the classifiers can learn the properties of the classes better, it is likely that their performance will show a positive increase. Moreover, the neural networks can be applied instead of *SVM* if enough training samples can be collected.

Another method which might increase the results is instead of using low-level descriptors directly, attribute based approaches can be adopted where the clothing items are labeled with high-level concepts such as “Does this item have buttons ?”, “Does it have collar ?”. Using these high-level concepts, it can be easier to create the abstract relationships among the clothing types and the cloth itself. However, the trade-off for such an approach would be having a dataset with all the samples are labeled with the high-level concepts mentioned above.

References

[1] "Phrases.org," [Online]. Available: http://www.phrases.org.uk/meanings/a-picture-is-worth-a-thousand-words.html. [Accessed 18 February 2014].

[2] "FilmAndMedia," [Online]. Available: http://www.filmandmedia.ucsb.edu/academics/undergraduate/production/104/bare-bones1-20.pdf. [Accessed 18 February 2014].

[3] J. Grepstad, "Pinhole," 1996. [Online]. Available: http://photo.net/learn/pinhole/pinhole. [Accessed 18 February 2014].

[4] N. J. Wade, "Perception Web," 2005. [Online]. Available: http://www.perceptionweb.com/perception/perc0505/editorial.pdf. [Accessed 18 February 2014].

[5] "GTown Niepce," [Online]. Available: http://www.geog.ucsb.edu/~jeff/115a/history/niepce.html. [Accessed 18 February 2014].

[6] "UTexas Niepce," [Online]. Available: http://www.hrc.utexas.edu/exhibitions/permanent/firstphotograph/niepce/#top. [Accessed 18 February 2014].

[7] R. Dhaliwal, "The Guardian," 9 July 2013. [Online]. Available: http://www.theguardian.com/artanddesign/picture/2013/jul/09/first-colour-photograph. [Accessed 18 February 2014].

[8] "Zauberklang," [Online]. Available: http://zauberklang.ch/filmcolors/timeline-entry/1321/. [Accessed 18 February 2014].

[9] M. McFarlane, "Digital Pictures Fifty Years Ago," *Proceedings of the IEEE*, vol. 60, no. 7, pp. 768-770, 1972.

[10] R. C. Gonzalez and E. R. Woods, in *Digital Image Processing*, New Jersey, Prentice Hall, 2008, pp. 3-7.

[11] P. Felber, "Website of Illnois Institute of Technology," 2 May 2002. [Online]. Available: http://www.ece.iit.edu/~pfelber/ccd/project.pdf. [Accessed 18 February 2014].

[12] J. Good, "1000memories Blog," [Online]. Available: http://blog.1000memories.com/94-number-of-photos-ever-taken-digital-and-analog-in-shoebox. [Accessed 23 April 2014].

[13] "Merriam-Webster," [Online]. Available: http://www.merriam-webster.com/dictionary/classification. [Accessed 22 April 2014].

[14] M. J. Swain and D. H. Ballard, "Color Indexing," *International Journal of Computer Vision,* vol. 7, no. 1, pp. 11-32, 1991.

[15] Y. Lecun, "Gradient-based Learning Applied to Document Recognition," *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998.

[16] A. W. Senior and R. M. Bolle, New York.

[17] R. Chellappa, C. L. Wilson and S. Sirohey, "Human and Machine Recognition of Faces: A Survey," *Proceedings of the IEEE,* vol. 83, no. 5, pp. 705-741, 1995.

[18] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," in *Computer Vision and Pattern Recognition, 2001*, Kauai, 2001.

[19] S. Yang and B. Bhanu, "Understanding Discrete Facial Expressions in Video Using an Emotion Avatar Image," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 42, no. 4, pp. 980-992, 2012.

[20] B. Prasad and A. Krishna, "Classification of Medical Images Using Data Mining Techniques," in *Advances in Communication, Network, and Computing*, Springer Berlin Heidelberg, 2012, pp. 54-59.

[21] I. King, "A feature-based image retrieval Database for the Fashion, Textile, and Clothing Industry in Hong Kong," *Proceedings of International Symposium Multi-Technology Information Processing,* pp. 233-240, 1996.

[22] C.-I. C. Liu and D. S.-M. Liu, "An Intelligent Clothes Search System Based on Fashion Styles," in *2008 International Conference on Machine Learning and Cybernetics*, Kunming, 2008.

[23] X. Chao, M. J. Huiskes, T. Gritti and C. Ciuhu, "A Framework for Robust Feature Selection for Real-time Fashion Style Recommendation," in *Proceedings of the 1st International Workshop on Interactive Multimedia for Consumer Electronics*, New York, 2009.

[24] Q. Tu and L. Dong, "An Intelligent Personalized Fashion Recommendation System," in *2010 International Conference on Communications, Circuits and Systems (ICCCAS),* Chengdu, 2010.

[25] X. Wang and T. Zhang, "Clothes Search in Consumer Photos via Color Matching and Attribute Learning," in *Proceedings of the 19th ACM international conference on Multimedia - MM '11*, New York, 2011.

[26] Z. Song, M. Wang, X.-s. Hua and S. Yan, "Predicting Occupation via Human Clothing and Contexts," in *2011 International Conference on Computer Vision*, Barcelona, 2011.

[27] M. Yang and K. Yu, "Real-Time Clothing Recognition in Surveillance Videos," in *2011 18th IEEE International Conference on Image Processing*, Brussels, 2011.

[28] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz and T. L. Berg, "Parsing Clothing in Fashion Photographs," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, 2012.

[29] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-art Superpixel Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, no. 11, pp. 2274-2282, 2012.

[30] H. Chen, A. Gallagher and B. Girod, "Describing Clothing by Semantic Attributes," in *Proceedings of the 12th European conference on Computer Vision - Volume Part III,* Firenze, 2012.

[31] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu and S. Yan, "Street-to-Shop: Cross-Scenario Clothing Retrieval via Parts Alignment and Auxiliary Set," in *2012 IEEE Conference on Computer Vision and Pattern Recognition,* Providence, 2012.

[32] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack and L. V. Gool, "Apparel Classification with Style," in *Proceedings of the 11th Asian Conference on Computer Vision - Volume Part IV,* Daejeon, 2012.

[33] S. Ayache, G. Quénot and J. Gensel, "Classifier Fusion for SVM-Based Multimedia Semantic Indexing," in *ECIR'07 Proceedings of the 29th European conference on IR Research,* Rome, 2007.

[34] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Volume:1),* San Diego, 2005.

[35] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.

[36] T. Ojala, M. Pietikäinen and D. Harwood, "Performance Evaluation of Texture Measures with Classification Based on Kullback Discrimination of Distributions," in *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision &amp; Image Processing., Proceedings of the 12th IAPR International Conference on (Volume:1),* Jerusalem, 1994.

[37] J. Chen, V. Kellokumpu, G. Zhao and M. Pietikäinen, "RLBP: Robust Local Binary Pattern," in *{Proc. the British Machine Vision Conference (BMVC 2013), Bristol, UK,* Bristol, 2013.

[38] G. Zhao, T. Ahonen, J. Matas and M. Pietikäinen, "Rotation-Invariant Image and Video

Description with Local Binary Patterns," *IEEE Transactions on Image Processing,* vol. 21, no. 4, pp. 1465-1467, 2012.

[39] M. Heikkilä, M. Pietikäinen and C. Schmid, "Description of Interest Regions with Local Binary Patterns," *Pattern Recognition,* vol. 42, no. 3, pp. 425-436, 2009.

[40] X. Xianchuan and Z. Qi, "Medical Image Retrieval Using Local Binary Patterns with Image Euclidean Distance," in *Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on*, Wuhan, 2009.

[41] T. Ahonen, A. Hadid and M. Pietikäinen, "2006," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, no. 12, pp. 2037 - 2041, 2006.

[42] N. Jiang, J. Xu, W. Yu and S. Goto, " Gradient Local Binary Patterns for Human Detection," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, Beijing, 2013.

[43] G. Zhao and M. Pietikäinen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, no. 6, pp. 915 - 928, 2007.

[44] T. Ojala, M. Pietikänen and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 7, pp. 971-987, 2002.

[45] T. Sikora, "The MPEG-7 Visual Standard for Content Description - An Overview," *IEEE Transactions on Circuits and Systems For Video Technology,* vol. 11, no. 6, pp. 696 - 702, 2001.

[46] D. S. Messing, P. v. Beek and J. H. Errico, "The MPEG-7 Colour Structure Descriptor: Image Description Using Colour and Local Spatial Information," in *IEEE International Conference on Image Processing (ICIP 2011)*, Thessaloniki, 2001.

[47] L. Cieplinski, *The MPEG-7 Color Descriptors Jens-Rainer Ohm (RWTH Aachen, Institute of Communications Engineering).*

[48] E. Kasutani and A. Yamada, "The MPEG-7 Color Layout Descriptor: A Compact Image Feature Description for High-Speed Image/Video Segment Retrieval," in *Image Processing, 2001. Proceedings. 2001 International Conference on (Volume:1)*, Thessaloniki, 2001.

[49] W.-H. Chen, C. H. Smith and S. C. Fralick, "A Fast Computational Algorithm for the Discrete Cosine Transform," *IEEE Transactions on Communications,* vol. 25, no. 9, pp. 1004 - 1009, 1977.

[50] R. Balasubramani and Dr.V.Kannan, "Efficient use of MPEG-7 Color Layout and Edge Histogram Descriptors in CBIR Systems," *Global Journal of Computer Science and Technology,* vol. 9, no. 4, 2009.

[51] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995.

[52] A. Farhadi, I. Endres, D. Hoien and D. Forsyth, "Describing Objects by Their Attributes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition,* Miami, 2009.

[53] B. Willimon, I. Walker and S. Birchfield, "Classification of Clothing Using Midlevel Layers," *ISRN Robotics,* vol. 2013, 2013.

[54] C.-C. Chang and C.-J. Lin, "LIBSVM : A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology,* vol. 2, no. 3, pp. 27:1--27:27, 2011.

[55] A. Ben-Hur and J. Weston, "PyML - machine learning in Python," [Online]. Available: http://pyml.sourceforge.net/doc/howto.pdf. [Accessed 4 March 2014].

[56] G. Bradski, "OpenCV_Library," *Dr. Dobb's Journal of Software Tools,* 2000.

[57] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations,* vol. 11, no. 1, 2009.

[58] O. D. Team, "About OpenCV," [Online]. Available: http://opencv.org/about.html.

[59] D. A. Lisin, M. A. Mattar and M. B. Blaschko, "Combining Local and Global Image Features for Object Class Recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops 2005*, San Diego, 2005.

[60] B. Loni, M. Menendez, M. Georgescu, L. Galli, C. Massari, I. S. Altingovde, D. Martinenghi, M. Melenhorst, R. Vliegendhardt and M. Larson, "Fashion-Focused Creative Commons Social Dataset," in *MMSys '13 Proceedings of the 4th ACM Multimedia Systems Conference*, New York, 2013.

[61] "Flickr," [Online]. Available: https://www.flickr.com/. [Accessed 6 March 2014].

[62] "Index of Fashion Article," [Online]. Available: http://en.wikipedia.org/wiki/Index_of_fashion_articles. [Accessed 8 April 2014].

[63] "Amazon Mechanical Turk," [Online]. Available: https://www.mturk.com/mturk/welcome.

[64] "Test Statistics," [Online]. Available: http://groups.bme.gatech.edu/groups/biml/resources/useful_documents/Test_Statistics.pdf. [Accessed 6 March 2014].

[65] K.-P. Wu and S.-D. Wang, "Choosing the Kernel parameters of Support Vector Machines According to the Inter-cluster Distance," in *International Joint Conference on Neural Networks, 2006. IJCNN '06*, 2006.

[66] C.-W. Hsu, C.-C. Chang and C.-J. Lin, "A Practical Guide to Support Vector Classification," 15 April 2010. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf. [Accessed 3 March 2014].

[67] "San Jose State University," [Online]. Available: http://www.sjsu.edu/faculty/watkins/transist.htm. [Accessed 18 February 2014].

[68] J. Li and R. Abraham, "COBOL," 2002. [Online]. Available:

http://www.csee.umbc.edu/courses/graduate/631/Fall2002/COBOL.pdf. [Accessed 18 February 2014].

[69] "The History of the Integrated Circuit," 5 May 2003. [Online]. Available: http://www.nobelprize.org/educational/physics/integrated_circuit/history/. [Accessed 18 February 2014].

[70] "Intel," [Online]. Available: http://www.intel.com/content/www/us/en/history/museum-story-of-intel-4004.html. [Accessed 18 February 2014].

[71] "Image Net Large Scale Visual Recognition Challenge 2014," 2014. [Online]. Available: http://image-net.org/challenges/LSVRC/2014/index. [Accessed 20 February 2014].

[72] Princeton University, "About WordNet," 2010. [Online]. Available: http://wordnet.princeton.edu. [Accessed 20 February 2014].

[73] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, South Lake Tahoe, 2012.

[74] R. Eigenmann and D. J.Lilja, "TUE," 30 January 1998. [Online]. Available: http://www.idemployee.id.tue.nl/g.w.m.rauterberg/PRESENTATIONS/VON-NEUMANN-COMPUTER[2].PDF. [Accessed 18 February 2014].

[75] A. Jain, J. Huang and F. Shiaofen, "Gender Identification Using Frontal Facial Images," in *Multimedia and Expo, 2005*, Amsterdam, 2005.

[76] S.-L. Chang, L.-S. Chen, Y.-C. Chung and S.-W. Chen, "Automatic License Plate Recognition," *IEEE Transactions on Intelligent Transportation Systems,* vol. 5, no. 1, pp. 42-53, 2004.

[77] "Boost C++," [Online]. Available: http://www.boost.org/. [Accessed 27 February 2014].

[78] B. G. Dawes, "Boost C++ Initial Proposal," 6 May 1998. [Online]. Available:

http://www.boost.org/users/proposal.pdf. [Accessed 27 February 2014].

[79] Open Std, "Library Technical Report," 02 July 2003. [Online]. Available: http://www.open-std.org/jtc1/sc22/wg21/docs/library_technical_report.html. [Accessed 27 February 2014].

[80] "A Practical Guide to Support Vector Classification," [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

[81] "Support Vector Machines," [Online]. Available: http://www.tristanfletcher.co.uk/SVM%20Explained.pdf.

[82] A. Jeffries, "Engadget," 20 May 2013. [Online]. Available: http://www.theverge.com/2013/5/20/4341388/imrsv-rolls-out-cheap-face-detection-software-cara. [Accessed 1 April 2014].

[83] A. Vrankulj, "Biometric Update," 4 November 2013. [Online]. Available: http://www.biometricupdate.com/201311/tesco-uses-face-detection-to-target-ads-in-gas-stations. [Accessed 1 April 2014].

[84] "Creative Applications," [Online]. Available: http://www.creativeapplications.net/openframeworks/google-faces-searching-earth-using-facial-detection/. [Accessed 1 April 2014].

[85] R. Padilla, "Mac Rumors," 3 December 2013. [Online]. Available: http://www.macrumors.com/2013/12/03/apple-awarded-patent-detailing-facial-detection-and-recognition-system-for-devices/. [Accessed 1 April 2014].