ABSTRACT

OBJECT DETECTION AND ACTIVITY RECOGNITION IN DIGITAL IMAGE AND VIDEO LIBRARIES

by
Ibrahim Burak Ozer

This thesis is a comprehensive study of object-based image and video retrieval, specifically for car and human detection and activity recognition purposes. The thesis focuses on the problem of connecting low level features to high level semantics by developing relational object and activity presentations. With the rapid growth of multimedia information in forms of digital image and video libraries, there is an increasing need for intelligent database management tools. The traditional text based query systems based on manual annotation process are impractical for today's large libraries requiring an efficient information retrieval system. For this purpose, a hierarchical information retrieval system is proposed where shape, color and motion characteristics of objects of interest are captured in compressed and uncompressed domains. The proposed retrieval method provides object detection and activity recognition at different resolution levels from low complexity to low false rates.

The thesis first examines extraction of low level features from images and videos using intensity, color and motion of pixels and blocks. Local consistency based on these features and geometrical characteristics of the regions is used to group object parts. The problem of managing the segmentation process is solved by a new approach that uses object based knowledge in order to group the regions according to a global consistency. A new model-based segmentation algorithm is introduced that uses a feedback from relational representation of the object. The selected unary and binary attributes are further extended for application specific algorithms. Object detection is achieved by matching the relational graphs of objects with the reference model. The major advantages of the algorithm can be summarized as improving the object extraction by reducing the dependence on the low level segmentation process and combining the boundary and region properties.

The thesis then addresses the problem of object detection and activity recognition in compressed domain in order to reduce computational complexity. New algorithms for object detection and activity recognition in JPEG images and MPEG videos are developed. It is shown that significant information can be obtained from the compressed domain in order to connect to high level semantics. Since our aim is to retrieve information from images and videos compressed using standard algorithms such as JPEG and MPEG, our approach differentiates from previous compressed domain object detection techniques where the compression algorithms are governed by characteristics of object of interest to be retrieved. An algorithm is developed using the principal component analysis of MPEG motion vectors to detect the human activities; namely, walking, running, and kicking. Object detection in JPEG compressed still images and MPEG I frames is achieved by using DC-DCT coefficients of the luminance and chrominance values in the graph based object detection algorithm. The thesis finally addresses the problem of object detection in lower resolution and monochrome images. Specifically, it is demonstrated that the structural information of human silhouettes can be captured from AC-DCT coefficients.

OBJECT DETECTION AND ACTIVITY RECOGNITION IN DIGITAL IMAGE AND VIDEO LIBRARIES

by
Ibrahim Burak Ozer

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Electrical Engineering

Department of Electrical and Computer Engineering

January 2001

APPROVAL PAGE

OBJECT DETECTION AND ACTIVITY RECOGNITION IN DIGITAL IMAGE AND VIDEO LIBRARIES

Ibrahim Burak Ozer

Ali N. Akansu, Dissertation Advisor — Date
Professor of Electrical and Computer Engineering, NJIT

Wayne Wolf, Dissertation Co-Advisor — Date
Professor of Electrical Engineering, Princeton University, Princeton, NJ

Dr. Richard Haddad, Committee Member — Date
Professor of Electrical and Computer Engineering, NJIT

Dr. Bede Liu, Committee Member — Date
Professor of Electrical Engineering, Princeton University, Princeton, NJ

Dr. Yun-Qing She, Committee Member — Date
Associate Professor of Electrical and Computer Engineering, NJIT

BIOGRAPHICAL SKETCH

Author: Ibrahim Burak Ozer

Degree: Doctor of Philosophy in Electrical Engineering

Date: January 2001

Undergraduate and Graduate Education:

- Doctor of Philosophy in Electrical Engineering,
  New Jersey Institute of Technology, Newark, NJ 2001
- Master of Science in Electrical Engineering,
  Bogazici University, Istanbul, Turkey, 1995
- Bachelor of Engineering in Electrical and Communications Engineering,
  Istanbul Technical University, Istanbul, Turkey, 1993

Major: Electrical Engineering

Publications and Presentations:

Ibrahim Burak Ozer, Wayne Wolf, Ali N. Akansu
"Activity Recognition from MPEG Sequences", *IEEE Human Motion Workshop*, December 2000, Austin.

Ibrahim Burak Ozer, Wayne Wolf, Ali N. Akansu
"Relational Graph Matching for Human Detection and Posture Recognition", *SPIE, Photonic East 2000, Internet Multimedia Management Systems*, November 2000, Boston.

Ibrahim Burak Ozer, Mahalingam Ramkumar, Ali N. Akansu
"A New Method for Detection of Watermarks in Geometrically Distorted Images", *ICASSP 2000*, June 2000, Istanbul.

Ibrahim Burak Ozer, Wayne Wolf, Ali N. Akansu
"A Graph Based Object Description for Information Retrieval in Digital Image and Video Libraries", *Submitted to Journal of Visual Communication and Image Representation.*

Ibrahim Burak Ozer, Wayne Wolf, Ali N. Akansu

"A Graph Based Object Description for Information Retrieval in Digital Image and Video Libraries", *Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries*, Colorado, June 1999.

Ibrahim Burak Ozer, Isil Bozma and Bulent Sankur

"Discretization and Localization Of Defective Surfaces Using Deformable Surfaces", *4.Conf. on Signal Processing and Applications*, April 1996, Kemer, Turkey.

Ibrahim Burak Ozer, Isil Bozma and Bulent Sankur,

"Using Deformable Surfaces in Quality Inspection", *3.Conf. on Signal Processing and Applications*, April 1995, Kapadokya, Turkey.

To my parents and grandmothers

ACKNOWLEDGMENT

I express my sincere gratitude to my advisors, Prof. Ali N. Akansu and Prof. Wayne Wolf. I would like to thank Prof. Ali N. Akansu for the constructive atmosphere in our laboratory and for all his creative suggestions.

I would like to thank Prof. Wayne Wolf, who has supported me in every step of my thesis and whose guidance has been invaluable. It has been my joy and privilege to work with him.

I extend my sincere gratitude to Professors Richard Haddad, Bede Liu and Yun-Qing Shi for the stimulating discussions and suggestions. I would like to thank my professors in Turkey, who served as role models and motivated me to achieve my educational goals and Prof. Khalid Sayood, who played a major role in my decision to pursue my graduate studies in the United States.

I thank fellow researchers in our laboratory, for all the help they have offered towards completion of this thesis. I would especially like to thank Taha, Ramkumar, Selahattin, Alper, Xiaodong, Kadir and Zafer.

Personally, I would like to thank my family for always having confidence in my abilities and for their unconditional love, support, and friendship: my father, Selahattin, my mother, Semra and my sister, Berrak.

Finally, I would like to thank my wife and closest friend Şebnem Zorlu, for her understanding and for all her love and care throughout this process while we were both stressfully finishing our theses.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

Figure **Page**

LIST OF FIGURES (Continued)

CHAPTER 1

INTRODUCTION

With the rapid growth of multimedia information in forms of digital image and video libraries, there is an increasing need for intelligent database management tools. Although, the visual information is widely accessible, technology for extracting the useful information is still restricted. The traditional text-based query systems based on manual annotation process are impractical for today's large libraries requiring an efficient information retrieval system.

Multimedia information retrieval is a multidisciplinary area that is at the intersection of artificial intelligence, information retrieval, human interaction, and multimedia computing. It enables users to create, index, present, summarize, interact with (e.g. query, browse), and organize information within and across media such as text, audio, image, graphics, and video. Intelligent multimedia information retrieval includes those systems which go beyond hypertext environments.

A significant amount of effort has been devoted recently to develop content-based retrieval systems where the images/videos are indexed by their intrinsic visual features. The major blocks of the proposed retrieval system are feature extraction, object description, and retrieval engine.

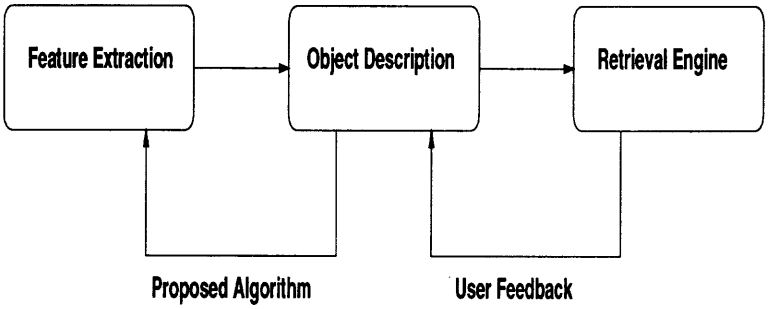


Figure 1.1 A general retrieval system.

Among those inter-dependent blocks, MPEG-7 (Multimedia Content Description Interface) activity is focused on the specification of a standard set of descriptors and description schemes. Automatic annotation of images where an object of interest is present faces two major problems. One is the dependence of the object description on the feature extraction process which is a complex task especially for cluttered scenes. The other is that the visual properties of images, that are described by feature vectors, are difficult to describe automatically with text. Therefore, the similarity retrieval connecting these vectors to high level semantics and using high level knowledge to improve feature extraction become an important issue.

This thesis is a study of theory and applications of object based information retrieval from compressed and uncompressed still images and video. A hierarchical retrieval system is proposed where the graph-based object detection is implemented in compressed and uncompressed domains. This hierarchical scheme enables working at different levels, from low complexity to low false rates. The finest details in the images and video sequences are obtained from the uncompressed domain via model based segmentation and graph matching for the analysis of cars and human bodies. The DCT coefficients and coefficient differences from JPEG images and MPEG sequences are used for object and activity detection. Available motion vectors are also used to detect human activity by comparing it with known human activity patterns.

The content is modeled by a hierarchical system (Figure 1.2) where the lowest level of information consists of pixels with color or brightness information. Features such as edges, corners, lines, curves and color/intensity regions are extracted next. In the higher level, these features are combined to describe objects and their attributes. The low level features that form the object descriptors are connected to the high level semantics via a graph-based description scheme, namely relational graph matching.

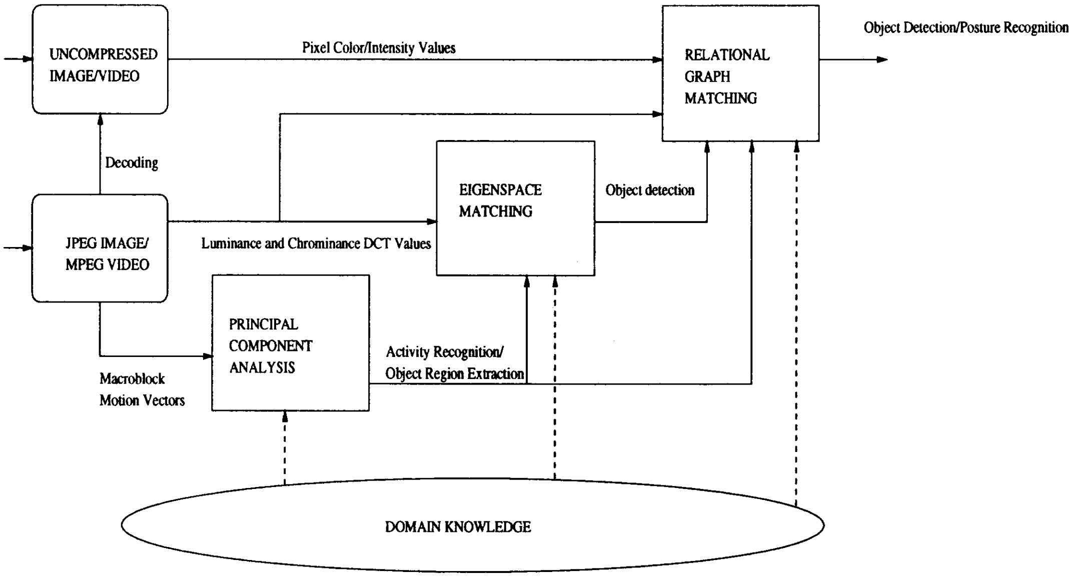

UNCOMPRESSED IMAGE/VIDEO
Pixel Color/Intensity Values
RELATIONAL GRAPH MATCHING
Object Detection/Posture Recognition
Decoding
EIGENSPACE MATCHING
Object detection
JPEG IMAGE/ MPEG VIDEO
Luminance and Chrominance DCT Values
PRINCIPAL COMPONENT ANALYSIS
Activity Recognition/ Object Region Extraction
Macroblock Motion Vectors
DOMAIN KNOWLEDGE

Figure 1.2 Overall algorithm.

The detailed algorithm of the graph matching process is given in Figure 1.3. Another important issue in digital libraries is the query representation which is related to the user interface. Query by example (QBE) is a method of query specification that allows a user to specify a query condition by giving image examples, such as a photo of the object in the database that contains the shape to be retrieved. Main features of an image can be given as shape, spatial relation, color and texture. Another method is to draw the shape of the object. Sketch based retrieval is a special case of shape retrieval. Here, the user describes a single object or a whole image by the layout of objects in it. Images are also retrieved by specifying colors and their spatial distribution in the image. User can specify the movement of an object for video retrieval. If textual descriptions representing the content of images are available then a query by keyword can be performed. The proposed retrieval system is used for video sequences, images and sketches enabling text based queries.

This thesis is organized as follows. Chapter 2 is a review of existing literature devoted to content-based retrieval systems. Shape similarity methods, namely,

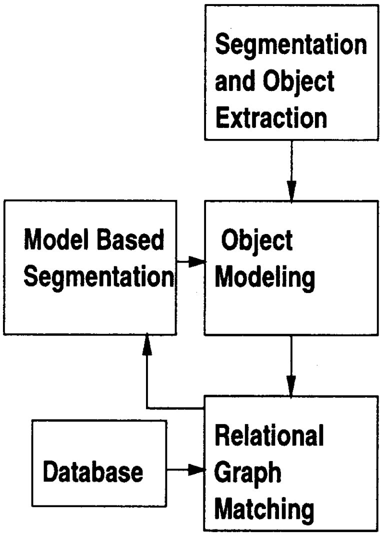


Figure 1.3 Relational graph matching algorithm.

contour and region based techniques, are also studied in this chapter. The theories of shape recognition postulate some form of internal and external representations for each object. The aim of the shape similarity method is to be able to correctly retrieve that shape during recognition. Current models introduce some form of representation which attempts to capture the supposed invariant properties of each object in various positions, sizes, rotations, and even under various lighting conditions. The representation that provides the best match using shape similarity measure is taken to be the object recognized.

The segmentation and object extraction are explained in chapter 3. First part corresponds to video applications where the moving rigid and non-rigid objects are extracted. Motion is a powerful cue used by human beings to extract objects of interest from a background of irrelevant detail. One of the approaches for detecting changes between two image frames is to compare the two images pixel by pixel which is computationally an expensive task. In our algorithm, instead of making a pixel by pixel comparison, only important feature points, which correspond to high activity regions and to sudden changes in images like corners and edges, are compared.

Next step is the segmentation of these rigid and non-rigid objects into smaller unique parts by region based segmentation, followed by curvature segmentation. In order to remove the noise and fine details, the boundaries of the region based segmented parts, are smoothed. Two smoothing techniques are studied: Gaussian approximation for human body parts, and line approximation for rigid bodies such as cars. For Gaussian approximation, a Gaussian kernel is used, which is suitable for smooth human body parts. Since the rigid body parts have sharp corners and sudden changes, the line approximation preserves the characteristics of the rigid body contours better than the Gaussian approximation. Curvature segmentation helps to partition complex shapes into more primitive ones, where the concave and convex segments are determined and used for segmentation. Last part in this chapter is the modeling of smoothed human body parts by superellipses. It is shown that 2D approximation of parts by fitting superellipses with shape preserving deformations provides satisfactory results for human detection.

Chapter 4 investigates the unary and binary shape attributes. For each segment the unary attributes; Hu moment invariants (only for rigid objects) , circularity, eccentricity, boundary shape code (only for rigid objects), color, and binary attributes; ratio of areas, relative position and orientation, adjacency information, are computed. The basic idea of moment invariants is to define a set of measures which are invariant to scale, rotation, translation, and contrast changes in a 2D plane. They are valid for recognizing a part from a particular pose which may be rotated, translated or scaled from the original training pose. The general shape attributes, such as circularity and eccentricity, are used to capture an intuitive measure of shape. Although the unary attributes are used to discriminate between different types of regions, often it is the relationship between the regions which identifies a particular class or structure. Relative position of one region with respect

.to the other one, relative size of two regions, and overlap of the boundaries of two regions are used as binary attributes.

Chapter 5 is devoted to graph matching. We study two graph matching algorithms for representation of complex objects. The chapter is split into two parts, where the graph matching algorithms for rigid and non-rigid objects are discussed separately. Graph matching algorithm (description scheme) combines low level features (descriptors) to high level semantics (car or human) by using a model based segmentation.

As the amount of information that is needed, increases, the need for compression increases as well. Chapter 6 addresses the problem of object and activity recognition in the compressed domain in order to reduce computational complexity and processing time. The first section of this chapter covers object detection in JPEG compressed still images, where the non-rigid and rigid objects are investigated in two separate subsections. In the subsection for non-rigid objects, possible human areas in the image are detected by using the JPEG coefficients and principal component analysis. The second part corresponds to the principal component analysis of MPEG motion vectors to detect the human activities; namely, walking, running, and kicking.

The performance of each algorithm block in the uncompressed domain and compressed domain results are given in Chapters 7 and 8 for rigid and non-rigid objects, respectively. Chapter 7 corresponds to the detection of car images in sketches, real images and video frames. The algorithm is also tested on quantized JPEG coefficients. The following chapter includes detection of human followed by posture recognition and activity detection in still images and video frames in uncompressed and compressed domains. Conclusions, and suggestions for future research are offered in Chapter 9.

CHAPTER 2

PREVIOUS WORK

Neural organization in the retina seems to be designed to provide information about the presence of discontinuities in the optical projection on the retina. It seems reasonable that the presence of borders, edges and contours in the stimulus would be the minimal information necessary for pattern perception since they could provide the building blocks for the perception of stable segregated portions of background and foreground [2]. Zusne [3] states the basic theories of visual form. The visual forms are transposable without loss of identity and will always be as good (regular, symmetric, simple, uniform) as the conditions allow. In our case, OOI (car or human) is the foreground object in an image or a moving object in a video frame. Car and human are complex objects formed by several simple visual parts (top and bottom parts of mainbody, windows, tires, etc. for car and head, torso, etc. for human).

The importance of high curvature points for visual perception is mentioned by many researchers. Hoffman and Richards [4] investigated the significance of corners for perception. Their main point is that one can represent common objects by first indicating points at which contours change direction and secondly connecting appropriate ones with a straight line. Another remark is the fact that one can sketch the essence of a thing with a very few lines separated at corners. Thus polygonal approximation of a contour after eliminating small discontinuities provides the essentials of an object shape. The learning aspect of perception is studied by Hebb [5]. He states that the organization and mutual spatial relationship of object parts must be learned for successful recognition. The learning of the shape of OOI is then related to the learning of the organization of simple visual forms that make up OOI with different attributes and spatial relationships among themselves.

2.1 Shape Retrieval

The search algorithms for the objects of interest related to shape similarity in a video or image library are implemented by various researchers. Shape based image retrieval is one of the hardest problems in general mainly due to the difficulty of segmenting objects of interest in images. The preprocessing algorithm determines the contour of an object depending on the application. Once the object is detected and located, its boundary can be found by using edge detection and boundary following algorithms [6]. The detection of the objects becomes a more difficult problem for complex scenes with busy background or many objects with occlusions and shading.

Once the object border is determined its shape can be characterized by its shape features. These feature vectors are generated by using a shape description method to characterize a shape. The required properties of a shape description scheme are invariance to translation, scale, rotation, luminance, and robustness to partial occlusion. Afterwards, shape matching is used in model-based object recognition where a set of known model objects is compared to an unknown object detected in the image using a similarity metric. Our description scheme is motivated by the well-known human perception theory and shape analysis techniques. The following subsections describe the related work.

2.1.1 Shape Analysis Techniques

Shape similarity methods can be classified into two parts namely contour and region based techniques. Birchfield [7] stated that every closed set in a plane can be decomposed into its two disjoint sets; the boundary and the interior according to elementary set theory. Since these two sets are complementary ("in the true, mathematical sense"), they claim that the failure modes of a tracking module focusing on the object's boundary will be orthogonal to those of a module focusing on the object's interior. Since the same concept can be applied to shape analysis, the combination of contour and region based shape descriptors are used in the proposed system.

2.1.2 Contour-based Techniques

For 1-D representation of shapes, Bennet and McDonald [8] use a tangent angle versus arc length function, that is also called turning function. The tangent angle at some point is measured relative to the tangent angle at the initial point. It is used by Arkin [9] for comparing polygonal shapes. The total turn (global curvature) is used for digital arcs by Latecki [10].

A signature of a boundary may be generated by computing the distance from the centroid to the boundary as a function of angle. Chang [11] constructs the distance function from the centroid to the feature points that are the points of high curvature. Template matching [12], chain coding [13], Fourier transform [14] and line segment moments [15] are other 1-D shape descriptors.

Another boundary representation technique is the curve approximation by utilizing polygonal and spline approximations. Polygonal approximations are used to approximate the shape boundary using the polygonal line. This is performed by using split-and-merge techniques based on some criteria. One approach is to merge points to form lines until exceeding a threshold. Bengston and Eklundh [16] proposes a hierarchical method where the shape boundary is represented by a polygonal approximation. Splines have been very popular for the interpolation of functions and the approximation of curves. They possess the beneficial property of minimizing curvature [17, 18].

Scale space techniques rely on the scale space representation. Witkin [19] proposes a scale space filtering approach which provides a useful representation for significant features of an object filtered by low-pass Gaussian filters of variable variance. Asada and Brady [20] introduce a new representation called the curvature primal sketch that is obtained by computing curvatures at different scales. Mokhtarian and Mackworth [21] uses the scale space approach as a hierarchical shape descriptor. Latecki [22] uses an approximation of the segment contours in order to distinguish perceptually similar shapes. The main advantage of discrete contour evolution is that it does not cause shape rounding (as in the case of Gaussian blurring).

2.1.3 Region-based Techniques

2-D moment based methods are among the most popular ones for regional descriptors [23]. The use of moments for shape description was proposed by Hu [24] who showed that moment based shape description is information preserving. An alternative transform approach is the Fourier transform of the shape. One of the disadvantages of these descriptors is that they do not reflect local shape changes. High-order features are required for shape classification. They are not robust to noise unlike low-order descriptors and are also computationally intensive. However, if the object is decomposed to its simpler forms, the low order moment invariants can be used for subparts of the object. Therefore, the result will be unaffected by a partial occlusion of object.

Some other simple region-based shape descriptors can then be used for these simpler forms: The area of a region is measured as the number of pixels contained within its boundary. Compactness (circularity) is defined as the ratio of squared perimeter to the area and the eccentricity is computed from the principal axes of the region. Medial axis transform first proposed by Blum [25] extracts a skeletal figure from the object and uses it to represent a shape by using a graph [26]. Leymarie and Levine [27] find the medial axis transform using snakes for active contour representation, high curvature points on the boundary, and symmetric axis transform. In shape decomposition techniques, a shape is represented as a combination of component shapes. The idea is to represent complex shapes in terms of simpler components. They can be approximated by well defined, similar shapes. For example, elliptical or a rectangular block can be used to represent an irregular shape. Superquadrics are widely used for modeling three dimensional objects in computer vision literature by Barr [29] and Bajcsy [30]. Bennamoun [31] and Boashash [32] use single test objects with a uniform background, and model object subparts with 2D superquadrics. Even when human body is not occluded by another object, due to the possible positions of non-rigid parts a body part can be occluded in different ways. Parametric modeling of image segments helps to overcome this problem and reduces the effect of the deformations due to the clothing. In the next section, some of the major content based image and video retrieval systems are described.

2.2 Retrieval Systems

Content based image/video indexing and retrieval has been researched by the governmental [33, 34] and industrial [35, 36] groups as well as at the universities [37, 38, 39, 40, 41, 42, 43, 44]. They use different techniques based on image features such as shape, color, texture, motion or a combination of them. A survey of these retrieval systems can be found in Yoshitaka [45] and Gupta [46]. Some of these systems, described below, support query by keyword representing a semantic concept.

One of the systems is the Photobook [47, 48] which is a software tool for performing queries on image databases based on image content and textual annotation. It basically compares features associated with images. The content descriptions, where one is dependent on appearance, another uses shape, and the third one is based on textural properties; are combined with each other and with text based descriptions.

Cypress-Chabot [49] integrates the use of stored text and other data types with content-based analysis of images to perform "concept queries". In Webseek [50], the images and video are analyzed using visual features (such as color histograms and color regions) and the associated text utilized to classify the images into subject classes.

SEMCOG [51] system performs a semiautomatic object recognition. SEMCOG (SEMantics and COGnition-based image retrieval) aims at integrating semantics and cognition-based approaches to give users a greater flexibility to pose queries. COIR (Content-Oriented Image Retrieval), an object-based image retrieval engine based on colors and shapes is used. The main task of the COIR is to identify distinct image regions based on preextracted image metadata, colors and shapes. Since an object may consist of multiple image regions, COIR consults to the image component catalog for matching image objects.

One of the commercial systems is QBIC [35], which supports several basic image similarity measures such as average color, color histogram, color layout, shape and texture. QBIC is a research prototype image retrieval system that uses the content of images as the basis of queries. Queries are posed graphically/visually, by drawing, sketching or by keywords.

"Car" and "Human" are the major objects of interest to be retrieved in the content-based retrieval systems. The systems, given below, cover different applications based on the extraction of these objects. Previous work on shape analysis of car objects was mostly based on boundary representation of objects. Dubuisson et al. [52] propose a segmentation algorithm using deformable template contour models to segment a vehicle of interest. Their goal is to determine the average travel time between two points in a road network by matching vehicles based on their color and shape attributes. They use five side view vehicle templates for classifying vehicle shapes where these templates are tested only for the side view car images with a stationary background. The results show that the classification depends on the detected edges. In another similar work [53], a vehicle is matched with a previously observed vehicle using color and shape features. Jain [12] proposes and tests a two dimensional shape matching and similarity ranking of still objects by means of a modal representation for car images. They employ selected boundary/contour points of the object with a coarse-to-fine shape representation. The algorithm is based on the contour detection of OOI. Xu [54] uses a hierarchical content description scheme and a hierarchical content matching technique for object retrieval. Experimental results are shown for a collection of car images. Papageorgiou et al. [55] use an overcomplete dictionary of Haar wavelets for identification of frontal and rear views of car in static images. Their method is based on the work presented by Papageorgiou [59].

Great effort has been devoted to human recognition related topics such as face recognition in still images, and motion analysis of human body parts. Most of the previous works depend highly on the segmentation results and mostly motion is used as the cue for segmentation [56]. There has been very few work that are on the human recognition in still images and in compressed domain. Although Franke [57] and Papageorgiou [59] use a compact representation of the training sets that are suitable for cluttered scenes there is no direct correspondence between the low level features and body parts. Such a semantic representation is needed for high level applications and for occlusion problems. In another survey by Gavrila [60], the segmentation problem is again pointed out especially for detection of multiple and occluded humans in the scene. Object detection in the compressed domain is more restricted since this application requires more detailed information. Schonfeld [61] proposes an object tracking algorithm by using compressed video only with periodically decoding I-frames. The object to be tracked is initially detected by an accurate but computationally expensive object detector applied to decoded I-frames. Zhong et al. [62] automatically localize captions in JPEG compressed images and I frames of MPEG compressed videos. Intensity variation information encoded in the DCT domain is used to capture the directionality and periodicity of blocks. Wang [63] proposes an algorithm to detect human face regions from dequantized DCT coefficients of MPEG video. This method is suitable for color images with face regions greater than 48 by 48 pixels (3 by 3 MPEG macroblocks). Previous work on color and motion retrieval techniques are given in the following subsections, respectively.

2.2.1 Color Retrieval

There are two approaches for querying by color: by regional color and by global color [64]. Regional color corresponds to spatially localized colored regions within the scenes. Global color corresponds to the overall distribution of color within the entire scene. Color information can be represented as color sets that give selection of colors or color histograms that denote the relative amounts of colors. Different color space bases related to human color judgments can be used [65]: HSV color space by Malik [66], Smith [40], and Yu [42], LUV color space by Moghaddam [67], YES color space by Saber [68]. Color models play an important role in extraction of skin regions for human detection systems [7, 69, 70, 71]. Texture characteristics of an image is mostly used with the combination of color attribute. Several texture models have been investigated by Picard [72] by pointing out potential uses in digital libraries.

2.2.2 Motion Retrieval

Motion is mostly used to index videos according to their activity levels, to detect shot and scenes in compressed and uncompressed domain [73, 74, 75]. Human motion analysis is another main research area that uses motion for information retrieval [56, 60, 77, 78, 79, 80, 76]. Motion extraction in compressed domain and human activity recognition are reviewed in more detail in Chapter 6.

Some of information retrieval systems allow the user to make a query using motion as the key object attribute [81]. An arbitrary polygonal trajectory is used for the query object. The temporal attribute defines the overall duration of the object, which can either be intuitive (long, medium or short) or absolute. The off-line process includes the decomposition of individual videos into separate shots. Then within each shot, video objects are tracked across frames.

Motion is also used for several video content-based retrieval systems for sports video. Kurokawa [82] retrieves scenes of soccer plays from several soccer video sequences. Motion is used to describe action of objects, interactions between objects and events using spatial and temporal relationships. Miyamori et al. [83] annotate tennis video where the court layout knowledge is used assuming that shots including tennis courts are preextracted. The ball and players are tracked by adaptive template matching. The players actions are detected using the transition of players' silhouettes. In Tan [84], the authors use camera motion to analyze and annotate basketball videos. Kobla et al. [85] automatically distinguish sports clips from other clips. They detect slow motion action replays in sports videos by detecting a repetitive pattern of a non-zero number of still frames being followed by a non-zero number of shift frames.

CHAPTER 3

SEGMENTATION

The purpose of image segmentation is to group pixels into regions that belong to the same object or object parts based on image homogeneity, e.g., color, texture, motion. An object can be found from appropriate grouping of object parts represented after a proper segmentation. Recognition is achieved by using the attributes of these groups. However, segmentation algorithms using only low-level features fail in most case due to the image noise, different illumination conditions, reflection and shadows. Although, some approaches use image features invariant to these conditions and improve the grouping results, they are not sufficient to detect complex objects. Consider to implement such an improved segmentation algorithm to segment a truck with a colored advertisement. It can only group regions of local consistency. The solution for an automatic object segmentation is to manage the segmentation process by using object-based knowledge in order to group the regions according to a global constraint. In this thesis, a new model-based segmentation, where global consistency is provided by using the relations of pixel groups, is proposed. These groups are obtained from the combination or further segmentation of group results of a low level segmentation algorithm. Managing the segmentation process using a feedback from relational representation of the object improves the extraction result even if its interior or its boundary is changed partially.

Our overall segmentation algorithm has three steps. The first step, moving object extraction for video sequences. The extraction algorithm presented in this chapter is a modified version of Kanade-Lucas-Tomasi's tracking algorithm. Output of this algorithm is a set of rectangular regions including moving objects where rest of the segmentation is implemented only in these bounding boxes. The second step is the model based segmentation. Color image segmentation is combined with an edge detector where small segments are removed. Resulting segments produced from this initial segmentation are combined by using a bottom-up control. In this chapter, it is shown that proposed model-based segmentation increases the overall algorithm performance by eliminating the segments that belong to the background. The last segmentation step, curvature segmentation, helps to get the primitive segments by dividing the complex object parts into simpler ones. Attribute calculation for these primitive segments reduces the computational complexity and increases the accuracy of the classification. Contour approximation for rigid objects and modeling by superellipses for non-rigid objects reduce the noise effects on the rigid object segments and disregard the deformations and occlusions on the human parts.

The contribution of the overall segmentation algorithm can be seen in guiding the segmentation process using a feedback from relational representation of the object. The major advantages can be summarized as improving the object extraction by reducing the dependence on the low level segmentation process and combining the boundary and region properties. Furthermore, the features used for segmentation (i.e. color, motion, curvature) are also attributes for object detection in relational graph representation. This property enables to adapt the segmentation thresholds by a model-based training system.

3.1 Motion Segmentation

This part corresponds to video applications where moving objects are extracted. In a video sequence, the feature points of an object are tracked based on Kanade-Lucas-Tomasi tracking method [86].

A point (x, y) in the first image I moves to point (W_x, W_y) in the second image J, where:

$$J(W_x(x,y), W_y(x,y)) = I(x,y) \tag{3.1}$$

$$W_x(x,y) = \sum_p^n \sum_q^n a_{pq} x^p y^q \tag{3.2}$$

$$W_y(x,y) = \sum_p^n \sum_q^n b_{pq} x^p y^q \tag{3.3}$$

Given the successive frames I and J, the problem is to find the parameters in the deformation matrix W and $\mathbf{d}$, where $\mathbf{d} = [a_{00} \quad b_{00}]^T$. The problem is the choice of the parameters that minimize the dissimilarity ϵ.

$$\epsilon = \int\int_W [J(W_x(x,y), W_y(x,y)) - I(x,y)]^2 \mathrm{dxdy} \tag{3.4}$$

where W is the given feature window. After Taylor series expansion, $\mathbf{d}$ is determined by solving the equation $Z\mathbf{d} = \mathbf{e}$ where:

$$Z = \int\int_W \begin{bmatrix} g_x^2 & g_x g_y \\ g_x g_y & g_y^2 \end{bmatrix} \mathrm{dxdy} \tag{3.5}$$

The eigenvalues of Z determine the selection of feature points, where d provides information about the displacement of the feature points in the second frame. The feature points with large eigenvalues correspond to high texture areas that can be matched reliably. These points are grouped according to their moving directions and distances (Figure 3.1). Only the feature points with a velocity greater than a given threshold are considered. Next step is the determination of a rectangular region of interest by calculating the center of gravity and the eccentricity of these groups. If the area of this region is smaller than a threshold defined by the maximum object size in the frame, this region is not processed. The output of this step is a rectangular region with an object of interest.

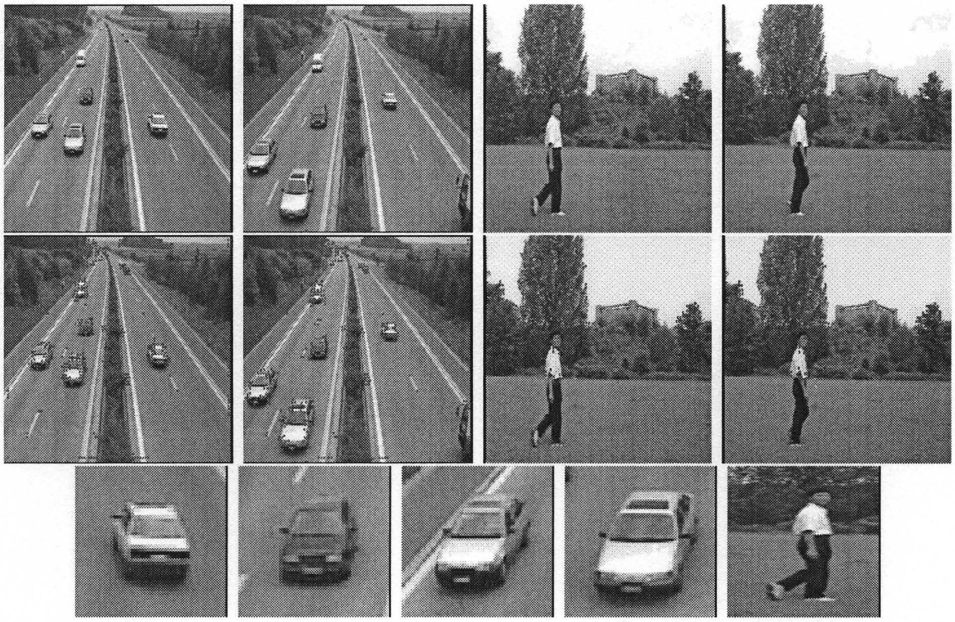


Figure 3.1 Extraction of moving objects in MPEG-7 video sequences. First Row: Initial and final video frames; Middle Row: Tracked features (motion threshold = 1pixel/frame, distance threshold = 15 pixels); Bottom Rows: Potential areas that contain OOI.

3.2 Region Based Segmentation with Model Based Segmentation

An object usually contains several sub-objects; such as wheels, windows, lights, etc. of a car or head, torso, limbs, etc. of a human, which can be obtained by segmenting the OOI hierarchically into its smaller unique parts. Here, the color image segmentation technique proposed in Harris [87] combined with an edge detector algorithm is used for rigid and non-rigid objects. For human detection, a skin color model is formed via Farnsworth nonlinear transformation.

The extraction of object of interest is a difficult task, especially in still images with a nonuniform background. As a result, the segmented image can contain regions corresponding to the background. However, these regions will not match to the regions of the template object. Semantic segments are created from the combination of low level edges or region based segments. If the object boundaries were segmented accurately, the shape descriptors for each object part could give satisfactory results for shape retrieval. However, a general automatic object segmentation without any user interface is almost impossible due to the illumination changes, shadows and occlusions especially for still images. Although using features invariant to illumination or reflection can improve the segmentation results, it is still not enough alone. Notice that handling the illumination and reflection changes in the car images is a more difficult task than in the human images.

Prior knowledge about the object to be retrieved should be used to segment the regions properly. One method is to perform rigid and deformable model based segmentations [89, 90, 91, 92]. The latter work differs from the previous works by enforcing global consistency. Local and global constraints should be used together for a segmentation that is robust to occlusions and variations in object shapes. These approaches try to extract the object boundary. Our approach differs from them at this point and will be explained in the next sub-section.

3.2.1 Proposed Model Based Segmentation

The combination of features related to the boundary and interior of the object along with the relationships between the parts is more robust since the other one works when one fails. For this reason, the proposed method and segmentation procedure are implemented iteratively. Closed regions are defined and small ones are removed. For each segment and the combinations of these segments formed by merging them according to the adjacency information, the attributes (unary and binary), that are given in detail in graph matching subsection, are computed. For comparison of the test and model data, the graph matching algorithm is implemented and the number of regions, that are matched, is checked. If sufficient number of regions is matched the unmatched regions are removed and the object regions are extracted.

The drawback of this approach for on-line applications is the computation of the various combinations of regions to be merged. The idea is to merge the neighboring regions. However, the connectivity constraint alone is not sufficient since testing all combinations to select the best match is impractical due to the computational complexity. Assume that one has four regions with the following structure; region 1 is a neighbor of region 2, region 2 is a neighbor of regions 1 and 3, and region 3 is a neighbor of regions 2 and 4. The possible combinations are (1,2), (1,2,3), (1,2,3,4), (2,3), (2,3,4), (3,4) (Figure 3.2). This number will increase exponentially with the number of regions under consideration. One way to handle this is to constrain the color difference between regions. However, although the natural object color does not change significantly (e.g., skin, fruits, animals, trees) it is not always true for objects such as cars. A meaningful part of a car can have several color components. For example, the mainbody of the car can be multi-colored. Best-first, or highest confidence first algorithms decrease the complexity [92] but also degrade the performance. For human images, a meaningful combination is the combination of adjacent segments on the same principle axis. For example, upper arm of a person with a shirt can be segmented into two parts, however it should be the combination of clothed and naked regions. The opposite of this example can also occur e.g., color and curvature segmentation can fail to segment arms from torso.

The segmented region boundaries can still be in complex forms. The boundaries are first smoothed. Concave and convex segments (landmarks) that are used for curvature segmentation are determined on the resulting contour. The main reason for finding boundary landmarks is that they can be used to partition complex parts into different domains. For example, these landmarks are used to partition the mainbody of a car into two subparts as well as partition the arm into upper-arm and lower-arm. However, color and curvature segmentation can fail in extracting the desired object

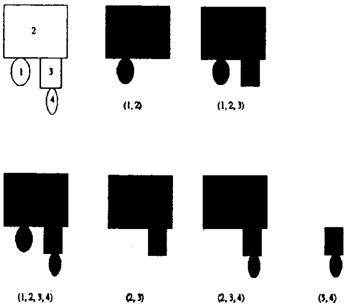


Figure 3.2 Combination of the example object segments.

parts. For example, two adjacent object parts in the image might correspond to one node in the model image. It is shown that this segmentation effect is removed by using possible combinations of the object parts. Curvature segmentation is explained more detailed in the following section.

3.3 Curvature Segmentation and Contour Approximation

Before the curvature segmentation, the contours with very small local deformations must be smoothed. In this subsection, two different contour approximation techniques are studied, Gaussian based smoothing and line approximation, that are followed by curvature segmentation. Both of the techniques are tested on the rigid car body parts. Since the parts include sharp corners and sudden changes, the line approximation preserves the contour information better than the Gaussian approximation. On the other hand, Gaussian smoothing is suitable for smooth human body parts in order to reduce the effect of image noise and clothing.

3.3.1 Gaussian Based Smoothing

The contour shape analysis is implemented to extract the convex parts of objects that determine visual parts separated by concavities. A method is the smoothing of

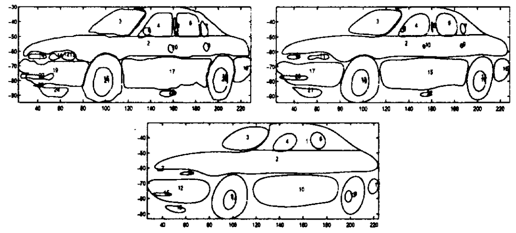


Figure 3.3 Multiscale representation of car segments. Top left: σ of the Gaussian kernel $= 1.5$, Top right: $\sigma = 5$, Bottom: $\sigma = 10$.

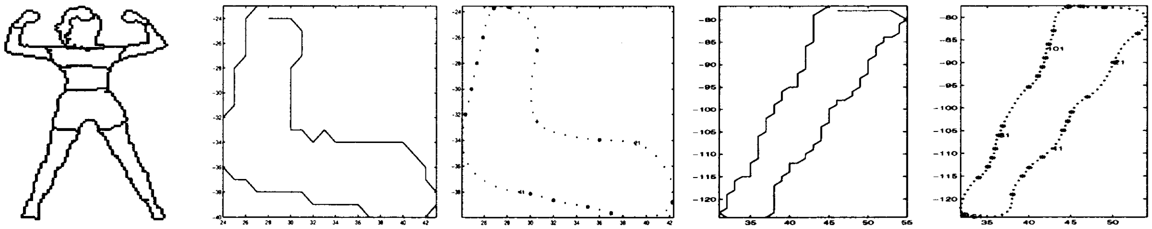


Figure 3.4 Gaussian smoothing results for the arm and leg segments of the example human body with the landmarks.

the boundaries by using a 1-D Gaussian kernel and then calculating the curvature of each boundary point [19]. The width of the kernel defines the scale at which curvature is estimated. Figure 3.3 shows a multiscale representation of car segments. The noise and fine details are smoothed at large width, leaving distinct extrema at positions of perceptually significant points on the boundary. These points are called "landmarks". Figure 3.4 shows the Gaussian smoothing result for the human body part. As an example, the arm and leg segments are smoothed with a Gaussian kernel and the landmarks are defined. Next step is the curvature segmentation regarding to these landmarks.

After the Gaussian smoothing operation, the concave points with high curvature K_s (greater than a threshold th_k) and arc lengths (greater than a threshold th_s relative to the segment length) are marked. A normal line is computed from this landmark until it reaches another point on the contour. Then, the segment is divided at these points and an interpolation is performed between these points to form closed

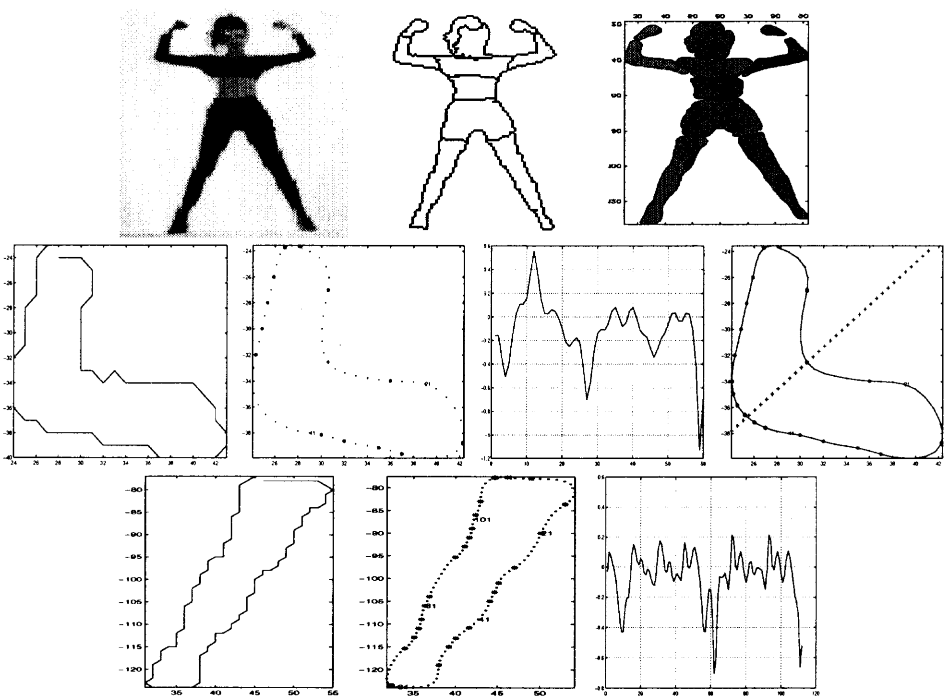


Figure 3.5 Top: First: Original image. Second: Segmentation result. Third: Curvature segmentation results. Middle: First: Arm segment. Second: Smoothed contours with landmarks ($th_k = 0.55$). Third: Curvature points. Four: Curvature segmentation. Bottom: First: Leg segment. Second: Smoothed contours with landmarks ($th_k = 0.55$). Third: Curvature points.

segments. As expected, experimental results show that the high curvature locations occur at the joints on the limbs. Since human body parts are smooth objects the smoothing factor is chosen very small (= 1.25). Curvature threshold is chosen the same for all the test images (= 0.55) and arclength threshold is 20%. In Figure 3.5, the curvature segmentation result for selected body parts is shown. Note that, since the arc length at the junction of the legs (belly) is small relative to the whole segment length, this part is not segmented. The graphs, given in Figure 3.5, show the curvature points. For the arm segment, there is one concavity point which is greater than the curvature threshold while for the leg segment, all the concave points are below this threshold. Figure 3.6 displays another example from a MPEG7 test sequence.

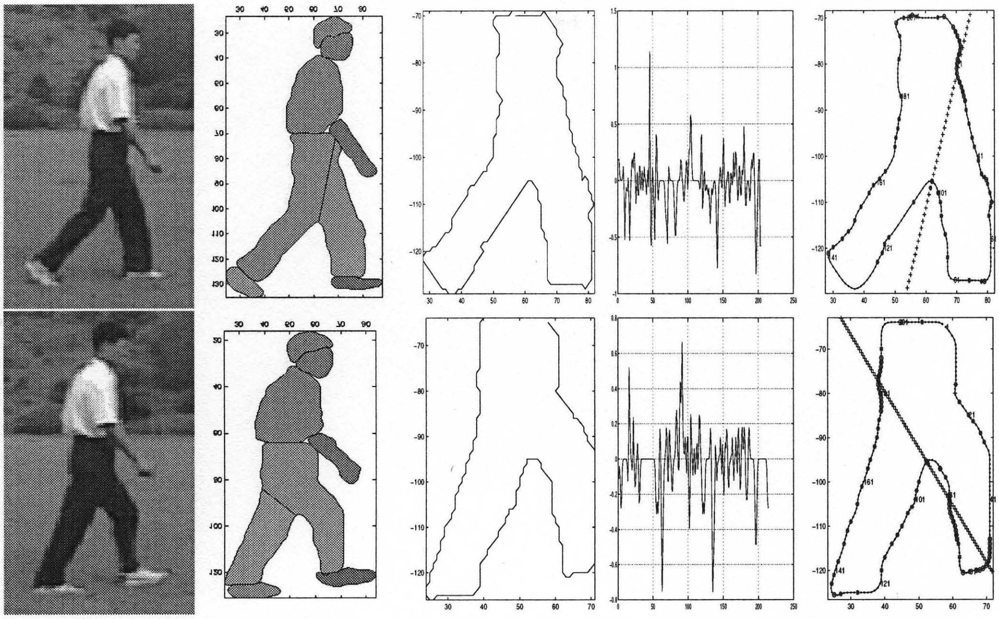


Figure 3.6 First column: KLT algorithm result for the MPEG7 test sequence. Second column: Segmentation results. Third column: Leg segment. Fourth column: Curvature of the segment($th_k = 0.55$). Fifth column: Curvature segmentation.

3.3.2 Line Approximation

In this method, digital curves which are composed of digital line segments are used. The idea is to decompose the digital curve into maximal digital line segments. In every evolution step, two consecutive line segments are replaced with a single line segment [22]. If the evaluation is continued, the curve shape will be simplified. For each line segment pair the cost function is calculated and consecutive line segments with the minimum cost function are replaced with a single one. For each adjacent line segment pair $s1$ and $s2$, the cost function $K(s1, s2)$, which represents the significance of the contribution of arc $s1 \cup s2$ to the shape of digital curve C, is determined (Figure 3.8) using Eq. (3.6):

$$K(s_1, s_2) = \frac{\beta(s_1, s_2) l(s_1) l(s_2)}{l(s_1) + l(s_2)} \tag{3.6}$$

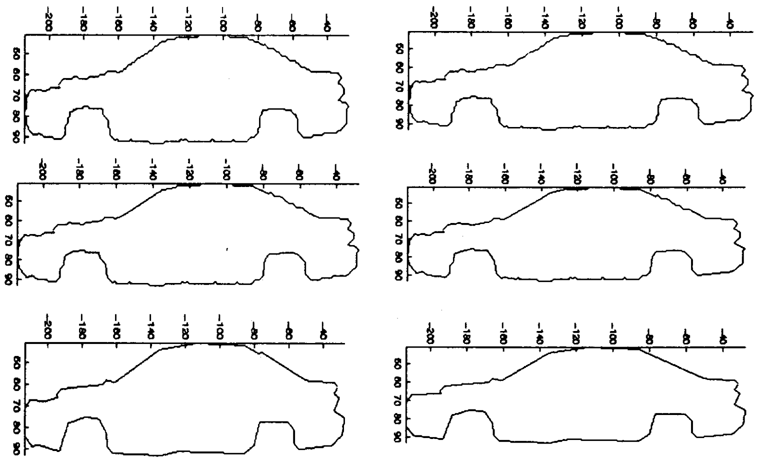


Figure 3.7 Selected stages of the discrete curve evaluation for the car mainbody.

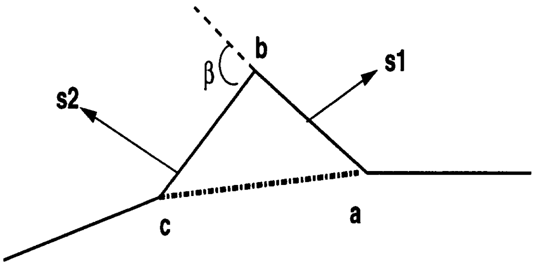


Figure 3.8 Curve evolution step.

In Eq. (3.6), l is the length function normalized with respect to C. $s_1 = \overline{ab}$ and $s_2 = \overline{bc}$ are the two adjacent line segments in the decomposition of curve C, so that b is their common edge point and $\beta = \beta(s_1, s_2)$ is the turn angle. If the cost function is above the threshold the line pair $s1$ and $s2$ are replaced by the single line $\overline{ac}$. This process continues until the cost function for each pair is above the threshold. In our experiments the threshold is 0.01. It is argued that parts are generally defined to be convex or nearly convex shapes separated from the rest of the object at concavity extrema. The length of these concave and convex line segments as well as the angle between the corresponding lines (turn angle) are used as descriptors.

An example for the discrete curve evaluation is given in Figure 3.7. Notice that the localization is not preserved in Figure 3.3 as in Figure 3.7 due to the global,

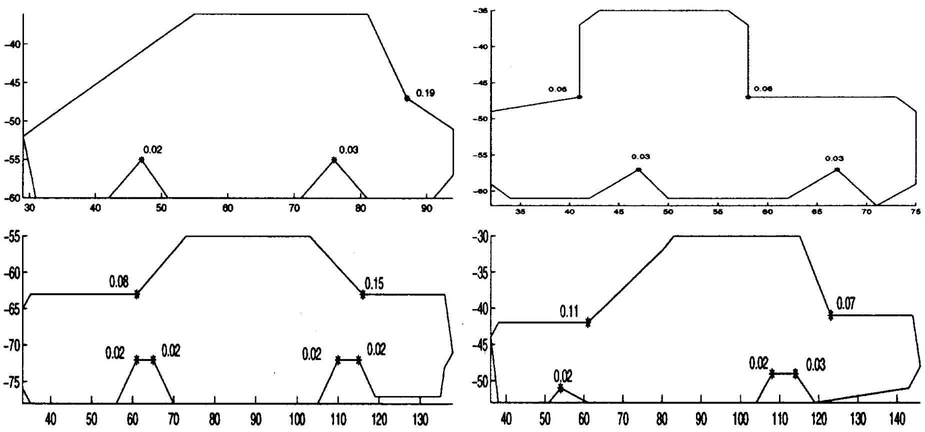


Figure 3.9 Concave points of mainbody for sketches.

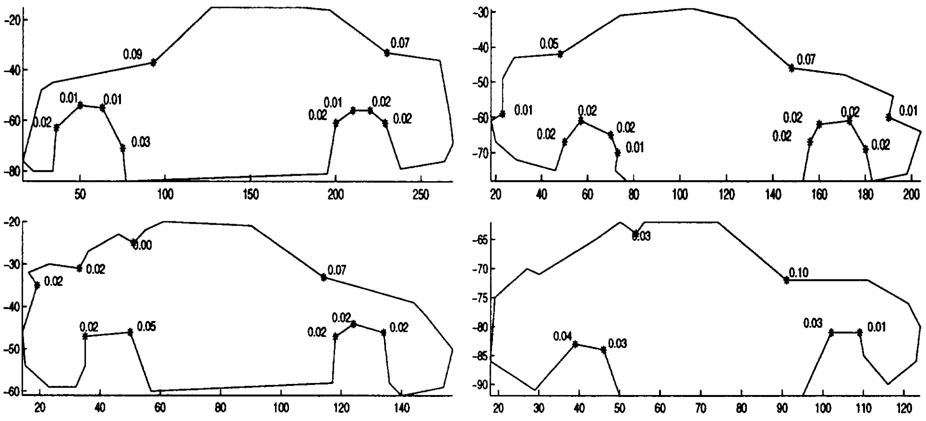


Figure 3.10 Concave points of mainbody for real images from side view.

independent smoothing of the spatial components in the former case. The concavity measures that are computed from the normalized length and angle of the concave axes are displayed in Figures 3.9 and 3.10 for sketches and real images, respectively. The highest concavity points correspond to the landmarks for the two main subparts of the mainbody. Note that this feature can also be used for the ranking purposes, e.g. sport cars with hatchback have one maximum concavity point while sedan type cars have two main concavity points of a similar order. Other major concave axes on the mainbody correspond to the location of tires where adjacent concavities with a similar cost function K are observed.

3.3.3 Surface Approximation (Modeling by Superellipses)

Even when human body is not occluded by another object, due to the possible positions of non-rigid parts a body part can be occluded in different ways. For example, hand can occlude some part of torso or legs. The contour approximations, used for rigid objects, is not efficient in this case and the combination of occluded part with hand is not meaningful. However, 2D approximation of parts by fitting superellipses with shape preserving deformations provides more satisfactory results. It also helps to disregard the deformations due to the clothing. Result of the global approximation which do not capture local deformations seems more appropriate for human body. Hence, instead of using region pixels it is better to use parametric representations to compute shape descriptors. In a similar work by Bennamoun et al. [31], a simple vision system, where the objects are modeled by superellipses, is proposed. Since their system performance highly depends on the initial segmentation results, they use single test objects with uniform backgrounds. The recognition stage compares the angles of the test object skeleton with the library object skeleton and decides if the same object is present in the library. Their algorithm can only be used for non-occluded objects with a certain orientation, where our system can overcome the initial segmentation problem with the model based segmentation, can work for occluded images without any orientation constraint and combine the object parts via graph matching algorithm and decide the human presence. The detailed procedure for superellipse description and fitting procedure is given below.

A superellipse can be described explicitly as:

$$x = f_x(\eta) = a_x cos(\eta)^\epsilon \tag{3.7}$$

$$y = f_y(\eta) = a_y sin(\eta)^\epsilon \tag{3.8}$$

In these equations, $-\pi < \eta < \pi$, a_x and a_y are two semi-axis, and ϵ is the roundness parameter. The curve intersects the x axis at a_x and $-a_x$ and intersects the y axis at a_y and $-a_y$. The inside-outside function of a two dimensional superquadric can be given as:

$$(\frac{x}{a_x})^{2/\epsilon} + (\frac{y}{a_y})^{2/\epsilon} = f(x, y, \mathbf{a}) \tag{3.9}$$

where **a** is the parameter set. There can be various deformations that can be implemented on the superellipses. Tapering and bending are sufficient deformations to represent human body. However, when for example legs are wide open they have to be segmented since no shape preserving deformation can represent them. Tapering along the y-axis is:

$$X = (\frac{K}{a_y} + 1)x \tag{3.10}$$

$$Y = y \tag{3.11}$$

where K is a constant. Circular bending:

$$X = x + sign(b)(\sqrt{y^2 + (a_y/b - x)^2} - (a_y/b - x)) \tag{3.12}$$

$$Y = sin(atan(y/(a_y/b - x))(a_y/b - x) \tag{3.13}$$

In these equations, b is the bending parameter, (X, Y) are the deformed (x, y) values where

$$(D \circ R \circ T)(x, y) \rightarrow (X, Y) \tag{3.14}$$

where D =Deformation, R =Rotation, T =Transformation.

In order to find superellipse parameter set $\mathbf{a} = [a_x, a_y, \epsilon, K, b, \theta, p_x, p_y]$, that fits best to the segment data (X, Y), Levenberg-Marquardt method is used [93] for nonlinear parameter estimation. First, the initial parameter set is used to find non-deformed world centered superellipse $(\overline{x}, \overline{y})$

$$(D \circ R \circ T)^{-1}(X, Y) \rightarrow (\overline{x}, \overline{y}) \tag{3.15}$$

The model to be fitted, the inside-outside function $f(\overline{x}, \overline{y}, \mathbf{a})$ forms the merit function χ in order to determine best fit parameters by its minimization. With nonlinear dependences, the minimization must proceed iteratively. The procedure is repeated until χ^2 stops decreasing.

$$(\frac{\overline{x}}{a_x})^{2/\epsilon} + (\frac{\overline{y}}{a_y})^{2/\epsilon} = 1 \tag{3.16}$$

$$\chi^2(\mathbf{a}) = \sum_{i=1}^{N}(1 - f(\overline{x}, \overline{y}, \mathbf{a}))^2 \tag{3.17}$$

The initial parameter set is taken as following:

$$\epsilon = 1 \quad K = 0 \quad b = 0 \tag{3.18}$$

$$px = 1/N \sum_{i=1}^{N} X_i \quad py = 1/N \sum_{i=1}^{N} Y_i \tag{3.19}$$

The initial values of $\theta, a_x,$ anda_y take different values regarding to the central moment μ.

If $\mu_{11} = 0$ and $\mu_{20} > \mu_{02} \rightarrow \theta = \pi/2, a_x = 2\mu_{20}^{1/2}, a_y = 2\mu_{02}^{1/2}$

If $\mu_{11} = 0$ and $\mu_{20} <= \mu_{02} \rightarrow \theta = 0, a_x = 2, \mu_{20}^{1/2}, a_y = 2\mu_{02}^{1/2}$

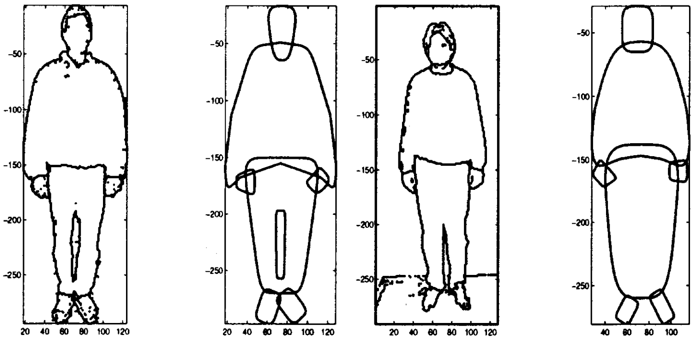


Figure 3.11 Approximations for two bodies

If $\mu_{11} \neq 0$ and $\mu_{20} <= \mu_{02} \rightarrow$

$$
\begin{aligned}
\theta &= atan(\frac{-2\mu_{11}}{\mu_{20} - \mu_{02} + ((\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2)^{1/2}}) \\
a_x &= (8(\mu_{20} + \mu_{02} + ((\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2)^{1/2}))^{1/2} \\
a_y &= (8(\mu_{20} - \mu_{02} + ((\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2)^{1/2}))^{1/2}
\end{aligned}
$$

If $\mu_{11} \neq 0$ and $\mu_{20} > \mu_{02} \rightarrow$

$$
\begin{aligned}
\theta &= atan(\frac{\mu_{20} - \mu_{02} + ((\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2)^{1/2}}{-2\mu_{11}}) \\
a_x &= (8(\mu_{20} + \mu_{02} + ((\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2)^{1/2}))^{1/2} \\
a_y &= (8(\mu_{20} - \mu_{02} + ((\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2)^{1/2}))^{1/2}
\end{aligned}
$$

Some examples for superellipse fitting are shown in Figure 3.11.

CHAPTER 4

INVARIANT SHAPE ATTRIBUTES

For object detection, it is necessary to select part attributes which are invariant to two dimensional transformations and are maximally discriminating between objects. Geometric descriptors for simple object segments, which correspond to the vectors in the graph nodes, such as area, circularity (compactness), weak perspective invariants [88], and spatial relationships are computed. These descriptors are classified into two groups: unary and binary features. For rigid and non-rigid objects the same binary features are used, the unary features are different.

In order to obtain high level semantics, a relational graph, where each node of this graph corresponds to a segmented part with its feature vector and each arc to their relationship, is built. Matching of the relational graphs of objects with the reference model yields to the detection of objects. The aspect graph of the reference object is formed according to the segmentation results of the training images.

Since the object is composed into its primitive subparts, simple attributes revisited in this chapter are sufficient to describe the segments characteristics. Furthermore, the following extensions are done for application specific algorithms: Since detection of skin regions in color images greatly increases the performance of human detection an elaborate skin color model based on a perceptually uniform color space is formed. For detection of rigid objects, e.g. car, a boundary shape code is developed enabling similarity ranking. Relative position and orientation obtained from the weak perspective invariants are used to detect human articulated movements.

4.1 Unary Features

The unary features for rigid objects are:

a) Hu moment invariants; b) compactness (circularity); c) eccentricity; d)boundary shape code (turnangle and length of concave axes).

Moment invariants are defined in [24]. The basic idea of moment invariants is to define a set of measures which are invariant to scale, rotation, and translation changes in a 2D plane. Given a 2D intensity distribution $f(x, y)$, the moments of this function are defined as:

$$m_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q f(x, y)\mathrm{dxdy} \qquad \text{for p, q = 0, 1, 2, ...}$$

These invariants can be modified to include translational invariance in the following way:

$$\mu_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \overline{x})^p (y - \overline{y})^q f(x, y)\mathrm{dxdy}$$

where $\overline{x} = \frac{m_{10}}{m_{00}}$, and $\overline{y} = \frac{m_{01}}{m_{00}}$. Scale invariant moments can be derived from the above to give a set of normalized central moments:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \qquad \text{where} \qquad \gamma = \frac{p+q}{2} + 1$$

A set of 7 functions can be defined which are invariant to translation, rotation, and scale changes in the image plane:

$$\begin{aligned} \phi(1) &= \eta_{20} + \eta_{02} \\ \phi(2) &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \end{aligned}$$

$$\begin{aligned}
\phi(3) &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\phi(4) &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\phi(5) &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{03} + \eta_{21})^2] \\
&\quad +(3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
\phi(6) &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{12} + \eta_{30})(\eta_{21} + \eta_{03}) \\
\phi(7) &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{03} + \eta_{21})^2] \\
&\quad +(3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned}$$

The eccentricity is calculated as the ratio of length of the minor axis to the length of the major axis, which is also the ratios of the eigenvalues of the principal components. The circularity (compactness) of the region provides a measure of how close the region is to a circle. The boundary shape code includes the turnangle and length of concave axes. This attribute can be used for ranking purposes. For example, the shape code of a sedan car body differs from the shape code of a sport car body (Section 4.3.2).

Eccentricity and circularity are defined as

$$\begin{aligned}
\text{Eccentricity} &\equiv \frac{\text{Major axis length}}{\text{Minor axis length}} \\
\text{Circularity} &\equiv \frac{\text{Perimeter}^2}{4\pi\text{Area}}
\end{aligned}$$

The unary features for non-rigid human bodies are:

a) compactness; b) eccentricity; c) color (hair and skin).

To represent the skin and hair color, perceptually uniform color system (UCS), proposed by Farnsworth [65] is used. Like other attributes, color attribute (c_j) of

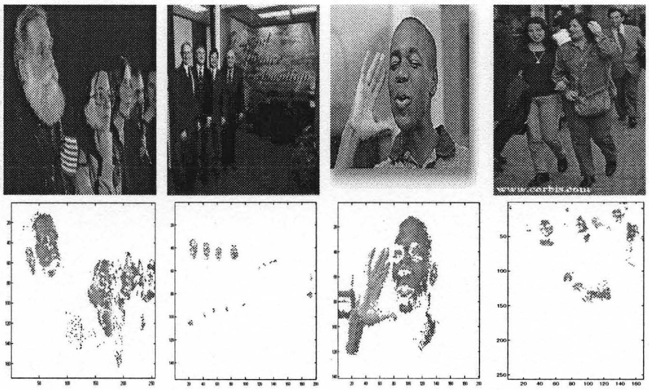


Figure 4.1 Skin color segmentation results for some test images.

an image segment will be separated by a distance from the model color (c_i) with tristimulus values (t_1, t_2, t_3). This color difference measure must reflect noticeable color differences in order to capture skin and hair color models and still be feasible to work in Euclidean space. Farnsworth nonlinear transformation produces uniform noticeable color differences that can be used in this approach. First RGB color information is converted to XYZ color system and the resulting chromaticity components are transformed using Farnsworth nonlinear transformation to the new chromaticity (u, v) values. The noticeable color differences in the XY chromaticity diagram can be fitted by ellipses, but these color differences become much more circular and tend to be uniform in the UV diagram [65]. These (u, v) values and the luminance are used to determine skin and hair locations in the image with adjacency and shape attributes (Figure 4.1). Our method relies mainly on the skin color model since the hair color model is not that reliable.

4.2 Binary Features

The same binary features are used for car and human detection. a) Ratio of areas; b) relative position and orientation; c) adjacency information between nodes with overlapping boundaries or areas. The relative position and orientation (Figure 4.2) are computed using the weak perspective approximation [88]:

$$\begin{aligned} u &= \frac{(\vec{p_3}-\vec{p_1}).(\vec{p_2}-\vec{p_1})}{|\vec{p_2}-\vec{p_1}|^2} \qquad v = \frac{(\vec{p_3}-\vec{p_1}).(\vec{p_2}-\vec{p_1})^{\perp}}{|\vec{p_2}-\vec{p_1}|^2} \\ \cos(\alpha) &= \frac{(\vec{p_2}-\vec{p_1}).(\vec{p_4}-\vec{p_3})}{|\vec{p_2}-\vec{p_1}||\vec{p_4}-\vec{p_3}|} \end{aligned}$$

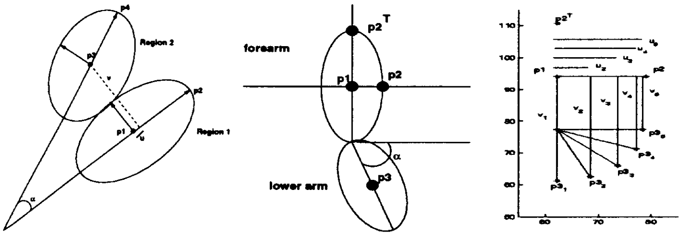


Figure 4.2 Left: Relative position(RP) and orientation(OR) of two regions. Middle: Arm model. Right: RP and OR changes of the forearm and lower arm with respect to each other.

CHAPTER 5

GRAPH MATCHING

Consider a human recognition application where head, arms, legs and torso are segmented and described by a set of unary and binary features. A system that contain unary and binary classification mappings must also be able to interpret the match and check the conditional rules in order to index the parts correctly. Our solution to this problem is to store the graph representation of the objects.

Although graph matching is widely used for representation of complex objects and scenes [98], [99], [100] and has a long history, it faces problems mostly due to the dependence on the segmentation results. For instance, a graph representation system called Acronym [101] that has been tested on aerial images to classify airplanes, failed when the extracted airplane features were not close enough to expected ones.

To overcome this problem, a new model based segmentation, that combines the initial segments or segments them to smaller parts using a feedback from graph representation of the object, is proposed. The reference graph representations of the objects are trained from the low level processing results. Extracted features for human detection differ also due to the different articulated movements and clothing. A graph matching algorithm with Bayesian framework is developed where conditional risk is minimized at every node of the branch to minimize the error rate.

Object detection is achieved by matching the relational graphs of objects (S regions) with the reference model. The input image graph O_n with N nodes ($N \geq S$) and a reference graph (O_r with N_r nodes) are matched. The aspect graph of the reference object is formed according to the segmentation results of the training images. For rigid and non-rigid objects, two slightly different algorithms are developed. The first proposed algorithm (Algorithm 1) is implemented for rigid objects (car). Then, this work is extended for non-rigid object (human) retrieval (Algorithm 2). In the graph matching algorithm for the rigid objects, during the decision step, hard decision is used. For non-rigid objects, in order to determine the body parts under the assumption that the unary and binary (relational) features belonging to the corresponding parts are Gaussian distributed, multi-dimensional Bayes classification is used. The graph matching algorithms for rigid and non-rigid objects are given below.

5.1 Graph Matching Algorithm 1

The reference graph for real images from a side view is given in Figure 5.1. A reference graph for sketches and three reference graphs for images; namely front-side view, back-side view, and side view images, are created. Our model graphs (for side-view sketches for three view angles of real cars) have the nodes for mainbody (upper and lower), windows (side, front-side, back-side, front, and back) and tires (front and back). To ease the understanding of matching algorithm we give the algorithm steps in a figure (Figure 5.2) for a sample test object and for a sample reference graph. It is assumed that input object has 5 nodes and the reference graph has 4 nodes.

Step 1: Begin from node of mainbody($i = 1$). Match this node to the O_n nodes by starting from a new branch for every possible match according to the unary descriptors (N branches at most). In this case the total matching cost between node pair $(1, j)$ for the branch b is calculated as

$$D^b(1,j) = d_{circ}(1,j) + d_{ecc}(1,j) + d_{mom}(1,j) + d_{curve}(1,j) \tag{5.1}$$

The differences for the unary descriptors (eccentricity, circularity, moment invariants, and turnangle and length of the concave axes of the boundary) are calculated according to the mean (μ) and deviation (δ) values as given in Eq. (5.2).

$$d_x(i,j) = \begin{cases} -w_x & \text{if } |x_j - \mu_{x_i}| \leq \delta_{x_i}; \\ w_x & \text{otherwise.} \end{cases} \tag{5.2}$$

where x_j is the corresponding attribute value of the image node, μ_{x_i} and δ_{x_i} are the mean and deviation values for this node attribute, respectively. These values are obtained from the training data set for sketches and for real images from side, front-side and back-side views. w_x is the weight for the penalty corresponding to this attribute. In our case it is 1.

In Figure 5.2, one can see that the matching cost between node 2 of the input graph and node 1' of the reference graph is above the threshold which is found during the training process, so there is no branch formed for this node pair.

Step 2: Increase i by 1. For every $j = 1, ..., N$ compute the matching cost $(D^b(i,j))$ between j^{th} and i^{th} node.

The total matching cost for a node pair (i,j) for a branch b $(D^b(i,j))$ is the sum of the unary and binary feature difference as

$$\begin{aligned} D^b(i,j) \quad = \quad & d_{circ}(i,j) + d_{ecc}(i,j) + d_{mom}(i,j) + d_{curve}(i,j) + \\ & d^b_{area}(i,j) + d^b_{adj}(i,j) + d^b_{position}(i,j) \end{aligned} \tag{5.3}$$

Binary feature differences are computed according to the previously matched nodes for every branch. For every matched node pair (n_i, n_j), the relative area, position and connectivity are computed between nodes j and n_j and between nodes i and n_i.

The difference of relational area, orientation, and position is calculated using already matched node pairs for the corresponding branch. Let image node n_i be matched to the model node n_j. Corresponding area distance between image node i and model node j for this branch is calculated as

$$d_x^b(i,j) = \begin{cases} -w_x & \text{if } \frac{\mu_{x_j} - \delta_{x_j}}{\mu_{xn_j} + \delta_{xn_j}} \leq \frac{x_i}{x_{n_i}} \leq \frac{\mu_{x_j} + \delta_{x_j}}{\mu_{xn_j} - \delta_{xn_j}}; \\ w_x & \text{otherwise.} \end{cases} \tag{5.4}$$

where x_i and x_{n_i} are the attribute values for the image nodes i and n_i, and μ and δ are the mean and deviation values, obtained from training, for the corresponding model nodes attribute, respectively.

Step 3: If the matching cost between nodes j and i for a branch is smaller than a threshold found in the training process, set (j, i) as the new matched node pair for this branch. Note that i must be different from the previous n_i's. It is observed that the relative distance of nodes vary highly for different type of cars and especially from sketch to sketch. However, the relative position at a coarser resolution does not change, e.g. the windows are in the upper mainbody (above the concave landmarks) and the tires are below the lower mainbody. The spatial relations (inside or adjacent nodes) are another relational distance. Also the relative position between the car parts and the mainbody is used as a checking step (Figure 5.3), e.g. the windows are in the upper mainbody and the tires are below the lower mainbody. The center of gravity of the window must be in the first part of the mainbody which is determined by the highest concavity points and the tires must be adjacent to the concavity points of the second part of the mainbody. Note that this information is not always available as the car can be a hatchback car, or the car can be sketched without these tire concavity points, or there can be occlusion due to other objects or to the view-point.

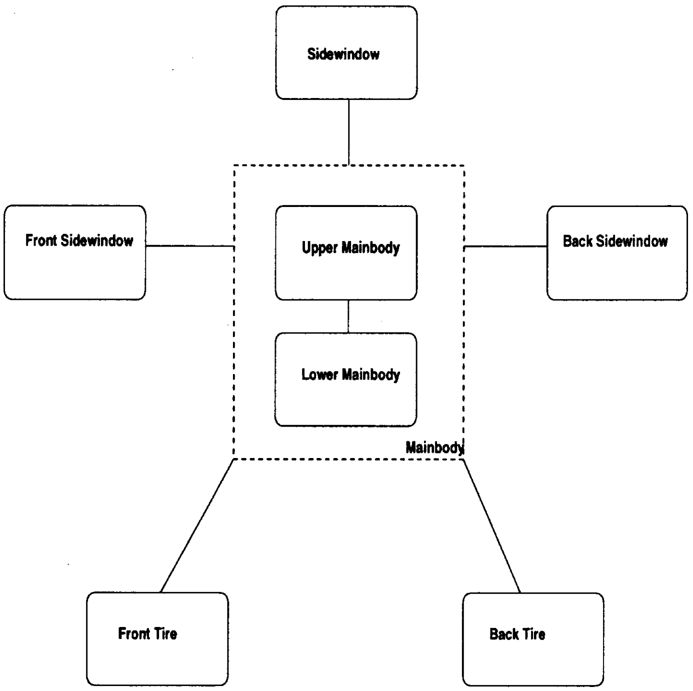


Figure 5.1 Reference graph from side view for real car images

In Figure 5.2, for each branch unary and binary feature differences are calculated. (For example, for branch 1, unary feature differences are calculated from node pairs 11' and 22', binary feature differences are calculated from the relation of 12 and 1'2').

Step 4: If all the $i = 1, ..., N_r$ nodes are taken into account, choose the branch with the maximum number of matched image nodes. If there are more than one resulting branches, choose one with the smallest total matching cost. In Figure 5.2, the winning branch is the fourth branch.

Step 5: If majority of reference graph nodes (75%) are matched, decide the presence of OOI, otherwise go to Step 1 for another view class and repeat Steps 1-5 until a match is found for a view class.

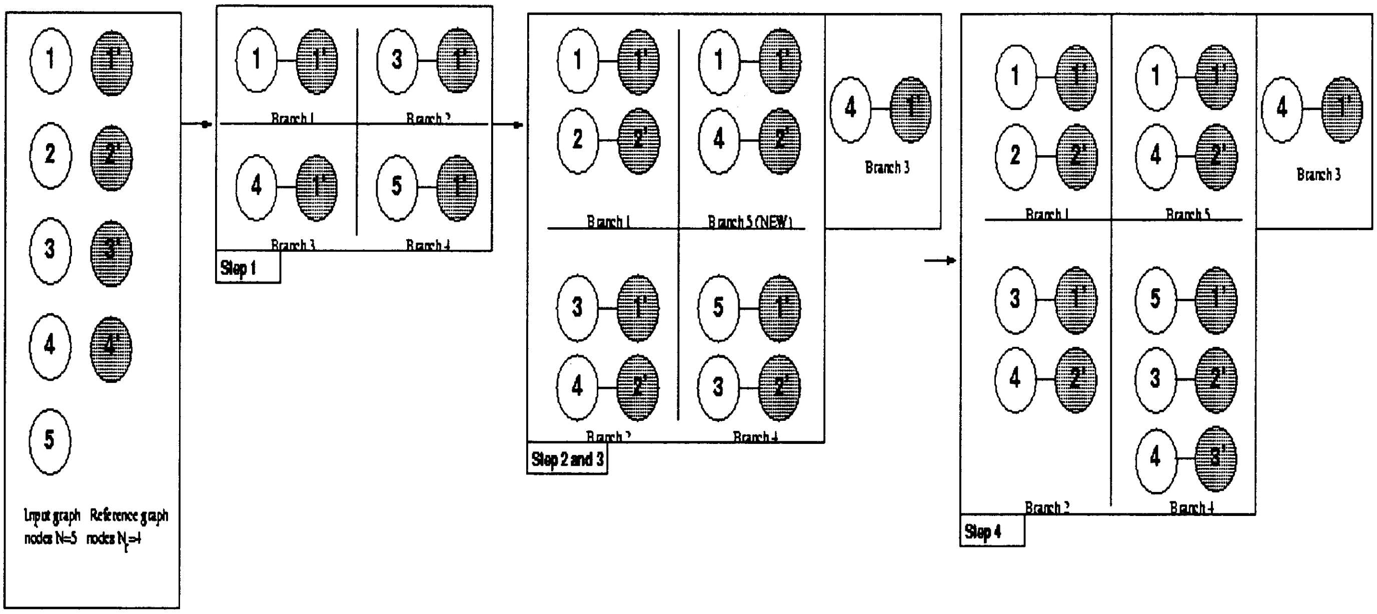


Figure 5.2 Graph matching for example input and reference graph nodes.

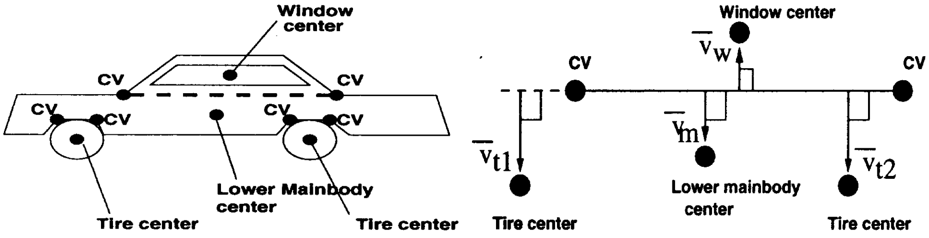


Figure 5.3 Left: Possible concavity and center points of a side car, Right: Normal vectors of the line between the maximum concavity points.

5.2 Graph Matching Algorithm 2

Two reference models namely front and side view models for human are used in the experiments. Our assumption is that human face (at least a part of it) must be seen since skin color is a dominant attribute for head (hair color model is also used but it is not that reliable)(Figure 5.4). Face detection allows to start initial branches efficiently and reduces the complexity. B_h represents the group of branches for the corresponding head area. Note that false face detection will result in a branch with single or very few matched nodes and will be eliminated. Relational graph matching would allow human detection without face part however it would increase the computational complexity significantly and it is left for future work. Each body part and meaningful combinations represent a class (ω). The combination of binary and unary features is represented by a feature vector (X). Note that feature vector elements change according to body part and the nodes of the branch under consideration. For example, for the first node of the branch, feature vector consists of unary attributes. The feature vector of the following nodes includes also binary features dependent on the previous matched nodes in the branch. For the purpose of determining the class of these feature vectors a piecewise quadratic Bayesian classifier is used. In our case, it is a multiclass and multifeature problem. For the reference model supervised learning is implemented using several test images. The features for each body part are assumed to be Gaussian distributed. From Bayes theorem:

$$k = arg \max_j P(\omega_j|X) = \max_j \frac{p(X|\omega_j)P(\omega_j)}{p(X)} \rightarrow X \in \omega_k \tag{5.5}$$

where $P(\omega_j)$ is a priori probability, $P(\omega_j|X)$ is a posteriori probability and ω represents a class. From [94], the discriminant function ca be written as

$$g_j(X) = log(p(X|\omega_j)) + log(P(\omega_j)) \tag{5.6}$$

For multifeature problems with arbitrary covariance the decision surfaces are hyperquadrics and the resulting discriminant functions are

$$g_j(X) = X^T W_j X + \omega_j^T X + \omega_{j0} \tag{5.7}$$

In Eq. (5.7)

$$\begin{aligned} W_j &= -1/2\Sigma_j^{-1} \\ \omega_j &= \Sigma_j^{-1}M_j \\ \omega_{j0} &= -1/2M_j^T\Sigma_j^{-1}M_j - 1/2log|\Sigma_j| + logP_{\omega_j} \end{aligned}$$

where M_j represents the class mean and Σ_j is the covariance matrix of each class. During supervised learning, for each reference model node that represents a class $p(X|\omega_j)$ is computed. $P(\omega_j)$ is computed with the assumption that each class is equal probable and parts such as arms represent two classes in the model file. Note that our problem differs from the classical Bayes classification method in the sense that one does not try to find the class of a given feature vector by minimizing the risk factor but tries to find the existence of a member for a given class. Our goal is to detect OOI in the image by matching the image segments to possible classes of OOI. Due to the generality of the problem "detecting human" and high variance of the within-class scatter matrices of unary feature vectors for different body parts, the relational features must be used. Relational attributes such as area ratio are elements of feature vector. Furthermore , conditional rule generation (r) eliminates the image segments that do not hold human body rules such as "face must be adjacent to torso", "if two arms are already matched in the branch there can not be another arm classification for that branch", and "angle between torso and face principal axis (α) can not exceed a certain threshold". Hence our problem is to find the existence of a member among image segments of a model class by maximizing the probability of feature vector for the given class in the corresponding branch.

The overall algorithm for the relational graph matching is given below.

```
for every model node $j \in O_r$ do
    for every branch $b$ do
        $(i_1, i_2)$ =match$(j, b)$
        copy branch $b$ and add node pair $(j,i_1)$ in the
        new branch and update $G^b$ by adding $g_j^b(X_{i_1})$
        copy branch $b$ and add node pair $(j,i_2)$ in the
        new branch and update $G^b$ by adding $g_j^b(X_{i_2})$
    end for
end for
choose $arg\,\max_{b \in B_h} G^b$
```

```
match$(j, b)$
for every image node $i \in O_n$ do
    for every matched node pair $(b_j, b_i)$ in the branch do
        if $\exists\, r(b_j, j)$ then
            if $r(b_i, i)$ holds then
                compute $g_j^b(X_{i,b_i})$
            else
                $g_j^b(X_{i,b_i}) = 0$
            end if
        else
                compute $g_j^b(X_i)$
        end if
        end for
    end for
    Return image nodes $i_1, i_2$ with two highest $g_j^b(x_i)$ values $>$ threshold
```

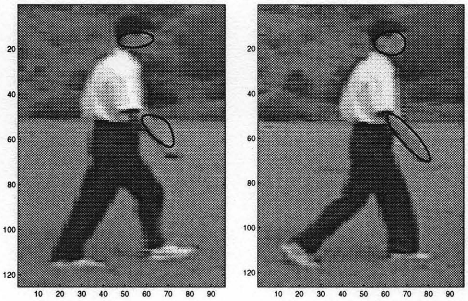


Figure 5.4 Modeling detected skin parts with superellipses.

CHAPTER 6

COMPRESSED DOMAIN TECHNIQUES

Our previous descriptors for rigid (car) and non-rigid objects (human) are defined in the uncompressed domain. The purpose of this chapter is to investigate object and activity recognition in the compressed domain in order to reduce computational complexity and processing time. For large libraries, compressed domain image/video processing for existing compression standards can solve the problem of bandwidth and intensive computing. In this thesis, new algorithms for object detection and activity recognition in JPEG images and MPEG videos are developed. It is shown that significant information can be obtained from the compressed domain in order to connect to high level semantics.

A hierarchical method for object detection and activity recognition at different resolution levels is proposed. The first and second parts are object and activity detection requiring minimal decoding of compressed data. The last part is graph-based object detection in uncompressed domain. Most object detection and human activity recognition techniques are done in the uncompressed domain and depend on proper segmentation of the body. The major contribution of the overall algorithm is to connect available data in compressed domain to high level semantics. The proposed hierarchical scheme enables working at different levels, from low complexity to low false rates. For instance, consider a recorded video sequence taken from a fixed camera surveying a passage. The first step would retrieve possible frames where people walk. If a walking person is detected to stop, second step would analyze the extracted region for posture recognition. If a suspicious movement is detected, the third step would be a more detailed investigation of the region in the uncompressed domain.

The first section of this chapter covers object detection in JPEG compressed still images. The algorithm uses DCT coefficients of the luminance and chrominance values obtained from JPEG algorithm. The second part corresponds to the principal component analysis of MPEG motion vectors from the P-frames to detect the human activities; namely, walking, running, and kicking. The motion vectors are grouped automatically according to velocity, distance and human body proportions.

6.1 Object Detection in JPEG Compressed Images

The JPEG still picture compression standard is simple to implement, is not computationally complex, and gets 10:1 to 15:1 compression ratios without significant visual artifacts. A single frame is subdivided into 8x8 sub-blocks, each of which is independently processed. Each block is transformed into DCT space, resulting in an $8x8$ block of DCT coefficients. These coefficients are then quantized and entropy coded into a compressed data stream. All the rigid and non-rigid object detection methods in still images, mentioned in the previous chapters, were applied on uncompressed images. However, many still images are usually stored in compressed form for efficient storage and transmission. First, one has to decompress these images in order to apply the proposed algorithms. In this section, the algorithms operate directly in the compressed domain on JPEG images, where the DCT coefficients are used as the input for our algorithm. Our proposed method operates on the I-frames of MPEG video or JPEG images, using DC-DCT coefficients of image blocks. DCT compressed images encode a two-dimensional image using the DCT coefficients (c_{uv}) of an NxN image region $(I_{xy}, 0 \leq x < N, 0 \leq y < N)$:

$$c_{uv} = \frac{1}{N} K_u K_v \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} I_{xy} cos \frac{\pi u(2x+1)}{2N} cos \frac{\pi v(2y+1)}{2N} \tag{6.1}$$

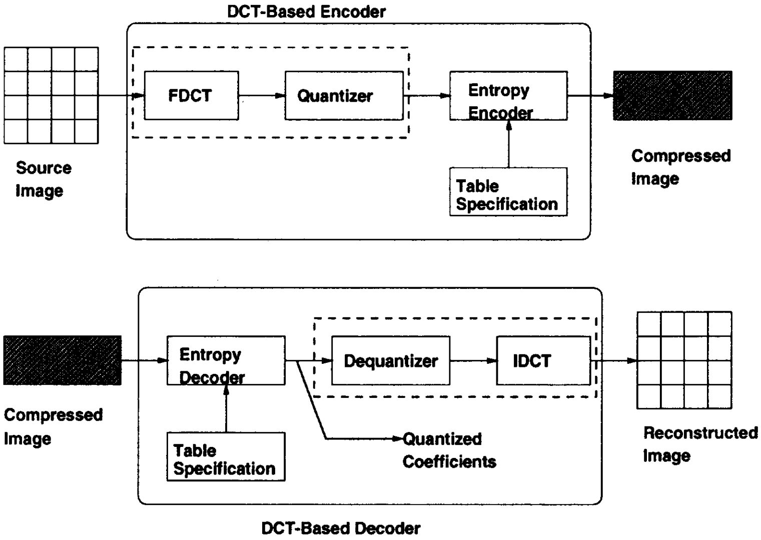


Figure 6.1 Overall algorithm.

In Eq. 6.1, u and v denote the horizontal and vertical frequencies and $K_u = 1/\sqrt{2}$ if $u = 0$ and $K_u = 1$, otherwise. The AC components ($c_{uv}, u \neq 0$ or $v \neq 0$) capture the spatial frequency and directionality properties of the image block. The DC component, which is the averaged and sub-sampled version of the input image, is used in our algorithm. The quantized DCT coefficients can be readily extracted from a video stream and JPEG data. Although they are quantized, the rank information is preserved and they can be used without any decoding procedure.

Figure 6.1 shows the key processing steps of the DCT based JPEG encoder and decoder. At the input to the encoder, source image samples are grouped into 8x8 blocks, shifted from unsigned integers with range $[0, 2^P - 1]$ to signed integers with range $[-2^{P-1}, 2^{P-1} - 1]$. At the output from the decoder, the IDCT outputs 8x8 sample blocks to form the reconstructed image. The discrete cosine transform in the JPEG algorithm, is based on the fast 1-D DCT algorithm proposed in [96]. A two dimensional DCT can be obtained by applying first a 1-D DCT over the rows followed by a 1-D DCT to the columns of the input data matrix. The same idea is also implemented for the inverse transform. Because the dequantization and 1-D

IDCT are implemented simultaneously, the output of the entropy decoder (quantized DCT values) is used directly. In Figure 6.1, the extraction step of the quantized DCT coefficients in the decoder part are given.

From the regenerated array of quantized coefficients, that are found during the JPEG decompression, the DC-DCT coefficients are extracted. The processing speed of the proposed method is fast since it does not require a fully decompressed MPEG video or JPEG image. The DCT information which is used is readily available from the compressed image. The processing unit for the algorithm is a DCT block. The proposed graph matching algorithm operates on an image which is 1/8th of the original image size in each dimension when 8x8 DCT blocks are used.

In Figure 6.2, the original image and the segmentation results are displayed in the first row. In the second and third rows, the quantized DC-DCT images and segmentation results for different resolutions are given. The DCT image dimension for the second row is 128 by 48 where for the third row it is 64 by 24. Since the DC-DCT coefficients give the average intensity values of the blocks, one can get rid of the local luminance changes due to the reflection and other factors. Besides the processing speed, this method also smoothes the image to test the system performance for different resolution levels. Note that the JPEG images and MPEG sequences include human where in most cases the skin regions consist of one or two 8 by 8 DC-DCT blocks or 16 by 16 macroblocks. Usually, the skin information from the DCT values of color components can not be used for human detection since the resolution requirement is not met. If the skin regions are detected (Figure 6.3), the next step will be the segmentation and implementation of the proposed model based graph matching algorithm on the DC luminance blocks for each frame. As it is mentioned before, finding the head region from the skin color information is a crucial step for the performance of the matching algorithm.

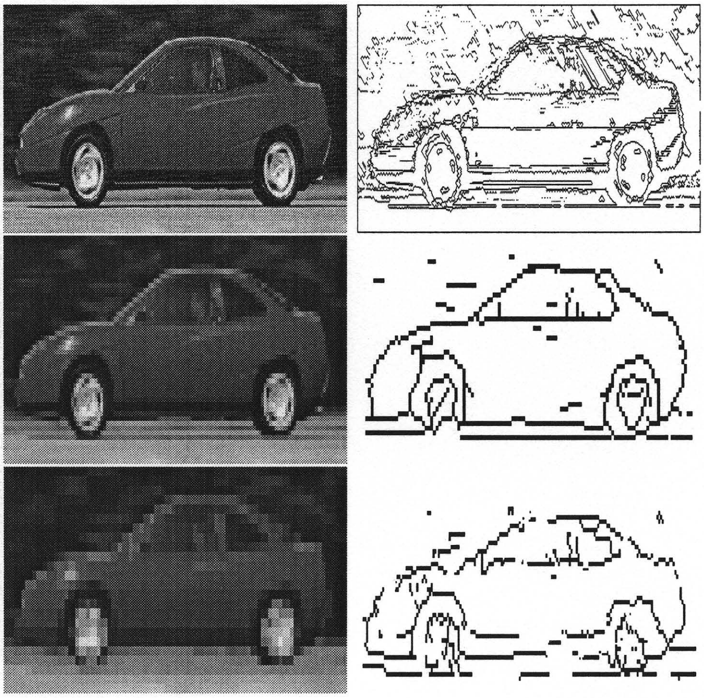


Figure 6.2 Top left: Original image, right: Segmentation result; Middle left: Quantized DC-DCT coefficients (128x48), right: Segmentation result; Bottom left: Quantized DC-DCT coefficients (64x24), right: Segmentation result.

For low resolution JPEG images where human is present, a new algorithm for human detection is proposed. The system detects people in arbitrary positions in the image and in different scales. This approach is given in the next section.

6.2 Human Detection in Lower Resolution and Monochrome JPEG Images

The major problem in low resolution JPEG images is the human skin color detection which leads us to use the DCT coefficients in a different way for human detection. In this new algorithm, the overall shape of a standing or walking person (from front or back-view) in still images is detected by using the AC-DCT coefficients. This work presents a new approach similar to an earlier work by Papageorgiou [59] where pedestrians are detected by using the wavelet templates and classification is achieved

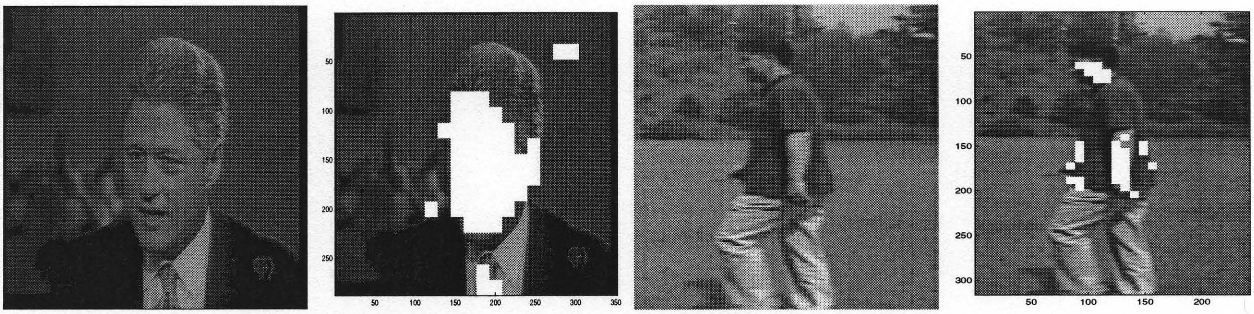


Figure 6.3 First and third images: Original frames (YCbCr: 4:2:0 and 4:4:4), Second and fourth images: Marked frames with macroblocks detected as skin regions.

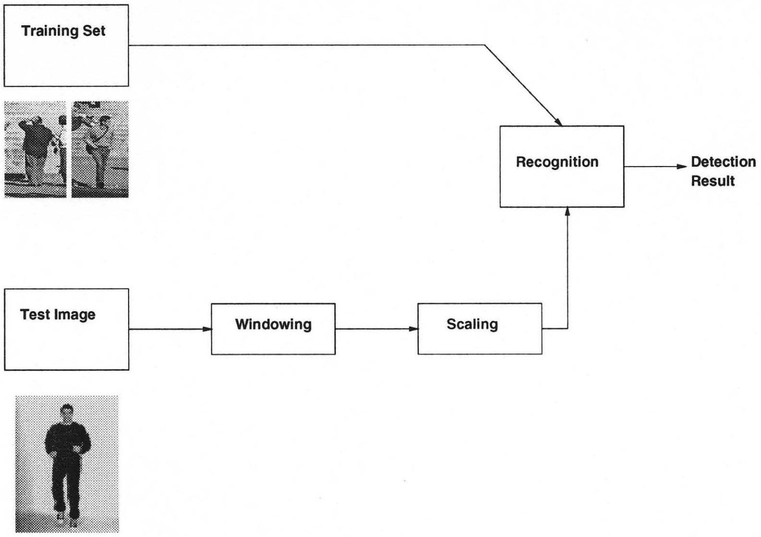


Figure 6.4 Human detection system in low resolution JPEG images.

by using support vector machine technique. The recognition part of our approach is similar to the face recognition algorithm developed by Turk, and Pentland [95]. Our proposed algorithm is given in Figure 6.4.

To capture the intensity variations, first order AC coefficients are used (Figure 6.5). To train our system, approximately 800 pedestrian images, that are centered in a 128x64 pixel window and are obtained from artificial intelligent laboratory at MIT, are used. Figure 6.6 shows some training images and corresponding AC coefficients. The windowing step in Figure 6.4 determines a 128x64 window and shifts it throughout the test image. The regions, which have a lower AC energy than a given threshold, are eliminated. The following step, scaling, resizes the image part in the 128x64 window to achieve multiscale detection. In our experiments, the window is scaled from 0.8 to 1.2 times its original size by 0.1 increments in the compressed domain [58]. Recognition step in the algorithm is given in the following subsection.

DCT coefficient values capture the local directionality and coarseness of the spatial image. The vertical (horizontal) edges in uncompressed image correspond to high frequency component in the horizontal (vertical) frequencies and diagonal variations correspond to channel energies around the diagonal harmonics. Our approach is based on the observation that the structural information of human silhouettes can be captured from AC-DCT coefficients. In particular, the energy of blocks, that is obtained by summing up the absolute amplitudes of the first order harmonics, is used. The sides of the human body have a high response to the vertical harmonics while AC coefficients of the horizontal harmonics capture head, shoulder and belt lines. Furthermore, the corner edges at shoulders, hands and feet contribute to local diagonal harmonics. In [59], the structural information of pedestrians is presented by a subset of wavelet coefficients and pedestrians are detected by the support vector machine classification method. Our work aims to retrieve information from images and videos compressed using standard algorithms such as JPEG and MPEG. This differentiates our approach from that of [59] where the compression algorithms are governed by characteristics of object of interest to be retrieved. Our results are compared with those of [59] for frontal and near-frontal poses since our system is trained only for these view angles. The authors in [59] use an overcomplete Haar dictionary of 16 x 16 pixels and train the system by using 564 positive examples that contain nonoccluded pedestrians and 597 negative examples that do not contain pedestrians. The detection rate for 141 nonoccluded pedestrian images in frontal or near-frontal images is 82% and false positive rate is 1 per 15000 windows.

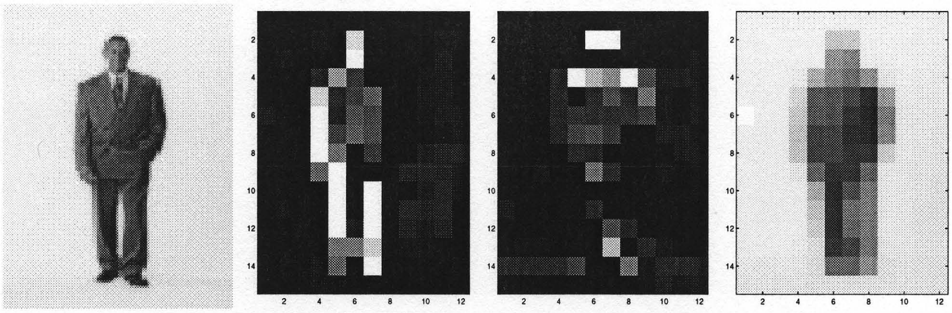


Figure 6.5 First image: Original image; Second: Horizontal AC coefficients; Third: Vertical AC coefficients; Fourth: DC coefficients

In our experiments, 800 positive examples and different negative examples with a bootstrapping algorithm are used in order to train the system. Our detection rate is approximately 80%. Although the resolution in [59] algorithm is higher and can be adaptive and therefore can provide better spatial resolution, our approach has the advantage of using the available data in standard compression algorithms and gives comparable detection results.

6.2.1 Recognition

Our goal is to find the principal components of the distribution of human bodies, or the eigenvectors of the covariance matrix of the human body images. These eigenvectors represent a set of features which together characterize the variation between human images. The number of eigenvectors is equal to the number of images in the training set. The bodies can also be approximated using only the best eigenvectors with the highest eigenvalues because of the computational efficiency.

Similarity measure in eigenspace representation for pattern matching in images is preserved under linear, orthogonal transformations. This implies that the principal component method gives exactly the same measure of match on transformed data as on pixel domain data. For lossy compression schemes such as JPEG and MPEG, the quantization of the transformed data is the cause for the degradation of the similarity

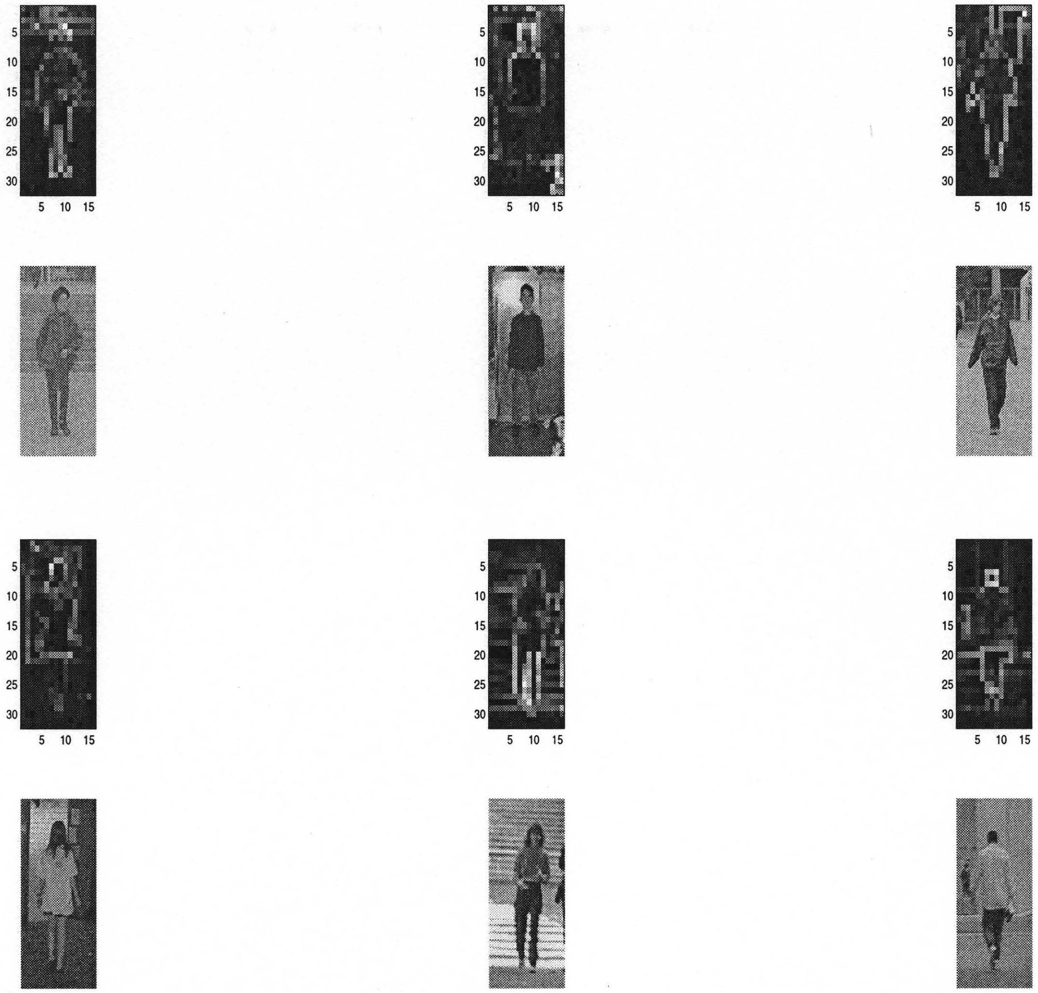


Figure 6.6 Training images and AC coefficients.

measure.Although the DCT coefficients are quantized, the essential information for matching purposes is preserved.

The following steps summarize the recognition:

- From the training set of human body images, calculate the eigenvectors and eigenvalues.

- When a new human image is tested, calculate a set of weights based on the input image and the M eigenvectors by projecting the input image onto each of the eigenvectors.

- Determine if the image is a human body by checking if the image is sufficiently close to the human body space.

The training set of human images is $\Gamma_1, \Gamma_2, ..., \Gamma_M$, and the average is $\Phi = (\Gamma_1 + \Gamma_2 + ... + \Gamma_M)/M$. The difference of a human image from this average image is $\phi_i = \Gamma_i - \Phi$. Our goal is to find a set of M orthonormal vectors, $\mathbf{u_k}$ and their eigenvalues $\beta_\mathbf{k}$ which best describes the distribution of the data by using the principal component analysis. $\mathbf{u_k}$ and $\beta_\mathbf{k}$ are the eigenvectors and eigenvalues, respectively, of the covariance matrix C:

$$C = \frac{1}{M} \sum_{n=1}^{M} \phi_n \phi_n^T = AA^T \tag{6.2}$$

where the matrix $A = \frac{[\phi_1 \phi_2 ... \phi_M]}{\sqrt{M}}$. The matrix C is a N by N matrix and the calculation of eigenvectors and eigenvalues of this matrix is a difficult task. To reduce the computational complexity, the eigenvectors x_k and eigenvalues λ_k of the matrix $A^T A$ are computed. It can be proven that the eigenvectors $\mathbf{u_k}$ of matrix C can be computed as:

$$\mathbf{u_k} = \frac{\sum_{l=1}^{M} \phi_l x_{kl}}{\sqrt{\lambda_k}} \tag{6.3}$$

and the eigenvalues are the same those matching x_k. The first 12 eigenimages obtained from 800 training images are shown in Figure 6.7. Creating the vector of weights for an image is equivalent to projecting the image onto the human body space. The distance ϵ between the image and its projection onto the body space is the distance between the mean adjusted input image $\phi = \Gamma - \Phi$ and $\phi_f = \sum_{k=1}^{M'} \omega_k \mathbf{u_k}$, its projection onto human body space, where $\omega_k = \mathbf{u_k^T}(\Gamma - \Phi)$ for $k = 1, ..., M'$. In our algorithm, M' eigenvectors corresponding to the largest eigenvalues from M human images are used. In Figure 6.8, two examples for the algorithm performance are given. The bounded areas have the minimum distance from the training data set. The system is also trained for background classification by using several images where human is not present. The overall system performance is tested on 40 images and some of the human classification results are given in Figure 6.9. The test images contain a total of 126 non-occluded frontal poses and the algorithm can detect 101 of them correctly.

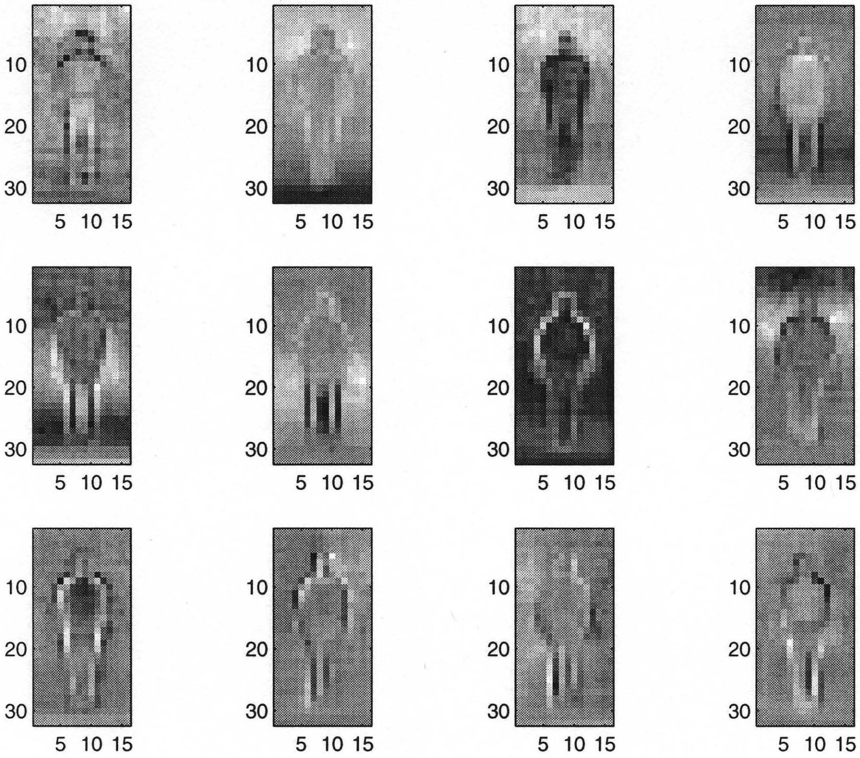


Figure 6.7 12 eigenimages.

6.3 Activity Recognition Using MPEG Motion Vectors

Great effort has been devoted to human recognition topics, such as face recognition in still images and motion analysis of human body parts. Most of the previous work is done in uncompressed domain. Since image and video applications are generally represented in the compressed domain, such as JPEG or MPEG, there is a need for image/video manipulation and automatic content extraction in the

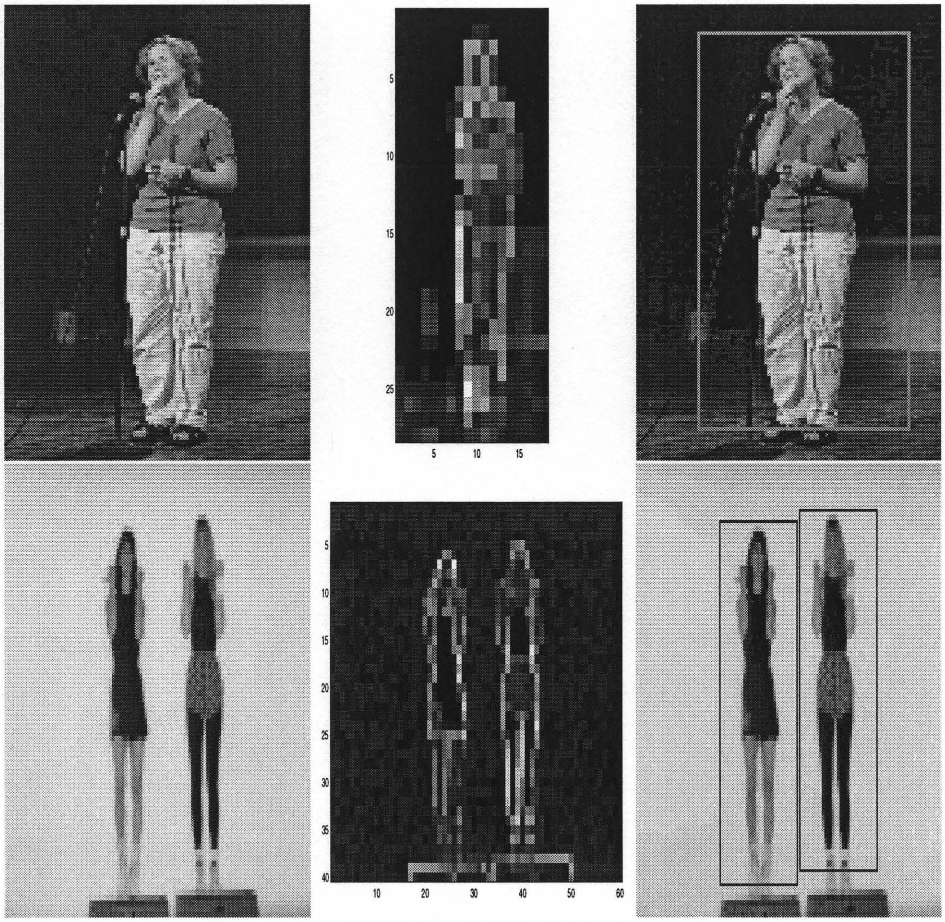


Figure 6.8 First row: Left: Original image, Middle: AC-DCT values, Right: Classification result; Second row: Classification for multiple human.

compressed domain. As stated in Chang [97], for existing compression standards the compressed-domain image/video manipulation techniques can be used to help to solve the bandwidth problem. Hence applications without expanding the coded visual content back to the large, uncompressed domain would reduce the need of large bandwidth and intensive computing. The use of available information in compressed video and images mostly has been investigated for video indexing, and shot and scene classification. In Yeung [73], hierarchical decomposition of a complex video is obtained using scaled DC coefficients in an intra coded DCT compressed video for browsing purposes. The technique combines visual and temporal information to

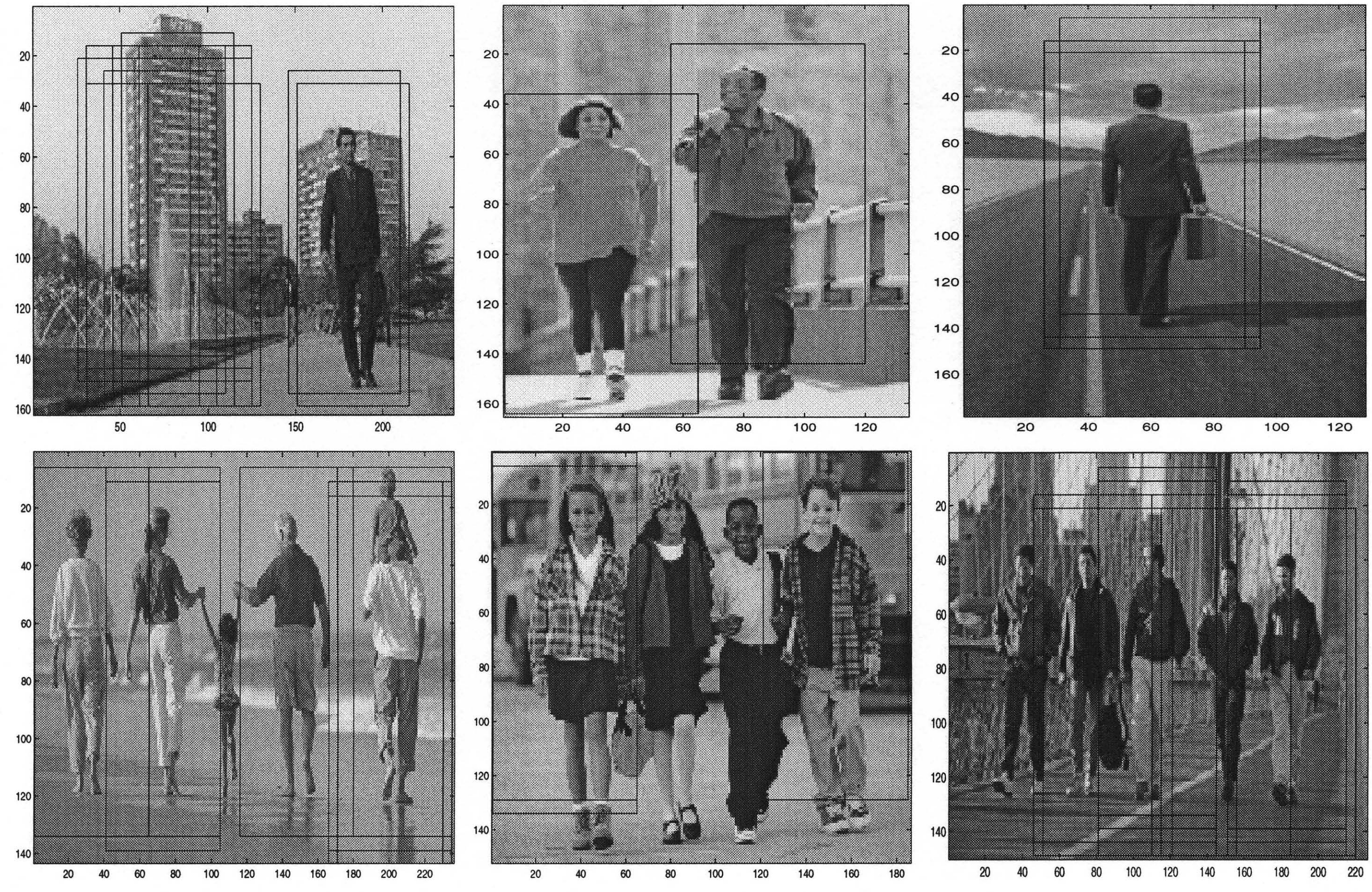


Figure 6.9 Human classification results.

capture the important relations within a scene and between scenes in a video. A general model of a hierarchical scene transition graph is applied for video browsing. In Yeo [102], the authors examine the direct reconstruction of DC coefficients from motion compensated P-frames and B-frames of MPEG compressed video. Their analysis and experimental results show that lower cost approximations can be used successfully for various image processing operations, such as shot detection, shot matching and clustering. In Dawood [74], an automatic scene classification scheme is proposed for MPEG videos. The scenes are divided into low, medium, and high texture and activity scenes. The bit rates of the I, P and B frames are used in shot texture classification while the percentage of macroblock types are used for shot motion classification.

MPEG motion vectors are used mostly to index videos (low-high activity) and track objects. The object detection in the compressed domain is more restricted since this application requires more detailed information. In Schonfeld [61], an object tracking algorithm is proposed using compressed video only with periodically decoding I-frames. The object to be tracked is initially detected by an accurate but computationally expensive object detector applied to decoded I-frames. In Wang [63], an algorithm to detect human face regions from dequantized DCT coefficients of MPEG video is proposed. The algorithm uses the DC DCT values of chrominance, shape, and energy distributions of the face area. This method is suitable for color images with face regions greater than 48 by 48 pixels (3 by 3 MPEG macroblocks). The authors extend their work in [103] in order to track and summarize faces from compressed video. The previous algorithm is used to detect faces and MPEG motion information is used with the Kalman filter prediction to track faces within each shot. The representative frames are then decoded for pixel domain analysis and browsing.

Activity recognition problem can be divided into two subparts: the first one is collecting satisfactory measurements and the second one is developing a recognition algorithm based on these measurements. Most of the related work use activity measurements from uncompressed images after a proper segmentation of human body parts. Our measurements are obtained from MPEG motion vectors for macroblocks in P-frames. Since the resolution of the motion vectors is one macroblock and there is no direct correspondence with the object parts and their motion, a robust and global model must be used. The corruption of data is another problem in MPEG motion vectors since some blocks can not be tracked during some frames. An overview of research on human motion analysis can be found in Aggarwal [56] and Gavrila [60]. The major problems in the activity recognition is the scale, shift and projection changes between the model and the test data and segmentation dependency. One of the activity modeling methods proposed in Walter [77] is based on first order Markov model descriptions and continuous propagation of observation density distributions. Hidden Markov Models are used to predict the state transitions. In Rangarajan [78], speed and direction components of 2D trajectories are represented by scale-space images that are invariant Euclidean transforms. A method based on time-frequency analysis is proposed in Cutler [79] to detect periodic human motion with self-similar characteristic. The outline of the human body is used to detect the periodical relative limb movement in Curio [80] by a template matching process. In these approaches, for each activity, a separate model is developed in order to compare with the observed activity. These approaches are robust to local transformations but lack a global detailed model to capture the variabilities.

Principal component analysis method is one of the global approaches. Our activity recognition model is based on the principal component analysis which has also been used by Yacoob and Black [76] for human activity recognition in uncompressed video sequences. The authors use the motion measurements for segmented human body parts. In our method, first the moving regions are detected and then the motion vectors are grouped automatically by using the ratio of the human body parts. Hence the measurements do not correspond to the actual human body parts but to macroblock groups corresponding to human region. For the classification of moving regions, the neighboring blocks with a velocity greater than a predefined threshold are classified as one moving object. The following subsection covers the principal component analysis (PCA).

6.3.1 Principal Component Analysis

PCA was successfully used for face recognition. A compact representation of facial appearance is described in [104], where face images are decomposed into weighted sums of basis images using a Karhunen-Loeve expansion. The eigenpicture representation has been used in [105] as eigenfaces for face recognition. PCA is a dimensionality reducing technique, used in pattern recognition. It reduces dimensionality by projecting the motion vectors to a new space spanned by the training data set. For training the system, several walking, running and kicking man sequences which are temporally aligned are used. For these sequences, the object region is extracted by grouping MPEG motion vectors. Then, the object is segmented to three parts (upper body, torso and lower body) according to the human body proportions. The mean of the motion vectors in horizontal and vertical direction is computed for the macroblocks corresponding to each part (6 parameters) for a number of sequences T. A training set of k different examples for each activity forms matrix A of dimensions $6T \times k$. Then the singular value decomposition of the matrix A is computed to get the approximated projection of the exemplar vectors (columns of A) onto the subspace spanned by the $q < k$ basis vectors. Hence activity basis with parameters m are computed.

$$A = U\Sigma V^T \tag{6.4}$$

where A is the motion parameter matrix, U represents the principal component directions, Σ includes the singular values, and V^T expands A in principal component directions. To recognize the activities, an unknown sequence, other than test sequences of an activity which can be shifted and scaled in time is compared with the training set. The transformation function κ might model uniform temporal scaling and time shifting to align observations with exemplars. Let $D(t)$ be an observed activity, $[D]$ be the nT column vector, obtained first by concatenating the n feature values measured at t, and then concatenating $D(t)$ for all t. Let $[D]_j$ denote the j-th element of vector $[D]$. By projecting this vector on the activity basis, a coefficient vector, $\overline{c}$, is recovered, which approximates the activity as a linear combination of activity basis. For recovering the coefficients, the error has to be minimized:

$$E(\overline{c}) = \sum_{j=1}^{nT} \rho(([D]_j - \sum_{l=1}^{q} c_l U_{l,j}), \sigma) \tag{6.5}$$

where $\rho(x, \sigma)$ is an error norm over x, and σ is a scale parameter. Let $\kappa(\overline{a}, t)$ denote a transformation with a parameter vector $\overline{a}$ that can be applied to an observation $D(t)$ as $D(t + \kappa(\overline{a}, t))$. After Taylor series expansion of $D(t + \kappa(\overline{a}, t))$, the error function becomes:

$$E(\overline{c}, \overline{a}) = \sum_{j=1}^{nT} \rho([D_t(t)\kappa(\overline{a}, t) + D(t))]_j - \sum_{l=1}^{q} c_l U_{l,j}, \sigma) \tag{6.6}$$

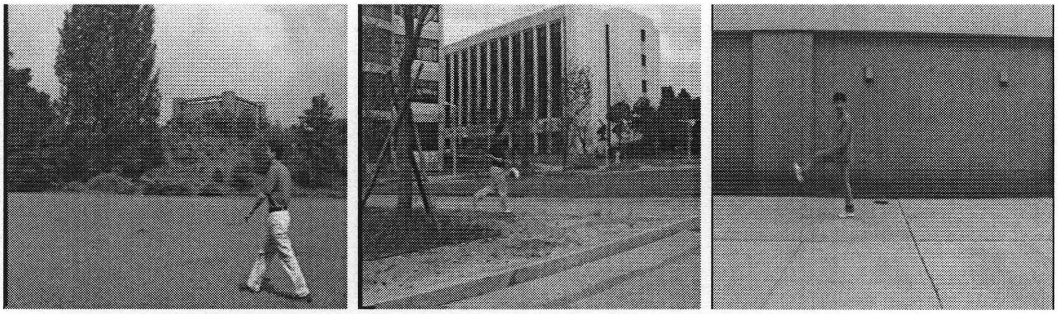


Figure 6.10 Frames from walking, running, and kicking man training sets.

Equation 6.6 can be minimized with respect to $\overline{a}$ and $\overline{c}$ using a gradient descent scheme with a continuation method that gradually lowers σ. The normalized distance between the coefficients m_i from the training data set and coefficients of exemplar activities c_i is used to recognize the observed activity that is transformed by the temporal translation, scaling and speedup parameters [76]. The Euclidean distance is given as

$$d^2 = \sum_1^q (c_i/||\mathbf{c}|| - m_i/||\mathbf{m}||)^2 \tag{6.7}$$

where $\mathbf{c}$ is vector of expansion coefficients of an exemplar activity. The algorithm is applied for recognition of three activity classes: walking, running, and kicking. 10 training test sequences for each class are obtained from various sources for the side-view. The camera motion is assumed to be zero. In Figure 6.10, some test frames from the activity training sets are displayed. The detection of the moving regions and the determination of the activities from the grouped MPEG motion vectors give a coarse information about the scene. Figure 6.11 displays the motion vectors obtained from the P-frames. Afterwards, these vectors are grouped by using the ratio of the human body parts.

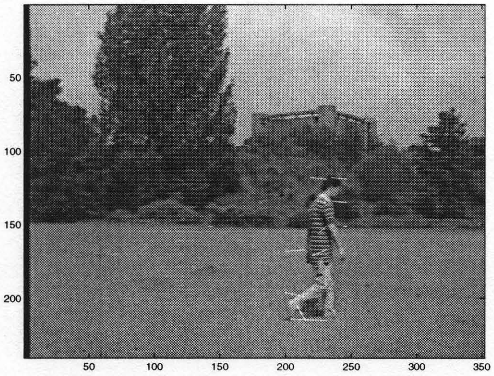


Figure 6.11 Motion vectors between P-frames for a walking man sequence.

For a more detailed investigation, one may need additional information. DC-DCT coefficients and coefficient differences obtained from MPEG sequences in the compressed domain are presented in the next subsection.

6.3.2 DC Differences

In this subsection, 8 by 8 block information (DC values) in the frames where human activity has been detected from the macroblock information (motion vectors), are used. The difference of the DC values for 8 by 8 blocks between consecutive frames are computed and the difference image is binarized by thresholding. To train our system, several human activity sequences from side-view with the similar camera distance, human motion direction and velocity are used. In order to find the template for each body position during one activity period, the mean of the moving regions, corresponding to these positions, are calculated. One of the templates is shown in Figure 6.12. The classification is done by using a basic template matching measure. Note that the mirror image of the template is also used. For every DC-DCT difference frame, the blocks are compared to the activity templates. For scale change invariance, the moving block regions with different scale parameters are scaled and the matching value for each scale factor is calculated.

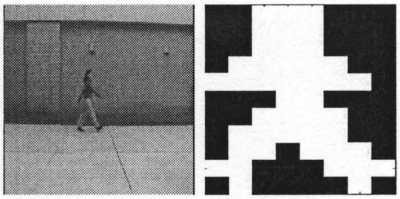


Figure 6.12 Left: Walking position in the uncompressed image, Right: Template corresponding to this position.

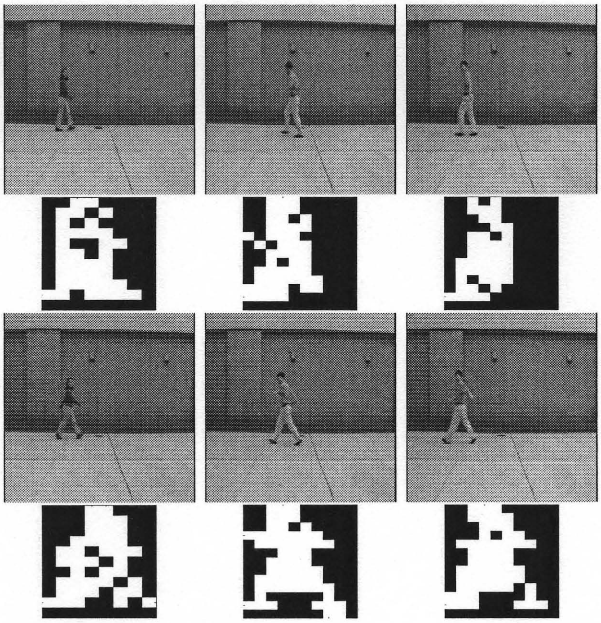


Figure 6.13 First and third rows: Left column: Walking position from the training set, Middle and right columns: Resulting frames with the minimum matching costs. Second and fourth rows: DCT coefficient difference for the corresponding frames.

6.3.3 Convolution

In the previous subsection, the difference of the DC values for 8 by 8 blocks between consecutive frames are used for activity period detection. To find the periods of an activity in lower resolutions without computing DC values, one can use the angle between the horizontal and vertical motion vectors obtained from 16x16 motion blocks. In Figure 6.14, two walking periods of the same walking man are given. If one convolves the activity parameters of one period with the whole sequence, the peaks in the convolution sequence (> 10) that correspond to other periods can be

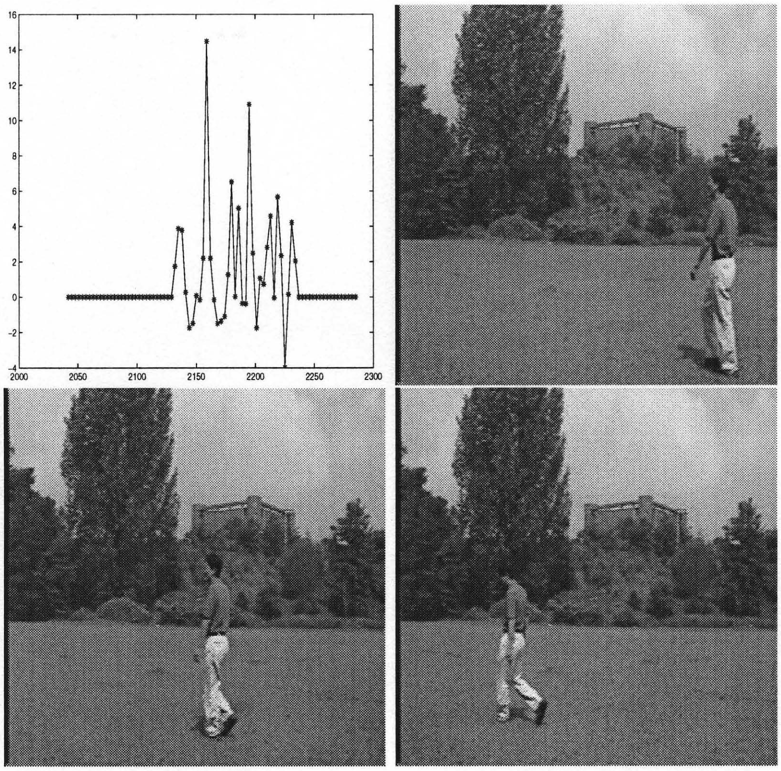


Figure 6.14 Top left: Convolution sequence for the same man (middle body), Top right: Beginning of the model period, Bottom left: End of the model period. Bottom right: The frame which is found from the peaks of the convolution sequence (end frame of the second period) .

obtained. In Figure 6.15, the model period of a walking man is convolved with the whole sequence where several persons are present. Note that image(b) and image(c) have the same characteristics as image(a) and found from the convolution sequence correctly.

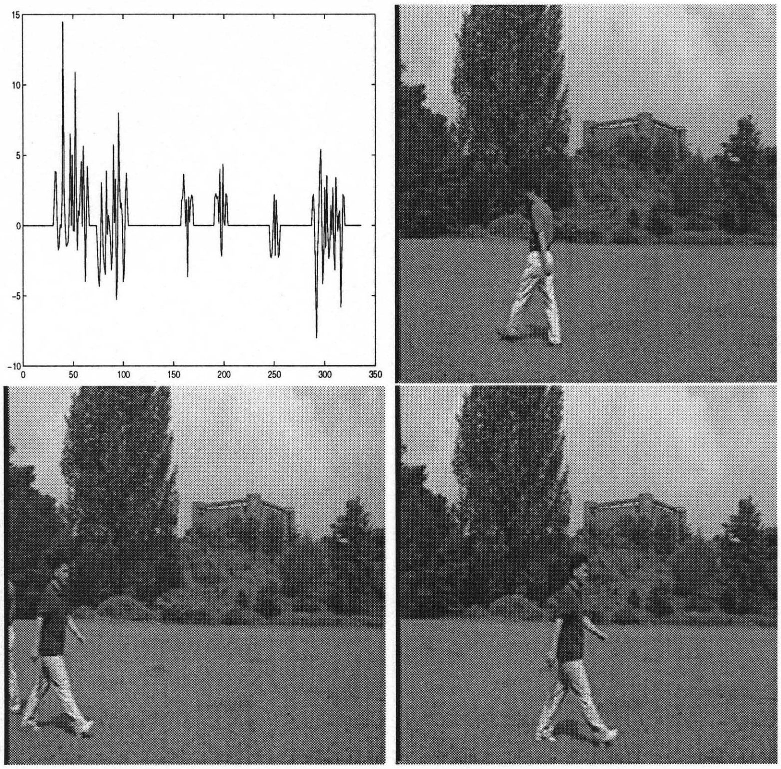


Figure 6.15 Top left: Convolution sequence (middle-body) for the whole sequence Top right: Image(a) is from the model. Bottom: Image(b) and (c) the frames which are found by using the peaks of the convolution sequence.

CHAPTER 7

EXPERIMENTAL RESULTS FOR RIGID OBJECTS

This chapter corresponds to the detection of car images in sketches, real images and video frames. The results for uncompressed and compressed domain techniques are given in the following sections, respectively. The algorithms are implemented on still images and video sequences.

7.1 Uncompressed Domain

The model graphs for sketches and for real images are obtained by computing the statistics of the node attributes for manually segmented parts. Some training images are displayed in Figure 7.1. Training data sets for sketches and for real images consist of 40 sketches from the side-view and 70 real images from side, front-side and back-side views. For each car part, the average mean values μ and maximum deviation δ of the unary and binary attributes are obtained from the training data sets. The use of the mean μ and maximum deviation δ on the algorithm are given in the graph matching chapter. In Figure 7.2, some of the real car images from the training data set and the resulting classifications are displayed.

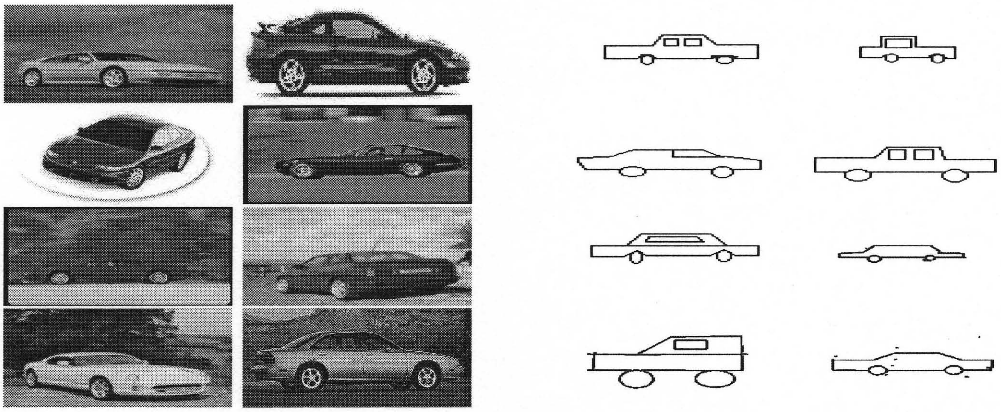


Figure 7.1 Left two columns: Training set images, Right two columns: Training set images obtained from student sketches.

In Figure 7.3, an example sketch image is shown. Sixth branch with the maximum number of matched nodes and minimum total matching cost is the result of matching. For this example the number of segments is 5, the number of nodes (number of combinations) is 8, and the number of candidate branches with node pairs having matching cost smaller than the threshold value (5) is 6. Figure 7.4 shows our approach on a test sketch. All the initial segments are represented with different colors. After graph matching three mainbody parts, namely upper, lower-front, lower-back, are combined together and represented with the same color. The windows and tires are also classified correctly. Some sketch images and corresponding segment, node and branch numbers are displayed in Figure 7.6.

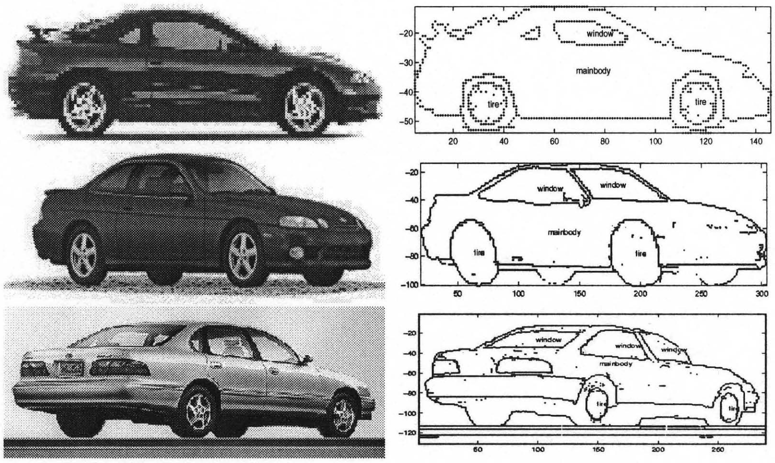
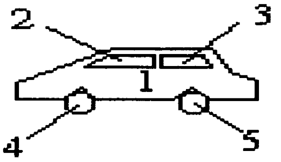


Figure 7.2 Training examples: Left: Original image, Right: Classification result

After the determination of the model graph attributes, the algorithm is tested on real images and sketches for several total matching cost values. In Figures 7.7, 7.8, 7.9 and 7.10 the percentage of correct, false detection and miss versus matching cost threshold are depicted for 18 sketches, 22 side, 18 front-side and 15 back-side images, respectively. In these four figures, the search of an optimum matching cost threshold is displayed. The presented percentage of classification corresponds to the classification results of the image segments.



Branch number:	1	2	3
	Mainbody 2, dif=3	Mainbody 3, dif=4	Mainbody 4 1, dif=2
	Total dif = 3	Total dif = 4	Total dif = 2
	4	5	6
	Mainbody 5 1, dif=2	Mainbody 5 1 4, dif=3	Mainbody 1, dif=2
	Tire 4, dif=-2	Window 5, dif=-1	Tire 5, dif=-3
	Window 2, dif=-2		Tire 4, dif=-3
	Window 3, dif=-3		Window 2, dif=-2
			Window 3, dif=-3
	Total dif = -4	Total dif = -4	Total dif = -9

Figure 7.3 Top: Original sketch with segment numbers, Bottom: Resulting branches, where the sixth branch has the maximum number of matched nodes and minimum total difference. The segment numbers 6, 7 and 8 are the combinations of segments 4 and 1; 5 and 1; 5, 1 and 4 respectively.

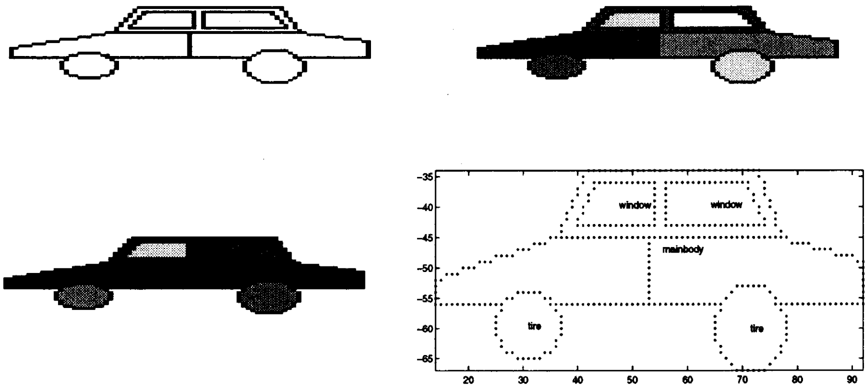


Figure 7.4 An example for classification of parts of a sketched car. Top Left: Original image; Top Right: Segments from closed regions; Bottom Left: Classification of nodes after graph matching; Bottom Right: Resulting nodes.

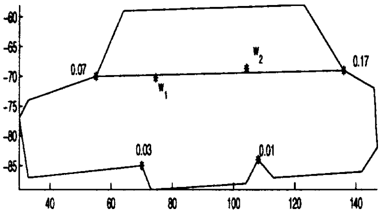


Figure 7.5 The center of gravity of the first window (w_1) is on the lower subpart of mainbody. This penalty is taken into account against false classification of parts that are on the lower mainbody as windows.

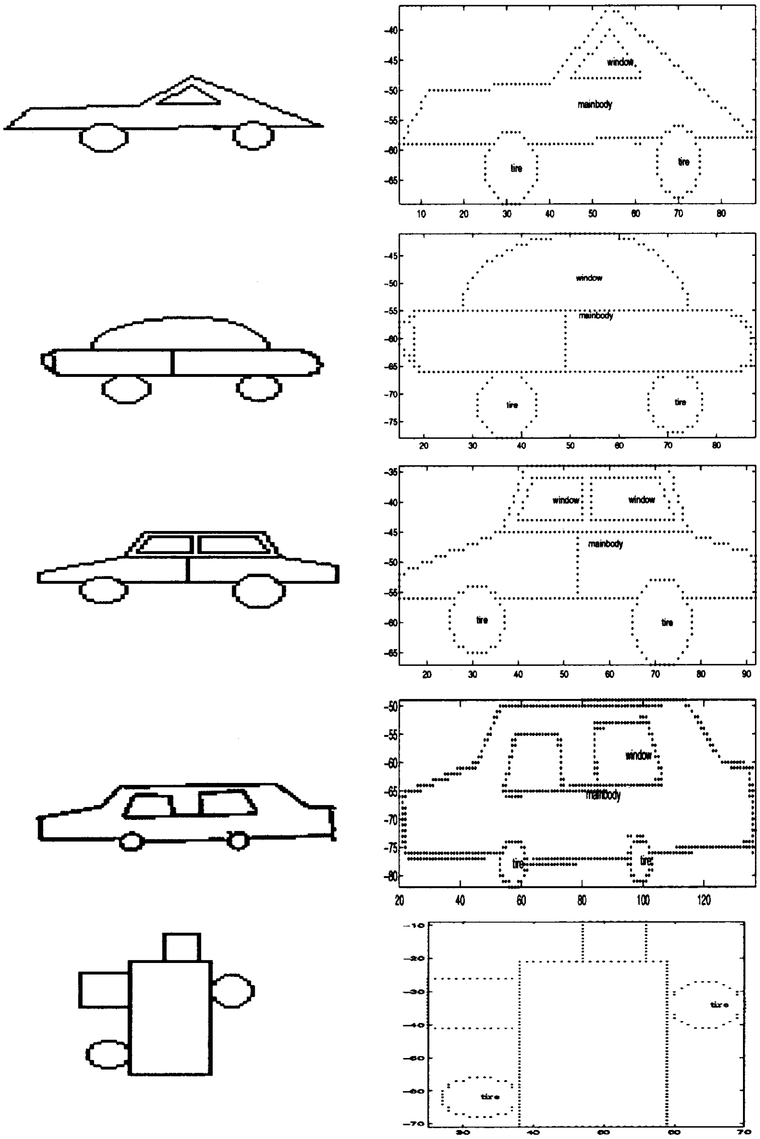


Figure 7.6 First row: Original sketch and matching result (all segments are correctly classified), number of segments(n_s)= 4, number of nodes(n_n)= 7, number of candidate branches(n_b)= 5, Second row: All segments are correctly classified, mainbody is the combination of three segments, n_s=6, n_n=19, n_b=19, Third row: All segments are correctly classified, mainbody is the combination of three segments, n_s=7, n_n=23, n_b=321, Fourth row: One window is missing, because the center of the gravity of the window is below the concavity line (Figure 7.5), n_s=5, n_n=8, n_b=5, Fifth row: For the non-car sketch the matched parts are not correct, the total segment number is 5 and the number of matched segments is 2 which is not sufficient to decide for the presence of OOI, n_s=5, n_n=20, n_b=48.

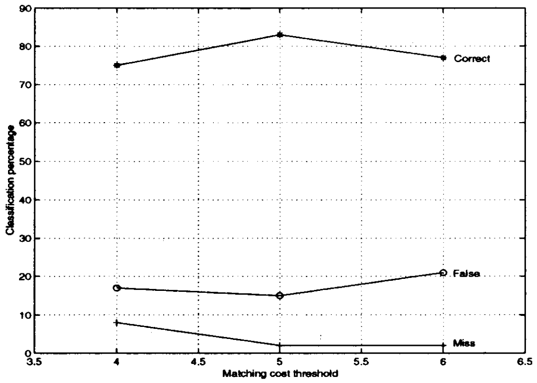


Figure 7.7 Classification percentage of correct, false detection and miss for nodes versus matching cost threshold for 18 sketches.

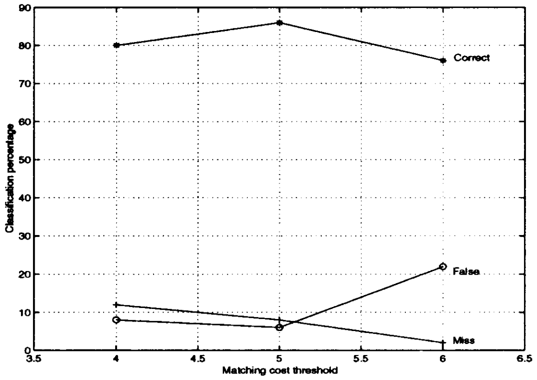


Figure 7.8 Classification percentage of correct, false detection and miss for nodes versus matching cost threshold for 22 real side images.

The result of the matching algorithm that gives the image segments and/or their combinations that are matched to car parts is observed. The number of correct matches divided to the total number of segments of the test images gives the percentage of correct classification of image segments. The falsely classified segments are the segments of the image that are not car parts matched to them. The miss percentage is computed from the observation of the algorithm failing to match the image segments that correspond to car parts that are used in the graph model. As seen in figures, the false classifications increase in average with increasing threshold while the miss percentage decreases. The optimum threshold for the given

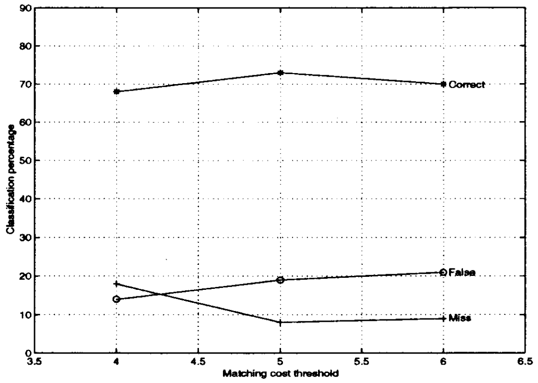


Figure 7.9 Classification percentage of correct, false detection and miss for nodes versus matching cost threshold for 18 real front-side images.

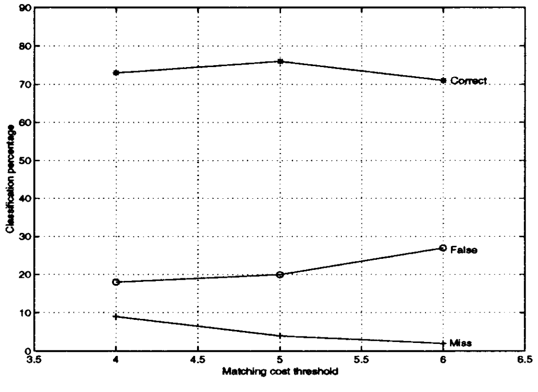


Figure 7.10 Classification percentage of correct, false detection and miss for nodes versus matching cost threshold for 15 real back-side images.

algorithm is found to be 5 for sketches and real images. Since the same weighting of dissimilarity is used, it is an expected result to have the same total cost threshold for every subgraph of aspect graph of the real images. Three examples for region classification for real images with busy and uniform backgrounds are displayed in Figure 7.11. Note the assumption that the foreground object is about the center of the image eliminates many background regions. However, there are still closed regions adjacent to the OOI. Due to the illumination and color changes the mainbody can be segmented to several regions but the combination of these regions gives the minimum matching cost in graph matching.

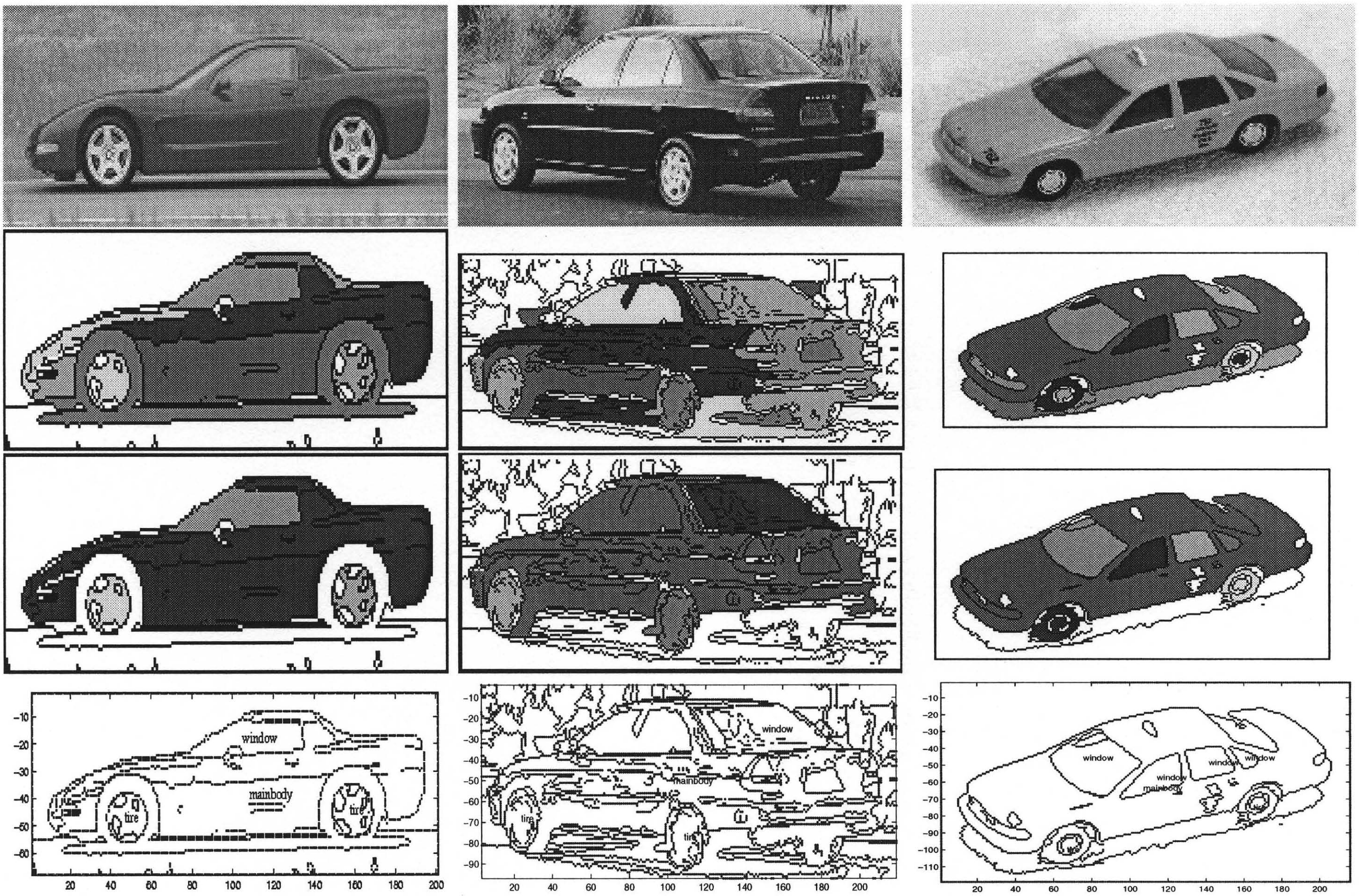


Figure 7.11 First row: Original images with busy and uniform backgrounds; Second row: Segmented images after taking only the closed regions at the center part of the image and after eliminating/merging small regions; Third row: Classification of nodes after graph matching; Fourth row: Resulting nodes.

In Figure 7.12, the performance of our method is shown for the Hamburg Taxi video sequence. Initial and final frames, extraction of the OOI from the sequence are shown in the first row. In the second row, the segmentation and matching results are displayed. In hierarchical object description, each segment helps us in handling the problems caused by occlusion. Figure 7.13 displays an example where the car is manually occluded to show the performance of the algorithm in the presence of partial occlusion.

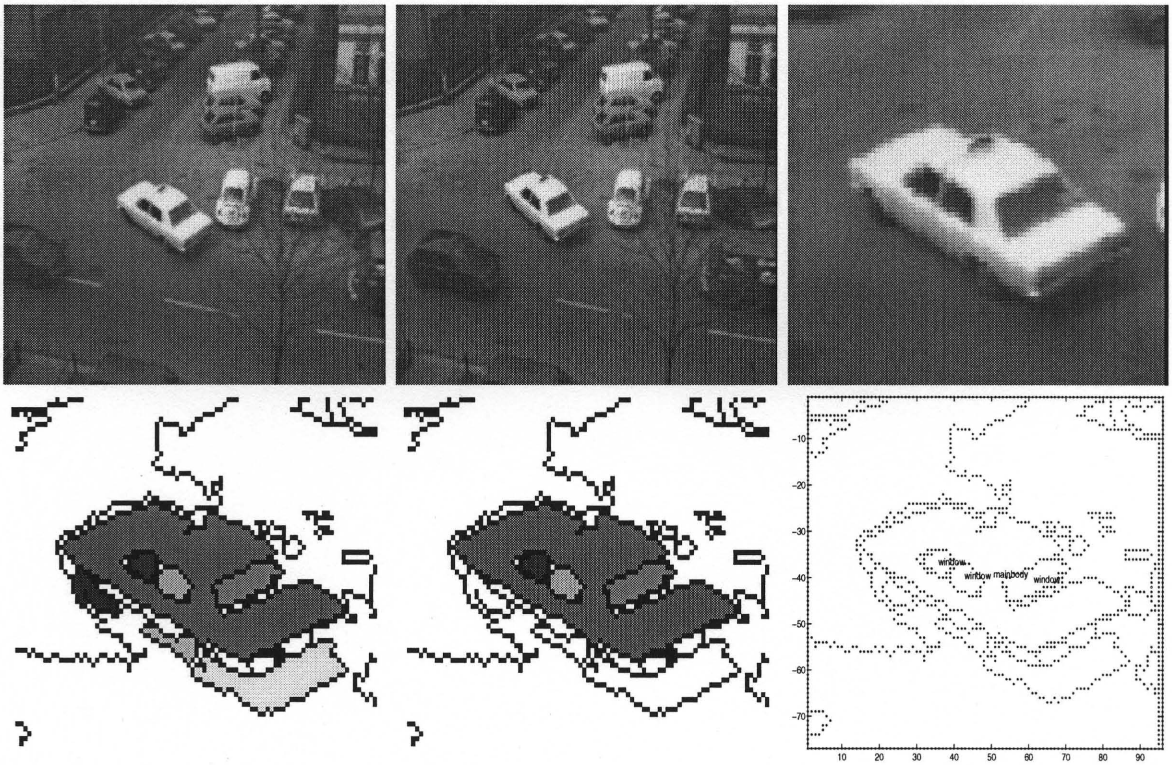


Figure 7.12 Separation of a moving car in Hamburg Taxi video sequence and matching result.

The most similar work to our graph-based object detection scheme is the object based retrieval system proposed by Xu et al [54]. The multi-level segmentation scheme used to create semantic features is similar to our model-based segmentation scheme. Our work mainly differentiates from the authors' algorithm at the bottom level of the segmentation tree. The authors form the root of the tree by grouping pixels similar in color therefore restricting the concept of homogeneity to color. Our approach is based on the observation that although complex objects can have shape and color variability within subparts of different objects, the relation between the subparts and the primitive shape characteristics are highly preserved. Therefore, our homogeneity concept is based on color and curvature, and the lowest level is formed of simplest visual parts in terms of curvature and color. Hence, the shape attributes

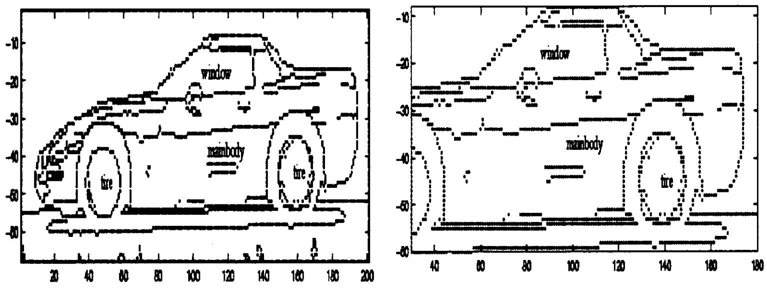


Figure 7.13 Manually occluded car and matching results. Left: No occlusion, all the parts are correctly classified; Right: 25% occlusion, all the parts except the front tire are correctly classified.

are chosen so that the fundamental shape characteristics are captured as opposed to B-spline fit of the complex region boundaries as described in [54]. The authors match a given query template (e.g. car main-body) to database images in a top-down fashion where the relation of other subparts is not used. However, although separating regions to primitive parts and combining them according to model-based segmentation increases the computational complexity, it enables to use the relations of basic subparts for a more robust detection scheme in terms of high-variability within class and occlusion. Furthermore, it is shown that the same graph-based object representation is suitable for non-rigid object detection i.e., human bodies with different postures since the lowest level of the tree can capture the articulated movements.

7.2 Compressed Domain

Figure 7.14 shows the line approximation and graph matching results for the 128 by 48 DCT image. Mainbody, window and one of the tires are correctly classified. Since the segmentation performance for the lowest resolution is not satisfactory, the graph matching algorithm also fails for this resolution level. Figure 7.15 shows our graph matching approach on the original and quantized DC-DCT coefficient images. The uncompressed, original image is given in the first row. Despite the highly accurate classification of the oversegmented object parts, the combination number and processing time increase exponentially. Classification result for the DCT image is given in the second row. In the experiments, 9 side-view car images are used. The graph matching threshold is set to 5. The classification percentage of correct and false detection and miss for these uncompressed side images are 82, 8, and 10, respectively. For DCT images, the correct classification decreases to 73%. 10% of the car parts are false and 17% of them are miss classified.

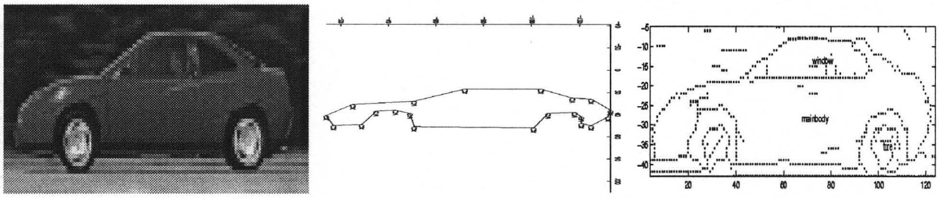


Figure 7.14 Left: Quantized DC-DCT coefficients; Middle: Line approximation and curvature points of the main body; Right: Graph matching result.

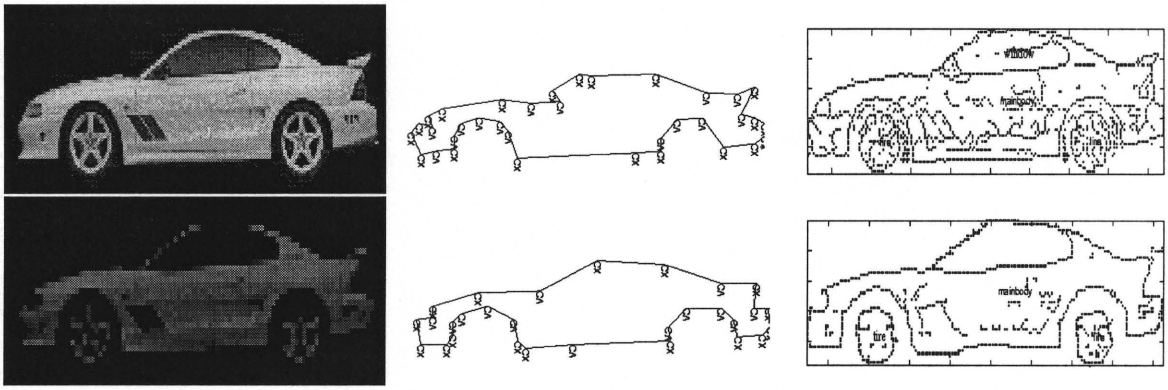


Figure 7.15 Top left: Original image, middle: Line approximation and curvature points of the main body, right: Graph matching result; Bottom left: Quantized DC-DCT coefficients, middle: Line approximation and curvature points of the main body, right: Graph matching result.

CHAPTER 8

EXPERIMENTAL RESULTS FOR NON-RIGID OBJECTS

This chapter includes detection of human followed by posture recognition and activity estimation in still images and video frames. The algorithms are implemented on the still images with OOI and on video sequences with moving OOI. The results for uncompressed and compressed domain techniques are given in the following sections, respectively.

8.1 Uncompressed Domain

The performance of the proposed algorithm for non-rigid objects is given for 42 test images with human bodies for front and side views which are chosen from different sources. Since bending deformation increases the computational complexity, its value is set to zero and the computations are done using the tapering deformation. An example model file is shown in Figure 8.2. In the model file, the adjacency information between parts is given as; head-torso, upper arm-torso, leg-foot, lower arm-hand, etc. For example, there is no adjacency restriction between hand and leg or hand and belly, since hand can be at any position near them. In the model file these combinations are also chosen: arm=upper arm+lower arm, legs=leg1+leg2, lowbody=legs+belly, upbody=torso+belly, armtorso=arm+torso. Another important issue in the model file generation is that the features, such as eccentricity, can show large deviations from person to person (thin-fat, big-small, etc.) for each body part. Furthermore, eccentricity of the limbs are close to each other. Hence, within-class scatter matrix can be large while between-class scatter matrix can be small which is the worst case for a classification. Under the assumption that feature vectors have Gaussian distribution, their mean and variance

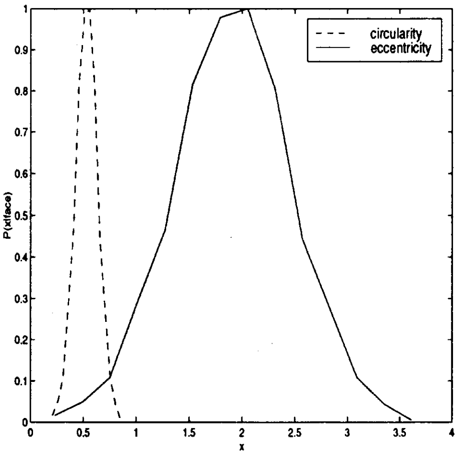


Figure 8.1 Distributions of two face features.

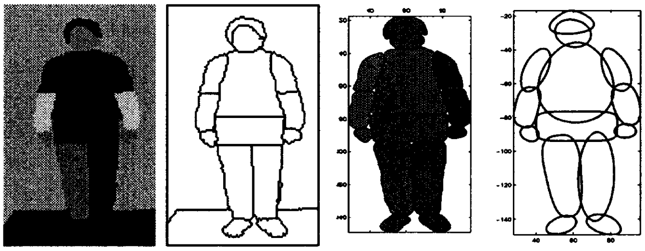


Figure 8.2 First: The skin areas are determined in the model color image. Second: Segmentation result. Third: Curvature segmentation results. Four: Fitted superellipses to the body parts.

are determined during supervised learning. Figure 8.1 displays the circularity and eccentricity distributions for face.

Results for segmentation and modeling with superellipses are displayed in Figure 8.3 for different test images. After graph matching, the classification results for three images in Figure 8.4 are given in Table 9. Note that, in Figure 8.4 d), an image with multi-persons is tested. Since the algorithm first determines the face regions, two separate branches for each face region are initialized. In the same image, the lower arms of the persons are folded on their upper arms where graph matching algorithm classifies them as upper arms. The overall algorithm performance is obtained by computing the correct, false, and miss detection of the body parts

model - image(a)	model - image(d)	model - image(e)
face - face	face - face(Right body (r.b.))	face - face
torso - torso	torso - torso(r.b.)	torso - torso
belly - belly	belly - belly(r.b.)	legs - legs
arm1 - arm1	uparm1 - lowarm1(r.b.)	
arm2 - arm2	uparm2 - lowarm2(r.b.)	
leg1 - leg1	leg1 - leg1(r.b.)	
leg2 - leg2	leg2 - leg2(r.b.)	
	face - face(Left body (l.b.))	
	torso - torso(l.b.)	
	belly - belly(l.b.)	
	uparm1 - lowarm1(l.b.)	
	uparm2 - lowarm2(l.b.)	

Table 8.1 Classification results for three test images.

in the test images. The preliminary results show that 70.27% of the body parts are correctly and 18.92% are falsely classified. The remaining 10.8% is the miss detection. In order to determine the posture of the persons in the still images and video sequences, the binary features of the corresponding matched node pairs are used after the classification. For example, the angle α between the image node matched to torso and image node matched to arm informs how much arms are open. Table 10 displays an example where both arms are open with an angle of 75-80 degrees, one leg is open with an angle of 40 degrees while other leg is approximately on the same axis as torso. Table 11 and 12 displays the angles between torso1-arms and torso2-legs for the multi-person image. Since the angles are very small, it can be easily determined that both of the persons have closed arms and closed legs where their arms and legs are approximately on the same axis of torso. Note that, posture recognition is a direct result of correct classification of the body parts.

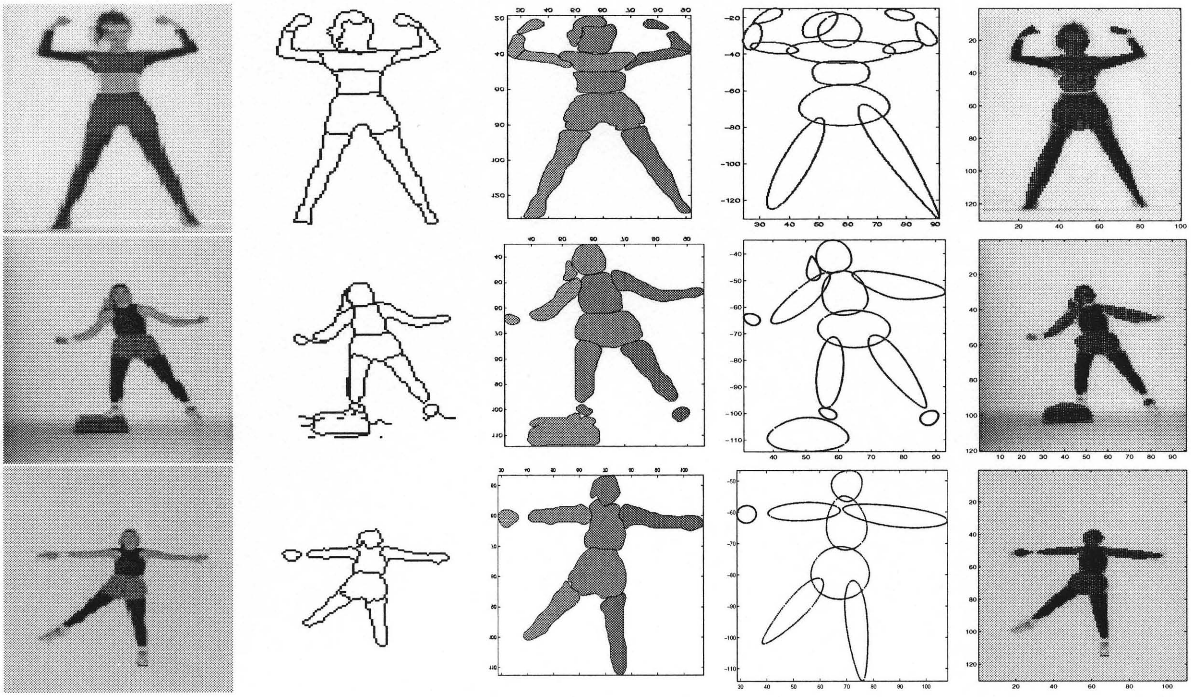
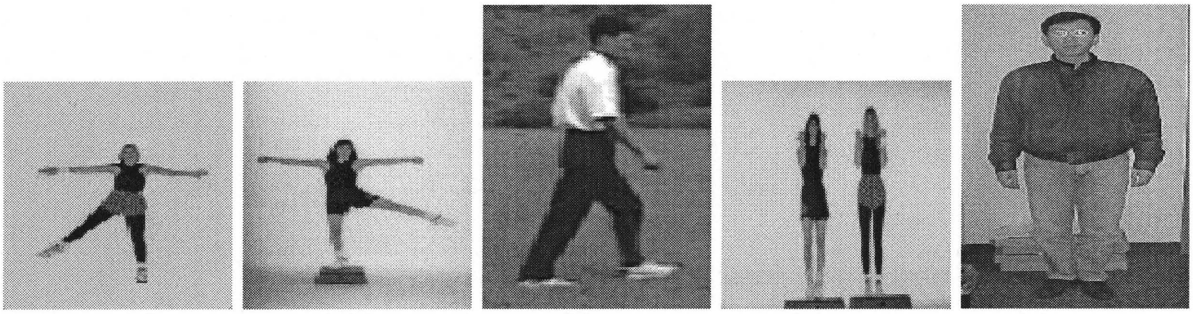


Figure 8.3 Column 1: Original image. Column 2: Segmentation result. Column 3: Part separation and curvature segmentation results. Column 4: Fitted superellipses. Column 5: Indexed superellipses on the body where superellipses with the skin areas are also determined.



(a) (b) (c) (d) (e)

Figure 8.4 Some test images. The detection performance for image a), d) and e) are given in Table 8.1

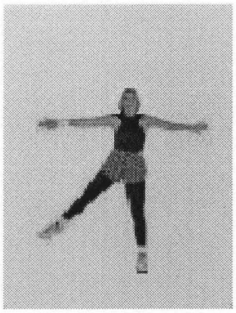


Figure 8.5 Test image.

part 1	part2	α
torso	arm 1	79.10
torso	arm 2	75.32
torso	leg 1	39.31
torso	leg 2	2.92

Table 8.2 α values ($\alpha = \Delta\theta$)

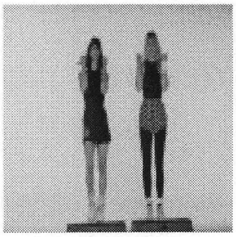


Figure 8.6 Test image.

part 1	part2	α
torso	arm 1	7.94
torso	arm 2	9.10
torso	leg 1	5.11
torso	leg 2	6.12

Table 8.3 α values for the left body.

part 1	part2	α
torso	arm 1	1.98
torso	arm 2	2.92
torso	leg 1	0.81
torso	leg 2	0.82

Table 8.4 α values for the right body.

8.2 Compressed Domain

To evaluate the system performance for the activity recognition in compressed domain, several sequences with different activities are used. Table 1 displays the resulting normalized distances (Eq. 6.7) between the activity sets and test sequences. The results show that MPEG motion vectors corresponding to three human body subregions can be used for detection and recognition of human activity. Each test sequence gives the minimum normalized distance with its corresponding training set. The last sequence is a MPEG car movie. Note that the distances are very high for each activity class. Another restriction for car sequences is that the human body ratio is not suitable for the car mainbody. The performance of the algorithm depends on the temporal duration of the observed activity. The results displayed in the table are given for sequences with two or more activity periods.

	Walking	Running	Kicking
walk1	0.001	0.0587	0.1543
walk2	0.0103	0.0929	0.0615
walk3	0.007	0.02	0.0784
walk4	0.0084	0.1218	0.1627
walk5	0.046	0.1506	0.1651
walk6	0.019	0.1298	0.208
run1	0.26677	0.0954	0.1688
run2	0.2525	0.0143	0.2519
run3	0.7665	0.027	0.1703
kick1	0.298	0.1253	0.0576
kick2	0.1901	0.109	0.0868
car	0.5362	0.4282	0.6922

Table 8.5 The normalized Euclidean distance between the activity sets and test sequences.

CHAPTER 9

CONCLUSIONS

This thesis is a comprehensive study of object-based image and video retrieval, specifically for car and human detection and activity recognition purposes. The thesis focuses in the problem of connecting low level features to high level semantics by developing relational object and activity presentations.

The thesis first examines extraction of low level features from images and videos using intensity, color and motion of pixels and regions. Local consistency based on these features and geometrical characteristics of the regions is used to group object parts. The problem of managing the segmentation process is solved by a new approach that uses object based knowledge in order to group the regions according to a global consistency. A new model-based segmentation algorithm is introduced that uses a feedback from relational representation of the object. The selection of shape attributes is addressed in Chapter 5. These unary and binary attributes are further extended for application specific algorithms: an elaborate human skin color model, boundary shape code for rigid objects and weak perspective invariants for articulated movements. Object detection is achieved by matching the relational graphs of objects with the reference model. The algorithm maps the attributes, interprets the match and checks the conditional rules in order to index the parts correctly. The major advantages can be summarized as improving the object extraction by reducing the dependence on the low level segmentation process and combining the boundary and region properties. Furthermore, the features used for segmentation are also attributes for object detection in relational graph representation. This property enables to adapt the segmentation thresholds by a model-based training system. The detection rate corresponds to correct classification of object parts. The detection rate is 83%

for free-hand car sketches and 86%, 73% and 76% for real car images viewed from side, front-side and back-side respectively. The test data set includes images and sequences from different sources and at different resolutions and occlusions. The detection rate for human body parts is 70.27% for images and sequences including human body regions at different resolutions and with different postures.

The thesis then addressed the problem of object detection and activity recognition in compressed domain in order to reduce computational complexity. A new algorithms for object detection and activity recognition in JPEG images and MPEG videos is developed and it is shown that significant information can be obtained from the compressed domain in order to connect to high level semantics. Since our aim is to retrieve information from images and videos compressed using standard algorithms such as JPEG and MPEG, our approach differentiates from previous compressed domain object detection techniques where the compression algorithms are governed by characteristics of object of interest to be retrieved. An algorithm is developed using the principal component analysis of MPEG motion vectors from the P-frames to detect the human activities; namely, walking, running, and kicking. The algorithm is tested for sequences without camera motion. The distances of expansion coefficients between six sequences of walking people, three sequences of running people and two sequences of kicking people are presented to demonstrate that the classification among activities is clearly visible.

Object detection in JPEG compressed still images and MPEG I frames is achieved by using DC-DCT coefficients of the luminance and chrominance values. The graph-based object detection is tested for JPEG side-view car images. The correct part classification is 73% while 10% is falsely classified and 17% is missed. The performance is dependent on the resolution especially for human detection where skin region extraction is crucial. For lower resolution and monochrome images it is demonstrated that the structural information of human silhouettes can be captured from AC-DCT coefficients. In order to train our system, 800 positive (human) examples and different negative (non-human) examples with a bootstrapping algorithm are used. The overall system performance is tested on 40 images that contain a total of 126 non-occluded frontal poses and the algorithm can detect 101 of them correctly. The major contribution of the overall algorithm is to connect available data in compressed and uncompressed domain to high level semantics. The proposed hierarchical scheme enables working at different levels, from low complexity to low false rates.

REFERENCES

1. I. B. Ozer, W. Wolf, and A. N. Akansu, "A Graph Based Object Description for Information Retrieval in Digital image and Video Libraries", CBAIVL, pp. 79-83, June 1999.

2. R. N. Haber, M. Hershenson, The Psychology of Visual Perception, Holt, Rinehart and Winston Inc, 1973.

3. L. Zusne, "Visual Perception of Form", Academic Press, B200, New York, 1970.

4. D. D. Hoffman and W.A. Richards, "Parts of Recognition", Cognition, vol. 18, pp. 65-96, 1984.

5. D. O. Hebb, "The Organization of Behaviour", Wiley, New York, 1949.

6. S. Loncaric, "A Survey of Shape Analysis Techniques", Pattern Recognition, vol. 31, no. 8, pp. 983-1001, 1998.

7. S. Birchfield, "Elliptical Head Tracking Using Intensity Gradients and Color Histograms", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, California, pp. 232-237, June 1998.

8. J. R. Bennet and J.S. McDonald, "On the Measurement of Curvature in a Quantized Environment", IEEE Trans. on Comput., vol. 24, pp. 803-820, 1975.

9. E. M. Arkin, L. P. Chew, D. P. Huttenlocher, K. Kedem and J. S. B. Mitchell, "An efficiently computable metric for comparing polygonal shapes", IEEE Trans. Pattern Anal. Mach. Int, vol. 13, pp. 209-216, 1986.

10. L. J. Latecki and A. Rosenfeld, "Supportedness and tameness: Differentialless geometry of plane curves", Pattern Recognition, vol. 31, pp. 607-622, 1998.

11. C. C. Chang, S. M. Hwang, and D. J. Buehrer "A Shape Recognition Scheme Based on Relative Distances of Feature Points from the Centroid", Pattern Recognition, vol. 24, pp. 1053-1063, 1991.

12. A. Jain, Y. Zhong, and S. Lakshmanan, "Object Matching Using Deformable Templates", *IEEE Trans. Pattern Analysis Mach. Intell.* , pp. 408-439, March 1996.

13. H. Freeman, "On the Encoding of Arbitrary Geometric Configurations", IRE Transactions, vol. 10, pp. 260-268, 1961.

14. E. Persoon and K. S. Fu, "Shape Discrimination Using Fourier Descriptors", *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 8, pp. 388-397, 1986.

15. S. Wang, P. Chen, and W. Lin, "Invariant Pattern Recognition by Moment Fourier Descriptor", Pattern Recognition., vol. 27, pp. 1735-1742, 1994.

16. A. Bengtsson and J. Eklundh, "Shape Representation by Multiscale Contour Approximation", IEEE PAMI, vol. 13, pp. 85-93, 1991.

17. F. S. Cohen, Z. Huang, and Z. Yang, " Invariant Matching and Identification of Curves Using B-splines Curve Representation", *IEEE Trans. on Image Processing*, vol. 4, pp. 1-10, 1995.

18. B. Gunsel and A. M. Tekalp, "Shape Similarity Matching for Query by Example", *Pattern Recognition*, vol. 31, no. 7, pp. 931-944, July 1998.

19. A. P. Witkin, "Scale Space Filtering", Proc. 8th Int. Joint Conf. on Artificial Intelligence, pp. 1019-1022, 1983.

20. H. Asada and M. Brady, "The Curvature Primal Sketch", IEEE PAMI, vol. 8, pp. 2-14, 1986.

21. F. Mokhtarian and A. K. Mackworth, "A Theory of Multiscale, Curvature-Based Shape Representation for Planar Curves", IEEE PAMI, vol. 14, pp. 789-805, 1992.

22. L. J. Latecki and R. Lakamper, "Convexity Rule for Shape Decomposition Based on Discrete Contour Evol.ution", *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 441-454, 1999.

23. J. L. Mundy and A. Zisserman, "Geometric Invariance in Computer Vision", MIT Press, 1992.

24. M. K. Hu, "Visual Pattern Recognition by Moment Invariants", IRE Trans. Inform. Theory, vol. 8, pp. 179-187, 1962.

25. H. Blum and R. Nagel, "Shape Description Using Weighted Symmetric Axis Features", Pattern Recognition, vol. 10, pp. 167-180, 1978.

26. K. Siddiqi, A. Shokoufandeh, S. J. Dickinson and W. Zucker, "Shock Graphs and Shape Matching", Technical Report, 1998.

27. F. Leymarie and M. D. Levine, "Simulating the Grassfire Transform Using an Active Contour Model", PAMI, vol. 14, pp. 56-75, 1992.

28. F. Mokhtarian and A. Mackworth, "Scale Based Description and Recognition of Planar Curves and 2D Shapes", *IEEE PAMI*, vol. 8, no. 1, pp. 34-43, 1986.

29. A. H. Barr, "Superquadrics and Angle Preserving Deformations," *IEEE Computer Graphics Applications*, vol. 1, pp. 11-23, 1981.

30. F. Solina and R. Bajcsy, "Recovery of parametric models from range images: the case for superquadrics with global deformations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 2, pp. 131-147, Feb. 1990.

31. M. Bennamoun, R. Boashash, "A Vision System for Automatic Object Recognition," *Proc. of IEEE International Conference on Systems, Man, and Cybernetics, 1994. Humans, Information and Technology.*, Oct. 1994.

32. M. Bennamoun, B. Boashash, "A Structural-Description-Based Vision System for Automatic Object Recognition", IEEE Transactions on Systems, Man, and Cybernetics-Part B, Cybernetics, vol. 27, no. 6, Dec. 1997.

33. R. Jain, "Workshop Report: NSF Workshop on Visual Information Management Systems", *Proc. SPIE Conf. on Vis. Commun. and Image Proc.*, 1993.

34. R. Jain, A. Pentland, and D. Petkovic, "NSF-ARPA Workshop on Visual Information Management Systems", Cambridge, MA, June 1995.

35. M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafine, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by Image and Video Content: The QBIC System", *IEEE Computer*, 1995.

36. J. Dowe, "Content-based Retrieval in Multimedia Imaging", *Proc. SPIE Conf. on Vis. Commun. and Image Proc.*, 1993.

37. B. Furht, S. W. Smoliar, and H. Zang, "Video and Image Processing in Multimedia System", Kluwer Academic Publishers, 1995.

38. R. W. Picard and T. P. Minka, "Vision Texture for Annotation", MIT Multimedia Laboratory Perceptual Computing Section TR no. 302, 1995.

39. S. Scraloff and A. Pentland, "Modal Matching for Correspondence and Recognition", *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 17, pp. 545-561, 1995.

40. J. R. Smith and S. F. Chang, "VisualSEEk: A Fully Automated Content-Based Image Query System", *Proc. ACM Multimedia Conf.*, pp. 87-98, Boston, 1996.

41. S. F. Chang, W. Chen, and H. Sundaram, "Semantic Visual Templates - Linking Visual Features to Semantics", *Proc. Int. Conf. on Image Proc.*, 1998.

42. H. Yu, W. Wolf, "A Visual Search System for Video and Image Databases", *Proc. IEEE Multimedia*, pp. 517-524, 1997.

43. J. Zhang, H. Krim, and X. Zhang, "Invariant Object Recognition by Shape Space Analysis", *Proc. Int. Conf. on Image Proc.*, 1998.

44. T. S. Huang, S. Mehrotra, and K. Ramchandran, "Multimedia Analysis and Retrieval System(MARS) Project", *Proc. of 33rd Annual Clinic on Library Application of Data Processing-Digital Image Access and Retrieval*, 1996.

45. A. Yoshitaka, T. Ichikawa, "A survey on Content-Based Retrieval for Multimedia Databases", IEEE Trans. on Knowledge and Data Eng., vol. 11, no. 1, pp. 81-92, Jan/Feb. 1999.

46. A. Gupta, R. Jain, "Visual Information Retrieval", Communications of ACM, vol. 40, no. 5, pp. 70-79, May 1997.

47. A. Pentland, R. Picard, and S. Sclaroff "Photobook: Tools for Content Based Manupulation of Image Databases", Storage and Retrieval of Image and Video Databases II, Paper no. 2185-05, San Jose, Calif., pp. 34-47, SPIE, Feb. 1994.

48. A. Pentland, R. Picard, and S. Sclaroff, "Photobook: Content-based Manipulation of Image Databases", International Journal of Computer Vision, 1996.

49. V. E. Ogle and M. Stonebraker, "Chabot: Retrieval from a Relational Database of Images", *Computer*, vol. 28, no. 9, Sept. 1995.

50. J. R. Smith and S. F. Chang, "Visually Searching the Web for Content", IEEE Multimedia, vol. 4, no. 3, pp. 12-20, July/Sept. 1997.

51. W. S. Li and K. S. Candan, "SEMCOG: A Hybrid Object-based Image Database System and Its Modeling, Language, and Query Processing", Proceedings of the 14th International Conference on Data Engineering, pp. 284-291, Feb. 1998.

52. M. P. Dubuisson, S. Lackshmanan, and A. K. Jain, "Vechicle Segmentation and Classification Using Deformable Templates", *IEEE Trans. Pattern Analysis Mach. Intell.* , pp. 293-307, March 1996.

53. M. P. Dubuisson, A. K. Jain, and W. C. Taylor, "Segmentation and Matching of Vehicles in Road Images", *Transportation Research Report*, no. 1412, pp. 57-63.

54. Y. Xu, E. Saber, and A. M. Tekalp, "Object Formation by Learning in Visual Databases Using Hierarchical Content Description", *Proc. Int. Conf. on Image Proc.*, Oct. 1999.

55. C. P. Papageorgiou, T. Poggio, "A Trainable Object Detection System: Car Detection in Static Images", Technical paper, MIT, CBCL Paper no. 180, Oct. 1999.

56. J. K. Aggarwal and Q. Cai, "Human Motion Analysis: A Review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428-440, March 1999.

57. U. Franke and D. Gavrila, "Autonomous Driving Goes Downtown," *IEEE Intelligent Systems*, vol. 13, no. 6, pp. 40-48, Nov. 1998.

58. S. F. Chang and D. G. Messerschmitt, "Manipulation and Compositing of MC-DCT Compressed Video," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 1, pp. 1-11, Jan. 1995.

59. C. P. Papageorgiou, M. Oren and T. Poggio, "Pedestrian Detection Using Wavelet Templates," *Proc. of CVPR*, Puerto Rico, June 1997.

60. D. M. Gavrila, "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82-98, Jan. 1999.

61. D. Schonfeld and D. Lelescu, "VORTEX: Video retrieval and tracking from compressed multimedia databases - template matching from MPEG2 video compressed standard", SPIE Conference on Multimedia and Archiving Systems III, Nov. 1998.

62. Y. Zhong, H. Zhang, A. K. Jain, "Automatic Caption Localization in Compressed Video", IEEE PAMI, vol.22, no. 4, pp. 385-392, April 2000.

63. H. Wang and S. F. Chang, "A Highly Efficient System for Automatic Face Region Detection in MPEG Video Sequences", IEEE Trans. on Circuits and Systems for Video Technology, special issue on Multimedia Systems and Technologies, vol. 7, no. 4, pp. 615-628, Aug. 1997.

64. J. R. Smith and S. F. Chang, "Querying by Color Regions Using the VisualSEEK Content-Based Visual Query System", Intelligent Multimedia Information Retrieval, Editor M. T. Maybury, AAAI/MIT Press, 1997.

65. W. K. Pratt, "Digital Image Processing", J. Wiley and Sons, Second Edition, 1991.

66. J. Malik, D. A. Forsyth, M. M. Fleck, T. Leung, C. Carson, S. Belongie, C. Bregler, "Finding Objects in Image Databases by Grouping", IEEE, pp. 761-764, 1996.

67. B. Moghaddam, H. Biermann, D. Margaritis, "Defining Image Content with Multiple Regions of Interest", CBAIVL, pp. 89-93, June 1999.

68. E. Saber, A. M. Tekalp, "Integration of Color, Edge, shape, and Texture Features for Automatic Region-Based Image Annotation and Retrieval", ICIP, vol. 3, pp. 851-854, 1996.

69. H. Wu, Q. Chen, and Y. Yachida, "Face Detection From Color Images Using a Fuzzy Pattern Matching Method", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 6, pp. 557-562, 1993.

70. M. H. Yang, N. Ahuja, "Detecting Human Faces in Color Images", ICIP, vol. 1, pp. 127-130, 1998.

71. D. A. Forsyth, M. M. Fleck, "Identifying Nude Pictures", *3rd IEEE Workshop on Applications of Compure Vision*, pp. 103-108, 1996.

72. R. W. Picard, "A Society of Models for Video and Image Libraries", MIT Multimedia Laboratory Perceptual Computing Section Technical Report no. 360, 1996.

73. M. M. Yeung, B. L. Yeo, W. Wolf and B. Liu , "Video Browsing using Clustering and Scene Transitions on Compressed Sequences", SPIE vol. 2417 Multimedia Computing and Networking, pp. 399-413, 1995.

74. A. M. Dawood and M. Ghanbari, "Scene Content Classification From Mpeg Coded Bit Streams", IEEE Workshop on Multimedia Signal Processing, pp. 253-258, 1999.

75. H. Yu, W. Wolf, "Let's Video Freely- Automatic Video Indexing for Film and TV Program oriented Digital Video Library", pp. 217-222.

76. Y. Yacoob and M. J. Black, "Parameterized Modeling and Recognition of Activities", ICCV, pp.120-127, 1998.

77. M. Walter, S. Gong, A. Psarrou, "Stochastic temporal Models of Human Activities", *International Workshop on Modelling People*, pp. 87-94, 1999.

78. K. Rangarajan, W. Allen, M. Shah, "Matching Motion Trajectories Using Scale-Space", *Pattern Recognition*, vol. 26, no. 4, pp. 595-610.

79. R. Cutler and L. Davis, "Real-Time Periodic Motion Detection, Analysis and Applications", pp. 326-332, 1999.

80. C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, W. von Seelen, "Walking Pedestrian Recognition", *International Conference on Intelligent Transportation Systems*, pp. 292-297, 1999.

81. S. F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, " VideoQ: An Automatic Content-Based Video Search System Using Visual Cues, *ACM Multimedia '97 Conference*, Seattle, Nov. 1997.

82. M. Kurokawa, T. Echigo, A. Tomita, J. Maeda, H. Miyamori and S. Iisaku, "Representation and Retrieval of Video Scene by Using Object Actions and their Spatio-temporal Relationships, " ICIP, pp. 86-90, 1999.

83. H. Miyamori, S. Iisaku, "Video Annotation for Content-based Retrieval using Human Behavior Analysis and Domain Knowledge", *International Conference on Automatic Face and Gesture Recognition*, pp 320-325, 2000.

84. Y. P. Tan, D. D. Saur, S. R. Kulkarni, and P. J. Ramadge, "Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 10, no. 1, pp. 133-146, Feb. 2000.

85. V. Kobla, D. DeMenthon, D. Doermann, "Identifying Sports Videos Using Replay, Text, and Camera Motion Features", *Workshop on Multimedia Signal Processing*, 1999.

86. J. Shi and C. Tomasi, "Good Features to Track", CVPR, 1994.

87. K. Harris, S. N. Efstratiadis, N. Maglaveras, and A. K. Katsaggelos, "Hybrid Image Segmentation Using Water Sheds and Fast Region Merging", *IEEE Trans. on Image Processing*, vol. 7, pp. 1684-1699, 1998.

88. J. B. Burns, R. S. Weiss and E. M. Riseman, "View Variation of Point-Set and Line-Segment Features", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 1, pp. 51-68, 1993

89. M. Nagao, T. Matsuyama, Y. Ikeda,"Region Extraction and Shape Analysis in Aerial Photographs", CGIP, pp. 195-223, 1979

90. M. Kass, A. P. Witkin and D. Terzopoulos. "Snakes: Active contour models", Inter. Journal on Comp. Vision, vol. 1, no. 4, pp. 321-331, 1988

91. H. S. Ip and D. Shen,"An Affine Invariant Active Contour Model for Model-Based Segmentation", IVC, pp. 135-146, 1998

92. L. Liu, S. Sclaroff,"Deformable Shape Detection and Description via Model-Based Region Grouping", CVPR, 1999

93. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Numerical Recipes in C ", Cambridge University Press, Second Edition, 1995.

94. R. O. Duda, and P. E. Hart, "Pattern Classification and Scene Analysis ", John Wiley and Sons, 1973.

95. M. A. Turk and A. P. Pentland, "Face Recognition Using Eigenfaces"

96. C. Loeffler, A. Ligtenberg, and G. S. Moschytz, "Practical Fast 1-D DCT Algorithms with 11 Multiplications", ICASSP, pp. 988-991, 1989.

97. S. F. Chang, J. R. Smith, M. Beigi, and A. B. Benitez, "Visual Information Retrieval from Large Distributed On-line Repositories", Communications of the ACM, vol. 40, no. 12, pp. 63-71, 1997.

98. D. H. Ballard and C. M. Brown, "Computer Vision", Prentice-Hall, Englewood Cliffs, NJ, 1982.

99. T. Caelli and W. F. Bischof, "Machine Learning and Image Interpretation", Plenum Press, New York, NY, 1997.

100. R. M. Haralick and L. G. Shapiro, "Computer and Robot Vision", Addison Wesley Publishing Co., 1993.

101. R. A. Brooks, "Model-Based Three Dimensional Interpretations of Two Dimensional Images", PAMI, vol. 5, pp. 140-150, 1983.

102. B. L. Yeo and B. Liu , "On the extraction of DC sequence from MPEG Compressed Video", ICIP, Oct. 1995.

103. H. Wang, H. S. Stone, and S. F. Chang, "FaceTrack: Tracking and Summarizing Faces from Compressed Video", *SPIE Multimedia Storage and Archiving Systems IV*, 19-22 Sept., Boston.

104. M. Kirby and L. Sirovich, "Application of the Karhumen-Loeve Procedure for the Characterization of Human Faces", IEEE PAMI, vol. 12, no. 1, pp.103-108, 1990.

105. M. Turk and A. Pentland, "Face Recognition Using Eigenfaces", CVPR, pp. 586 -591, 1991.