WP-11: INTEGRATION:
State of the Art in Automatic Human Detection, Motion and Behaviour Analysis in Multimedia Data

1. Introduction

The goal of the NoE MUSCLE is to automatically extract semantic information from multimedia data about people and their interactions with complex and natural environments. Human behaviour understanding has a wide range of future applications ranging from intelligent surveillance systems to advanced user interface systems [1-8]. Surveillance cameras and sound recording systems are already available in banks, hotels, stores and highways, shopping centers and the captured video data are monitored by security guards and stored in archives for forensic evaluation. In a typical system, a security guard watches 16 video channels at the same time and may miss many important events. It would be desirable to have an intelligent system with capability of analysing the video in real time and warning the security staff or a semi-automatic system high-lighting possible regions of interest in the viewing area of the surveillance cameras for crime prevention, crowd control [8]. Surveillance of elderly and ill people in their homes is another important application area requiring automatic video analysis. Real time detection of a fallen person is extremely important for immediate emergency assistance and it may increase the quality of life of elderly and sick people. The sound of a fallen body may be necessary to achieve a robust automatic system. Other applications include various body parts segmentation to track the movement of joints in a video for medical diagnosis, treatment support, access control, and athletic, artistic, and musical performance improvement [9-15].

Development of digital libraries will require advanced user interface systems, which can interpret multimedia data in an automatic manner. Advanced user interface systems using speech and human motion understanding may provide control and command in high-level interaction with computerized systems and multimedia databases.

Early work on image retrieval systems is based on text input, in which the images are annotated by text and retrieval is performed on text [16]. However, manual annotation is labor-intensive and becomes impractical when the collection is large. Another problem is the subjective nature of keywords. The same image or video may be annotated differently with different annotators. Therefore, it is desirable to have a video database management system, including a Web-based graphical query interface, which can handle queries involving spatio-temporal and semantic properties of video data [17]. A spatio-temporal query may contain any combination of directional, topological, 3D-relation, object-appearance, trajectory-projection and similarity-based object-trajectory conditions. In this way, videos containing human activity and images containing humans in a database can be retrieved in an efficient manner. Such a system can interact with the user over the Internet through a graphical query interface as well.

It is also desirable to have a multimodal natural language interfaces for web search engines. The user will be able to perform web queries using natural language or speech input to the system. Such a system can engage the user in multimodal dialogue trying to further constrain or relax the query and better match the user's search requirements and it will be important to summarize the web query results and create visualizations effects.

Natural language processing and speech understanding has already been used in early human-machine interfaces. Vision is very useful to complement speech recognition and natural language understanding for more natural and intelligent communication between human and machines.

More detailed cues can be obtained by gestures, body poses, facial expressions, etc [18-22]. Hence, future systems must be able to independently sense the surrounding environment, e.g., detecting human presence and interpreting human behavior.

In order to extract semantic information humans in the video have to be detected as a first step. Humans can be stationary or moving. Detection of stationary persons in the video is equivalent to detection of persons in still images. This involves detection of human body parts including face, human body, hands etc. in a given image. If there are moving person(s) in the video then the problem of human detection is easier in the sense that one can extract moving object boundaries and perform human or body part detection inside the object boundary. In addition, object boundaries contain additional information for object classification and detection. In order to reach a robust decision, one can take advantage of the associated audio and text (if available) as well. If there is human speech in the audio portion of the video then this may be used as an additional clue to reach a final decision. Therefore, following problems should to be solved in a semantic information system before extracting useful knowledge from multimedia data:

- Stationary human and human body parts detection in image and video,
- moving object detection and classification in video,
- tracking moving objects in video and gait analysis,
- speech detection in audio, and annotation of video using associated speech data, and
- analysis of the available text for human behaviour understanding in video.

Although there are practical face detection and recognition systems in controlled environments above problems, unfortunately, are not completely solved in complex natural scenes. Therefore, solutions to above problems will be also studied within the NoE MUSCLE to achieve a human behavioural understanding system.

Stationary human detection in still images or video aims at segmenting regions corresponding to people from the rest of an image. The process of moving human detection in video process involves segmentation of moving regions in an image frame of the video and object classification. In this report, moving object and human detection process in video is reviewed in Section 2. In Section 3, human face and body detection in still images and video is described. Current research on multimodal data analysis for semantic information extraction from multimedia archives and semantic information extraction in video are reviewed in Sections 4 and 5, respectively. Future Research Directions are described in Section 6.

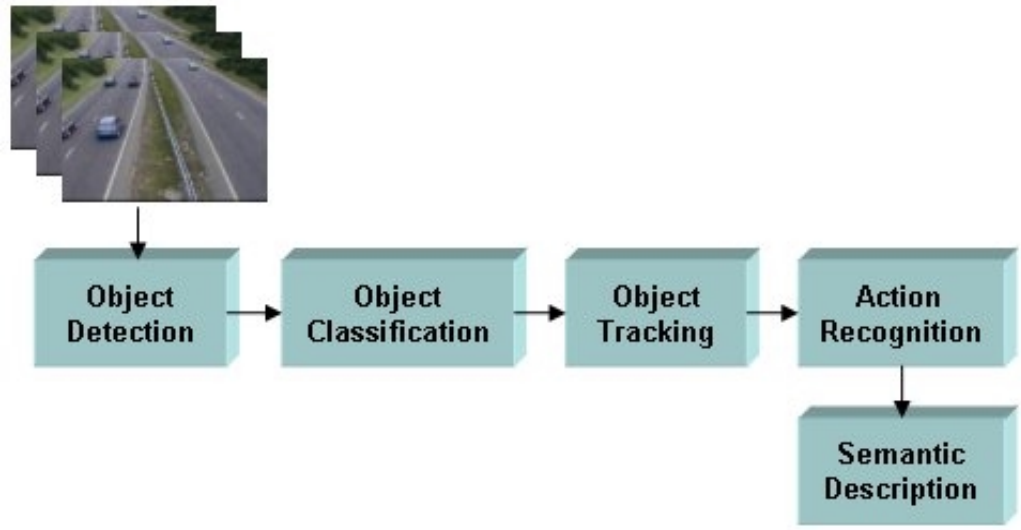

Figure 1: Block diagram of a typical smart video analysis system.

2. Human Detection in Video

A block diagram of a typical smart video analysis system is shown in Figure 1. Given a digital video signal the first step is to detect regions of interest, which are usually moving objects in many practical problems.

Detecting moving regions in the form of image blobs provides a narrower region of interest for later processes such as tracking and activity analysis because only the pixels of the moving region need be analyzed. This is not a trivial problem because of changes in illumination conditions, shadows in natural scenes.

Background estimation based human detection methods are widely used in moving object detection in video [1-4,23-27]. In these methods the background is adaptively estimated from past image frames and pixels of the moving objects are estimated by subtracting the current image frame of the video from the estimated background.

Background of the video is defined as the union of all stationary objects and the foreground consists of moving objects. A simple approach for estimating the background image is to average all the past image frames of the video because the effect of moving objects will be negligible in the long run with the assumption that the camera is stationary. In [1-3], the current background of the video is recursively estimated from past image frames using recursive first order Infinite-duration Impulse Response (IIR) filters acting on each pixel of the video in a parallel manner. By using two IIR filters with different update parameters in parallel two different background images can be estimated as well. A statistical background estimation method is described in the article by C. Stauffer et al. [23] in which, each pixel is modeled as a mixture of Gaussians and the model is updated in an iterative manner. This method results in a reliable, real-time outdoor tracker, which can deal with lighting changes and clutter. Order statistical methods were also developed for background estimation. Yang and Levine's [24] algorithm uses the median value of the current pixel over a series of past images. In [4], a statistical model is constructed by representing each pixel with its minimum and maximum intensity values, and the maximum intensity difference between consecutive frames observed during a training period. The model parameters were updated periodically. The numerous methods described in the literature to the background

estimation problem differ in the type of a background model and the procedure used to update the background model. In all of these methods, pixels of the foreground objects are estimated by subtracting the current image from the estimated background image. Moving blobs are constructed from the pixels by performing a connected component analysis.

If the camera is unstationary then a camera motion compensation algorithm has to be also used. Other class of moving object estimation methods includes optical flow estimation based methods [25-27]. Such methods use the optical flow vectors of moving objects over time to detect change regions in an image sequence.

2.1 Moving Object Tracking

Moving blobs should be tracked in video to achieve robust recognition results. Object tracking is necessary for human behavior analysis and recognition in video. Basically, tracking of moving blobs over sequence of images involves associating detected moving objects to each other in consecutive frames using various features of the blobs. Some tracking methods can also predict the location of the moving blob in the next image frame.

Standard tracking algorithms are based on statistical methods of correlation, Kalman filtering [28], condensation algorithm [29], mean-shift tracking [30], and dynamic Bayesian network [31], particle filtering [28] etc. Correlation based methods work well if there is no scale change in the tracked objects. Kalman filtering based methods are developed for radar tracking problems and they work very well in tracking point targets [28].

The condensation algorithm [29] was recently developed and it is based on conditional density propagation over consecutive image frames. In this tracking algorithm a posterior distribution estimated in the previous frame is extended to upcoming frames in an iterative manner. A dynamic model is combined with visual observations to achieve highly robust object tracking in video. However, it needs a large number of samples to estimate distributions to have good maximum likelihood estimates of the current state.

Mean shift tracking is another recently developed method for tracking moving objects in video [30],[126]. It is our belief that mean-shift tracking is an important tool in any tracking system. It is based on normalized and smoothed color histogram of the moving objects. Color histogram smoothing is essentially equivalent to so-called kernel density estimation process described in [30]. The standard mean-shift tracking algorithm determines the location of the moving object in the next image frame through an iterative process. The normalized histogram of a moving object in the n-th video frame is expressed as

$$H_n(k) = (1/N) \sum_{s \in O} \delta(q(s)-k)$$

where s represents the wavelet domain data, O represents a moving object, δ is the Kronecker-delta function, N is the number of data points in O, q is the quantizer function mapping the wavelet domain data into a 16 bit number. The histogram H_n, which is constructed from color pixel information, distinguishes the moving object O from the background and other moving objects in the scene.

Once the histogram representing a moving object is constructed manually or automatically from an initial image frame, n, in which the object appears for the first time, the location of this object in the next frame (n+1 frame) is estimated by determining the similarity of the histogram H_n of the object with a number of histograms extracted from the subsequent n+1 frame. In [30] an

initial histogram iterate is defined at the same location in the next frame. However, it may be also wise to estimate the initial histogram h_1 inside the nearest blob region R obtained by background subtraction and compare it with using the Bhattacharya measure. If the distance is larger than a predetermined threshold then this means that the blob obtained in the next frame is not related to the object O and the histogram is initialised in the n+1 image frame in the location of the object O in the n-th frame. The distance between the two histograms can be compared using other distance measures as well.

The histogram comparisons are repeated in an iterative manner until a satisfactory match is established. The article by Comaniciu et al. provides a procedure for determining an offset vector called the mean-shift vector. It is shown that iterates converge to a local extrema, which is assumed as the center of mass of the object O in the n+1 video-frame. Instead of the kernel density estimation process, it is experimentally observed that histograms can be smoothed using a low-pass FIR filter.

By continuing this process the trajectory of the object O is determined in the sequence of images forming the video. If there are several objects in the video the tracking process is carried out for each object, separately to estimate their trajectories. Relative distances between known background objects and moving objects are continuously calculated from their trajectories.

It should be pointed out that a specific tracking algorithm by itself might not perform well in all possible cases that can be encountered. A feedback mechanism to correct the mistakes has to be incorporated to the algorithm as shown in Figure 2. For example, in the above discussion background subtraction based motion detection and the mean-shift tracking are combined. Iterative mean-shift process is not only started from the original position of the blob in the previous frame but also from blobs determined by the motion detection method.

Researchers also developed tracking algorithms especially designed for tracking human body parts and human body [32-46].

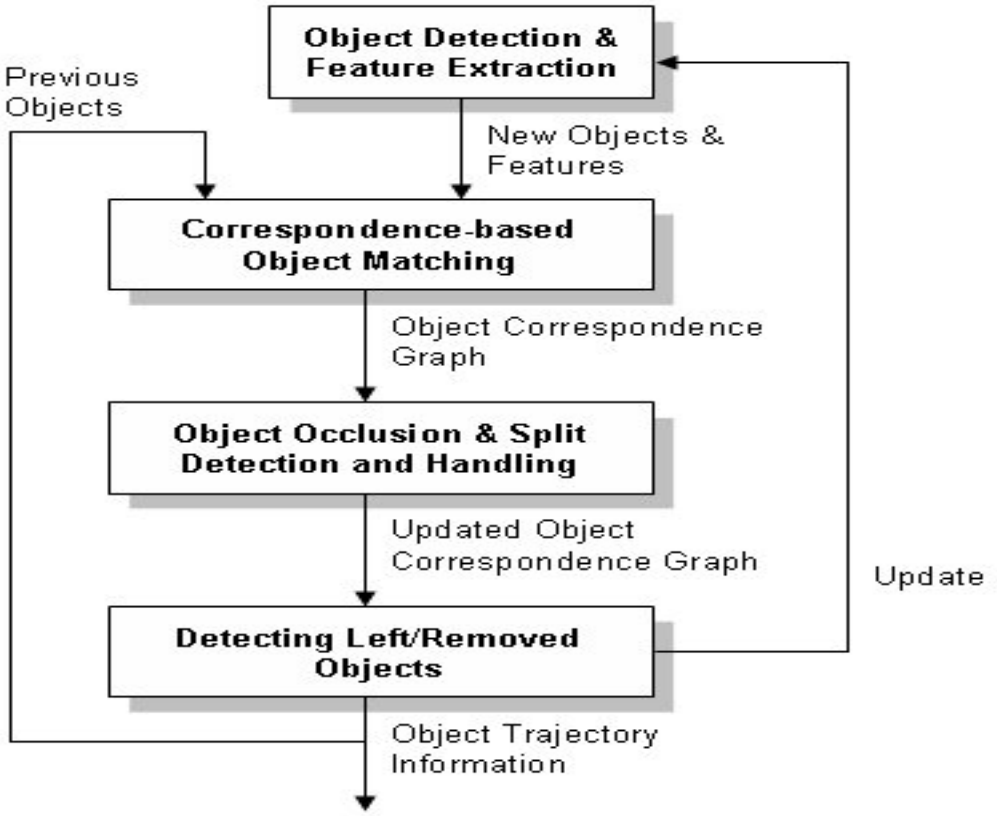

Figure 2: Block diagram of an object tracking algorithm

2.2. Object classification

The next step is to determine what estimated moving blobs are. Usually, the problem is posed as a classification problem [1,47-49]. For example, the video captured by an indoor surveillance camera may contain individual persons, groups of people, and pets and mice in some cases. Video captured by an outdoor surveillance camera include pedestrians, vehicles, birds, clouds, and other animals. To extract semantic information it is necessary to correctly distinguish people from other moving objects and track their motion in the video.

Moving objects are classified according to *shape, colour* and *motion.* Block diagram of a typical object detection scheme is shown in Figure 2. Shape of the moving regions can be characterized in many ways, including two-dimensional (2-D) contour [1], [124],[125], silhouette [50], skeleton [51], and blob box information [1]. In [1] moving object blobs are classified into four classes isolated human beings, vehicles, human groups and clutter using a three-layer neural network classifier. Classification is carried out using a set of features including a mixture of image-based and scene-based object parameters such as moving blob dispersedness, blob area, aspect ratio of the bounding box, and the camera zoom information.

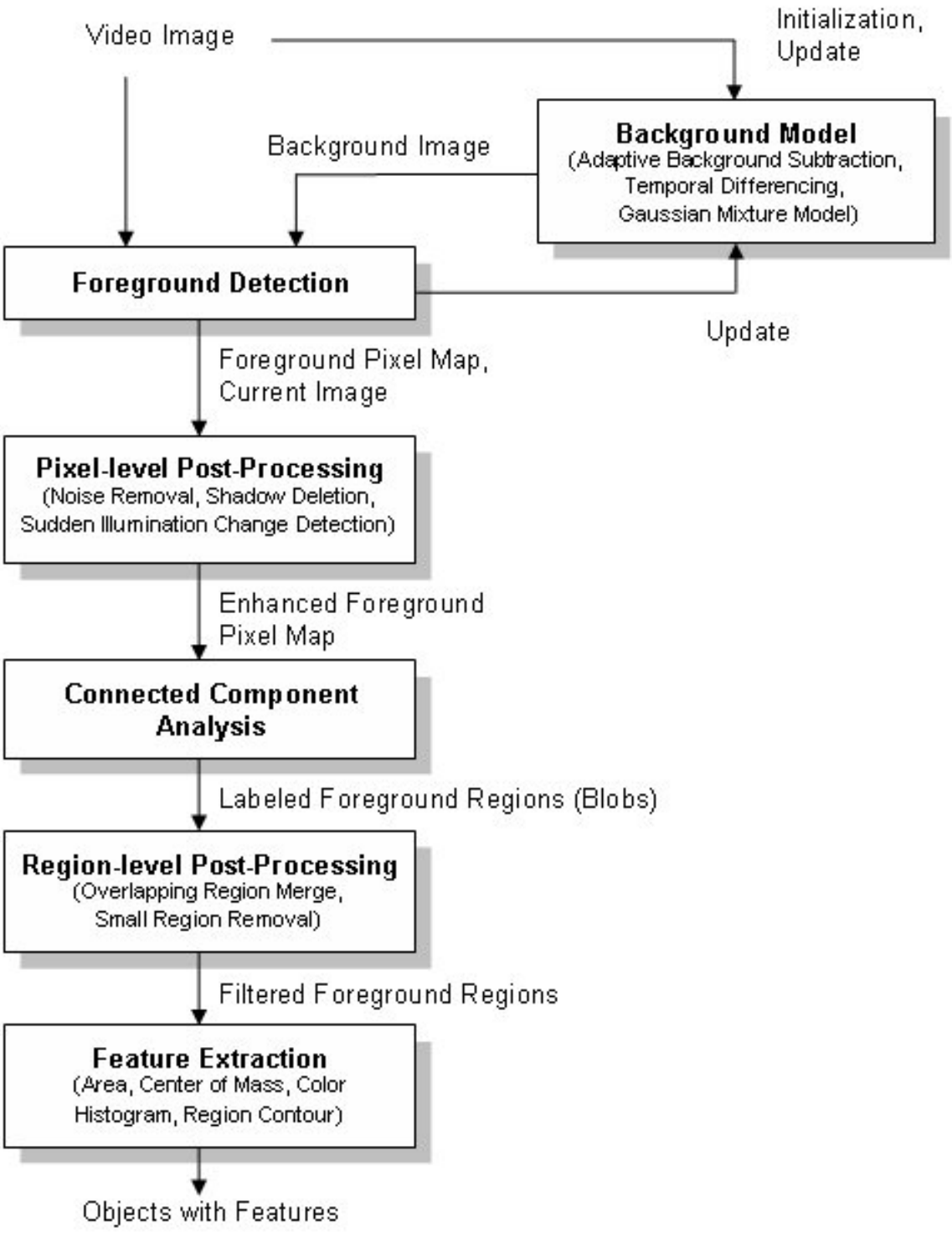

Figure 3: Feature extraction process in video

Moving objects are modeled using stick figures, so-called ribbons [52], 2-D ellipsoids [53] etc. Model parameters are estimated from blobs and various classification engines are used to recognize the moving blob.

Existence of human face and hand colors in the color histogram of the moving blob robustifies the decision process. Human skin color detection can be carried out in various color spaces including the red, green, and blue (RGB) color space, luminance and color difference or chrominance (YUV), and hue, saturation (HSV) color spaces. It is experimentally found that the simple YUV color space is as good as any other color representation method for human skin color detection.

The use of temporal information for object classification has been studied by many researchers as well [47,54-56]. In fact, the classification scheme described in [1] is repeated over several frames to reach a final decision with the help of a tracking algorithm. More sophisticated methods take advantage of the periodic behaviour exhibited by non-rigid articulated human motion [47,54-56]. One can also take advantage of the prior information about the scene monitored by the surveillance camera. For example, pedestrians move much slower than vehicles in a traffic monitoring case [56].

Data fusion concepts can be used to combine shape, colour and motion information for moving object classification. There are good recognition results in controlled environments however in natural settings it is still an unsolved recognition problem [57-59]. The problem is especially difficult if there is occlusion e.g., a person is sitting behind a table etc. Moving object tracking may provide partial solution to the occlusion problem.

It is probably impossible to solve the problem completely but a method consisting of fusion of several algorithms may provide a solution.

2.3. Model-based tracking and Classification

Model based human body modelling and related human body tracking methods include

- stick figure representation based methods,
- 2-D contour modeling based methods, and
- 3-D volumetric models

which represent the geometric structure of the human body in various mathematical ways.

Stick-figure representation of a human body is basically a combination of line segments linked by joints [60-64]. Such a representation is suitable to model the movements of the torso head and the limbs. Various human movements represented by a sequence of stick movements are modeled by Hidden Markov Models (HMMs) [115] for classification [65]. This approach can be used not only to recognize the existence of human beings in a video but also to classify the human movements. A hidden Markov model representing each typical human pose and movement can be designed and trained a priori using typical training video sequences. During the recognition phase the movement of human body is recognized according to the HMM producing the highest probability.

The weakness of this approach is the overly simplified stick representation of the human body. Classification errors are mainly due to the simplified model. More sophisticated models include not only simple sticks but also 2-D ribbons [66-69] to have better models of articulated motion of human limbs.

2-D contour modelling based approaches include snakes [70] or active contour models [44-46,70-72] and various 1-D functions representing the 2-D contours of the moving blobs.

Active contour based tracking and object boundary modelling have been studied especially in the context of video coding. Adaptive contour models, which are dynamically updated for each video frame, were also developed. Given an image sequence having a moving object, it is only necessary to initialize the snake around a contour of the moving object, which can be extracted using a background subtraction algorithm, or level sets can be used. For each new image, the snake is placed on the image at its position in the previous image. The snake then extracts the object's contour in the new image, if the velocity of the object is slow so that there is reasonable amount of overlap in object pixels between the two consecutive image frames. A prediction algorithm, such as a standard Kalman filter can be used to estimate the next position of the snake in a video sequence. This makes the tracker more robust and capable of tracking fast moving objects as well.

Neural networks are also used for classifying tracked moving objects as human being or not. The neural network can be trained using the so-called snaxels of the active contour. Due to the relatively small size of the feature vector consisting of snaxels computational load of classification is low. However, active contour based boundary estimation is an iterative process hence it is relatively computationally costly. The negative aspect of the active contour or in general object boundary based classification algorithms is the occlusion problem. If a human being is partially occluded then such methods lead to incorrect results.

Neural networks, HMMs and other classification engines can be also trained using 1-D functions extracted from 2-D contours [1]. For example, a 1-D function can be constructed by computing the distance from the center of mass of the object to the boundary of the object at various angles as shown in Figure 4. Another 1-D function can be constructed by projecting the moving object onto the horizontal and vertical axes and concatenating the 1-D projections etc.

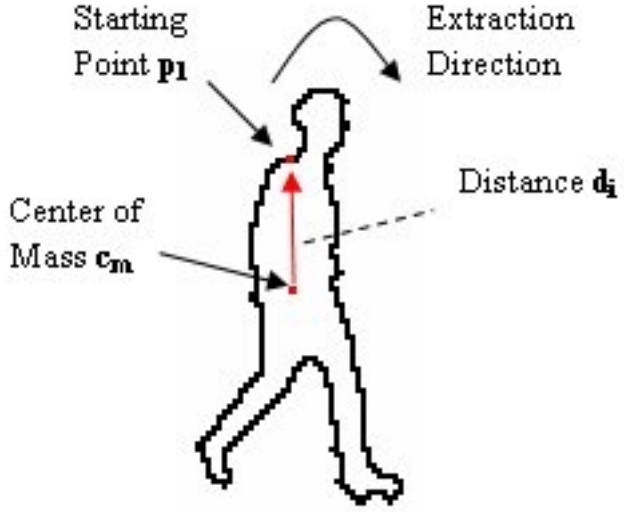

Figure 4-a: A 1-D function can be constructed by computing the distance from the center of mass of the object to the boundary of the object at various angles

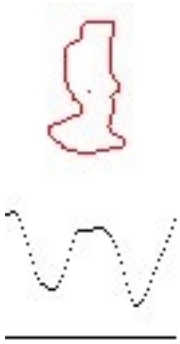

Figure 4-b: Boundary of a walking person and the corresponding 1-D function

3-D *volumetric models* including elliptical cylinders, cones, spheres and other more sophisticated models [73-78] are used to model human body. Research in this field is based on the human modelling methods developed in the field of computer graphics. If a 3-D model of the scene is also available or estimated then 3-D volumetric models provide an accurate description of moving objects and people. In order to estimate 3-D parameters it would be desirable to have stereo or multiple cameras monitoring the scene. Obviously, 3-D models require more parameters to be estimated and updated during the tracking process and this leads to computationally expensive methods compared to other human models during the matching process. In general, the number of parameters to represent the links of the human body is kept at minimum to reduce computational complexity. In the case of deformable links, the deformation parameters are reduced by restricting the deformations to linear deformations [75, 79].

There is a great deal of research to find suitable models for object recognition purposes, especially for human recognition and tracking in images and videos. Biederman [80] proposed parameterized part models that capture the structure of the objects in terms of a few parameters. He suggests that recovery of arrangements of a few major primitive components leads to faster object recognition. These models are computationally efficient since using a small number of parameters simplifies the problem to search in a small database of stored models for matching.

Terzopoulos, Witkin, and Kass [81] proposed a deformable cylinder for recovering 3D shape and motion. Their work exploits computational physics in the recognition process. However, the model with a small number of parameters is not enough to provide an abstract representation of object shape.

Gupta and Bacjy [82] use superquadric ellipsoid with parameterized deformations for part description and segmentation. Pentland [83,84] proposed a physics-based method for fitting deformable part models based on superquadric ellipsoids. They used modal analysis to reduce the computational cost of the fitting process. They eliminated high frequency components that do not have a significant effect. Metaxas and Terzopoulos combined the best aspects of Terzopoulos, Witkin and Kass models and Pentland's models to their deformable primitives to recognize articulated models composed of deformable parts. They coupled the rigid body deformation dynamics as in [81]. Since they use global deformations defined by nonlinear parametric equations, their models are more general.

3. Human Face and Body Detection in Images and Video

Human face and body detection can be carried out not only in video but in still images as well. In this case there is no motion information limiting the search space. This problem is important in applications such as face recognition for security application, emotion estimation, face image database management, and in some safety applications. The problem of human face and body detection also suffers from the illumination changes, cluttered background illumination, partial occlusion, scale and the pose variations. A human detection scheme should have a way to solve variations in scale, orientation, pose and illumination. In current systems, some of these variations are assumed to be fixed to reduce the complexity of the problem. In spite of the recent research human face and body detection the problem remains as an unsolved problem.

Although it suffers from illumination changes skin color detection is an important starting point or a clue in many human detection methods in still images [85,86]. Various color spaces are used for skin color modeling. But recent studies show that widely used Y,U and V (luminance and chrominance or color difference) color space is also as good as other sophisticated color representation methods.

The most effective way to cope with variations in scale is to use a multiresolution wavelet or pyramidal image analysis strategy. By representing the input image at various resolutions, it is possible to have a human face or body in different sizes so that a matched filter type object detection algorithm can detect it by running it over all multiresolution sub-images [87-91]. Or object features can be trained in a fixed scale and it is hoped that one of the subimages created by the multiresolution decomposition have a similar size object. Another related recent approach is to use Haar wavelet filters at various scales as matched filters over the image and extract some feature parameters to feed to a pattern classification engine. This approach is also robust to variations in face orientation.

Applying several filters or templates at various resolutions to an image is a dual version of obtaining multiresolution sub-images from a given image and processing them with a fixed filter. These two approaches are equivalent from a systems engineering point of view. In fact a similar approach can be used to cope with variations in face orientation as well. Template matched filters of various scale and orientation can be used [92],[93]. Another approach is to detect facial features and apply some affine geometric transformations to convert them into a fixed space [94],[94]. The problem with this approach is that the feature extraction stage may not produce good results or it may detect too many candidate regions.

Multiple still images can be also used to solve variations in pose and scale. But this requires the use of multiple cameras. Unfortunately, only one still image is available in many image databases.

It is probably impossible to have a single algorithm perfectly working in all possible cases. However results of several algorithms can be fused to achieve a robust human face and body part detection method from still images.

4. Multimodal Data Analysis for Semantic Information Extraction From Multimedia Archives

Early work on image retrieval systems is based on text input, in which the images are annotated by text and retrieval is performed on the associated text [16]. However, manual annotation is labor-intensive and becomes impractical when the collection is large. Another problem is the subjective nature of keyword, the same image/video may be annotated differently with different annotators. Therefore, it is desirable to have a video database management system, including a Web-based graphical query interface that can handle queries involving spatio-temporal and semantic properties of video data [17].

In content-based retrieval systems the aim is to browse and retrieve according to their visual features, such as color, texture or shape. Many systems are introduced for searching image and video archives using their visual contents, see e.g. [96] for a survey of image and video indexing and retrieval methods. Apart from [48], [97], [98], that try to identify faces and cars, people, and pedestrians, image matching in database systems is not usually directed towards object semantics but based on low-level features, like color and/or texture. User studies show that the users are interested semantics based image retrieval [99,100].

Due to the limitations of only text based and content only based systems, the use of multimodal data, if available, is a best strategy to get robust retrieval results. It is shown that performance of multimedia analysis and understanding systems can be greatly enhanced by combining different modalities such as image, video, audio and text [101- 107].

The goal is to develop useful multimedia systems that employ annotation and search technologies based on semantic concepts that are natural to the user. However, extracting the semantics from the images is a very difficult and long-standing problem. Learning the semantics linked to image features requires carefully labeled data, which is very difficult to require in large quantities. Instead, in recent studies it is shown that, such relationships can be learned from multimodal data sets that provide a loosely labeled data. Probabilistic models are proposed to learn the joint statistics of visual features and textual features [108-113]. It is shown that learning the associations between different modalities provides better annotation and retrieval performance and captures the hidden semantics. Examples include recognition of objects and faces in large databases [109, 112,114].

5. Semantic Information Extraction in Video

Once the so-called low-level image processing tasks such as moving object detection, tracking etc. are finished then the next step is semantic information extraction from video sequence. Understanding of human behaviour in video is one of the key goals of this NoE.

The current approach to this problem is to analyze the video to extract some feature vectors and to classify this time-varying feature data. During the recognition phase, extracted unknown test feature vector set is compared to a group of labeled reference feature vector sets representing typical human actions. The problem is basically equivalent to a selection of a phrase or a sentence from a finite set of sentences or phrases. Therefore, the training phase consists of obtaining reference action sequence of vectors corresponding to each possible “sentence” from training videos.

Feature extraction process is very important to achieve good results. It is impossible to train and perform classification using the currently available classification engines. The size of the feature vector should be as small as possible to have computational efficiency. At the same time it should represent each action very accurately. For example, the center of mass of the tracked object in image frame of the video can be used as a feature vector in a security application in which moving persons entering or leaving a building. In this simple problem, the “vocabulary” consists of a person leaving the building and a person entering a building and the center of mass information consisting of the horizontal and vertical coordinates in an image of the video may be good feature vector for this problem. On the other hand, detection of a fallen person in a room may require some other additional parameters in the feature set. For example, so-called snaxels of an active snake contour of the human body can be added to the feature vector to distinguish a fallen person from a person sitting on a couch. Compactness of the contour boundary, the speed of the center of the mass etc can be also used to distinguish the normal action of sitting on a

couch and the abnormal action of a fallen person. As a rule of tomb, model parameters are selected as entries of the feature vector in a model based tracking approach. Since a video consists of sequence of images a sequence of feature vectors are obtained to characterise the motion of a person(s).

General pattern classification methods used in human behaviour understanding in video include

- Dynamic Programming,
- Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) s, and
- Neural Networks (NN) and Support Vector Machines (SVM),
- Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

The above classification methods are originally used in speech and speaker recognition systems.

Dynamic programming is a well-known optimization method used in many practical problems. It was also used in finite vocabulary speech recognition in 1980's and the method is called as Dynamic Time Warping (DTW) in the field of speech processing. Dynamic Programming is an NP-complete problem and as a result when the size of the problem increases there is no way to reduce the computational cost. Therefore, it is not used in current practical speech recognition systems due to its relatively high computational cost compared to HMM based systems.

Dynamic Programming (DP) was used in matching of human movement patterns [115,116]. It can successfully used to establish matching between the test and the corresponding reference vector as long as time ordering constraints hold.

It is our belief that DP is not fully exploited in computer vision problems including contour boundary based object classification. Computational cost can be reduced by using some heuristic optimization solutions to DP problems.

Hidden Markov Models (HMM) s are stochastic finite state machines, see e.g., [117]. In the context of human motion analysis, a finite state Markov model is assigned for each possible scenario and its parameters are trained with feature vectors of this typical human action. The training process is an off-line iterative algorithm called Baum-Welch algorithm [117]. During the classification or recognition phase, the test feature vector set is applied to all of the Markov models and output probabilities are computed. The Markov model producing the highest probability is determined and the corresponding human action scenario is selected as the result. In finite vocabulary speech recognition DP and HMM based methods produce comparable results. Since HMM based approach is computationally more efficient than dynamic programming it is used in practical speech recognition systems. HMM's are not only used in human action classification [118-122] but also in other image and video analysis algorithms in automatic video indexing and segmentation [123] and in stochastic natural language processing.

Neural Networks (NN), Time-Delay Neural Networks (TDNN) [131], Cellular Neural Networks (CNN) [128] and other related schemes including the Support Vector Machines (SVM) are also widely used in pattern analysis and classification problems [1,2,8,13,128]. Applications include classification of human faces according to emotional expressions [130,122] and erratic human motion patterns [1,2,127].

Principle Component Analysis (PCA) [22,132], and Linear Discriminant Analysis (LDA) are also used in pattern classification. Well-known eigenface method developed for human face recognition is based on PCA.

6. Future Research Directions

Human motion analysis and action recognition are in early stages of research in spite of the large amount of work has been carried out in many distinguished research centers. In general the following problems deserve further attention:

- Shot segmentation problem in video,
- occlusion in video,
- scene modelling and 3-D object detection and tracking,
- feature extraction in of multiple camera systems,
- human action recognition, and automatic or semiautomatic learning techniques,
- automatic and semi-automatic image and video annotation and search methods based on natural semantic concepts,
- human face and body detection in still images and
- Performance evaluation and test databases.

Segmenting the video into semantically meaningful portions in real-time is still an active non-trivial research problem because of changes in illumination conditions, and shadows in natural scenes. This step is also very important to start further motion detection and tracking process.

Occlusions in natural scenes produce problems in many human tracking and detection algorithms. Therefore robust algorithms capable of handling occlusions are needed. This is especially a major problem in human-computer interaction systems based on gesture recognition. If the user's hand is partially blocked somehow during the interaction process then most hand gesture recognition systems fail. Algorithms predicting human pose and position and interpolating missing body parts or portions will be essential for robust systems.

Automatic or semi-automatic scene modelling and understanding using multiple cameras and determining the position of moving objects and human beings in a 3-D model of the scene are unsolved problems. Finding the exact 3-D position of a human being in a 3-D scene is very important in many applications including security, safety and human-computer interaction systems.

Early attempts to classification of human actions in video are based on the techniques based on classification systems developed for speech recognition. Such systems can classify human behaviour in very controlled environments and clearly segmented videos and they can assign the action into only one of the labeled cases. Automatic or semiautomatic learning techniques should be also developed for human behaviour analysis. In addition, data mining algorithms to find interesting human action patterns in an arbitrary video or in a video database and automatic video retrieval systems are still at their infancy.

Current face detection algorithms do not work in a perfect manner in all possible cases, in which human faces may vary in scale, orientation, illumination and pose. It seems that human face feature-based approaches have the best potential for handling variations in scale, orientation, illumination and pose.

Apart from recently established human face databases there are no standard multimedia databases for performance evaluation as in speech and speaker recognition. Such databases are necessary to determine the performance of a given algorithm. In this NoE a specific work package is allocated for this purpose.

Further Information:

Analysis of multimedia data to detect and track the people, and interpret human behavior is an important research area in computer vision. In recent years, human detection and motion analysis have been featured in a number of leading international journals including IJCV (International Journal of Computer Vision), CVIU (Computer Vision and Image Understanding), IEEE PAMI (IEEE Transactions on Pattern Recognition and Machine Intelligence) and IVC (Image and Vision Computing), IEEE Transactions on Image Processing, IEEE Transactions on Multimedia as well as prestigious international conferences and workshops such as European Signal Processing Conference (EURASIP), ICCV (International Conference on Computer Vision), CVPR (IEEE International Conference on Computer Vision and Pattern Recognition), ECCV (European Conference on Computer Vision), WACV (Workshop on Applications of Computer Vision) and IWVS (IEEE International Workshop on Visual Surveillance), International Conference of Pattern Recognition (ICPR), IEEE International Conference on Multimedia and Expo, (ICME). IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on Image Processing (ICIP).

References

1. R.T. Collins, A. Lipton, T. Kanade et al., “A system for video surveillance and monitoring,” Technical Report, CMU-RI-TR-00-12, , Carnegie Mellon University, 2000.

2- R.T. Collins, A.J. Lipton, T. Kanade, Introduction to the special section on video surveillance, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22 (8) (2000) 745-746.

3- R.T. Collins, A.J. Lipton, T. Kanade, "A System for Video Surveillance and Monitoring," in *Proc. American Nuclear Society (ANS) Eighth International Topical Meeting on Robotics and Remote Systems*, Pittsburgh, PA, April 25-29, 1999.

4- I. Haritaoglu, D. Harwood, L.S. Davis, W^4: real-time surveillance of people and their activities, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22 (8) (2000) 809-830.

5-Liang Wang, Weiming Hu, Tieniu Tan, Recent Developments in Human Motion Analysis, *Pattern Recognition*, Vol. 36, No. 3, pp.585-601, 2003.

6. D.M. Gavrila, The visual analysis of human movement: a survey, Computer Vision and Image Understanding, 73 (1) (1999) 82-98.

7- S. Maybank, T. Tan, Introduction to special section on visual surveillance, International Journal of Computer Vision, 37 (2) (2000) 173-173.

8- B.A. Boghossian, S.A. Velastin, Image processing system for pedestrian monitoring using neural classification of normal motion patterns, Measurement and Control, 32 (9) (1999) 261-264.

9- J.J. Little, J.E. Boyd, Recognizing people by their gait: the shape of motion, Videre: Journal of Computer Vision Research, The MIT Press, 1 (2), 1998.

10- J.D. Shutler, M.S. Nixon, C.J. Harris, Statistical gait recognition via velocity moments. Proc. of IEE Colloquium on Visual Biometrics. 2000, pp. 10/1-10/5.

11- P.S. Huang, C.J. Harris, M.S. Nixon, Human gait recognition in canonical space using temporal templates. Proc. of IEE Vis. Image Signal Process. 146 (2) (1999) 93-100.

12-H.M. Lakany, G.M. Haycs, M. Hazlewood, S.J. Hillman, Human walking: tracking and analysis. Proc. of IEE Colloquium on Motion Analysis and Tracking. 1999, pp. 5/1-5/14.

13- M. Köhle, D. Merkl, J. Kastner, Clinical gait analysis by neural networks: issues and experiences. Proc. of IEEE Symp. on Computer-Based Medical Systems. 1997, pp. 138-143.

14- D. Meyer, J. Denzler and H. Niemann, Model based extraction of articulated objects in image sequences for gait analysis. Proc. of IEEE Intl. Conf. on Image Processing. 1997, pp. 78-81.

15- W. Freeman et al., Computer vision for computer games. Proc. of Intl. Conf. on Automatic Face and Gesture Recognition. 1996, pp. 100-105.

16- S. Chang and A. Hsu. Image information systems: Where do we go from here? IEEE Trans. on Knowledge and Data Enginnering, 4(5):431-442, October 1992.

17- M.E. Dönderler, E. _aykol, Ö. Ulusoy, U. Güdükbay, BilVideo: A Video Database Management System, IEEE Multimedia, Vol. 10, No. 1, pp. 66-70, January/March 2003.

18- Yi Li, Songde Ma, Hanqing Lu, Human posture recognition using multi-scale morphological method and Kalman motion estimation. Proc. of IEEE Intl. Conf. on Pattern Recognition. 1998, pp. 175-177.

19- J. Segen, S. Kumar, Shadow gestures: 3D hand pose estimation using a single camera. Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition. 1999, pp. 479-485.

20- M-H. Yang, N. Ahuja, Recognizing hand gesture using motion trajectories. Proc. of IEEE CS Conference on Computer Vision and Pattern Recognition. 1999, pp. 468-472.
21- Y. Cui, J.J. Weng, Hand segmentation using learning-based prediction and verification for hand sign recognition. Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition. 1997, pp. 88-93.
22- M. Turk, Visual interaction with lifelike characters. Proc. of IEEE Intl. Conf. on Automatic Face and Gesture Recognition, Killington, 1996, pp. 368-373.
23- C. Stauffer, W. Grimson, Adaptive background mixture models for real-time tracking. Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition, Vol. 2, 1999, pp. 246-252.
24-Y.H. Yang, M.D. Levine, The background primal sketch: an approach for tracking moving objects, Machine Vision and applications, 5 (1992) 17-34.
25- A. Verri, S. Uras, E. DeMicheli, Motion Segmentation from optical flow. Proc. of the 5th Alvey Vision Conference. 1989, pp. 209-214.
26- J. Barron, D. Fleet, S. Beauchemin, Performance of optical flow techniques, International Journal of Computer Vision, 12 (1) (1994) 42-77.
27- H.A. Rowley, J.M. Rehg, Analyzing articulated motion using expectation-maximization. Proc. of Intl. Conf. on Pattern recognition. 1997, pp. 935-941.
28- Branko Ristic, Sanjeev Arulampalam, Neil Gordon, Beyond the Kalman Filter: Particle Filters for Tracking Applications, Artech House Radar Library, 2004
29- M. Isard, A. Blake, Condensation—conditional density propagation for visual tracking, International Journal of Computer Vision, 29 (1) (1998) 5-28.
30- Comaniciu, D., Ramesh, V. and Meer, P., "Real-Time Tracking of Non-Rigid Objects using Mean Shift," *IEEE Computer Vision and Pattern Recognition*, Vol II, 2000, pp.142-149.
31- V. Pavlovi_, J.M. Rehg, T-J. Cham, K.P. Murphy, A dynamic Bayesian network approach to figure tracking using learned dynamic models. Proc. of Intl. Conf. on Computer Vision. 1999, pp. 94-101.
32- J. Yang, A. Waibel, A real-time face tracker. Proc. of IEEE CS Workshop on Applications of Computer Vision. Sarasota, FL, 1996, pp. 142-147
33- C.R. Wren, A. Azarbayejani, T. Darrell, A. P. Pentland, Pfinder: real-time tracking of the human body, IEEE Trans. on Pattern Analysis and Machine Intelligence, 19 (7) (1997) 780-785. 996, pp. 142-147.
34- M-H. Yang, N. Ahuja, Recognizing hand gesture using motion trajectories. Proc. of IEEE CS Conference on Computer Vision and Pattern Recognition. 1999, pp. 468-472.
35- J. Rehg, T. Kanade, Visual tracking of high DOF articulated structures: an application to human hand tracking. Proc. of European Conference on Computer Vision. 1994, pp. 35-46.
36- D. Meyer et al., Gait classification with HMMs for Trajectories of body parts extracted by mixture densities. British Machine Vision Conference. 1998, pp. 459-468.
37- P. Fieguth, D. Terzopoulos, Color-based tracking of heads and other mobile objects at video frame rate. Proc. of IEEE CS Conf. on Computer vision and Pattern Recognition. 1997, pp. 21-27.
38- D-S. Jang, H-I. Choi, Active models for tracking moving objects, Pattern Recognition, 33 (7) (2000) 1135-1146.
39- S. Wachter, H-H. Nagel, Tracking persons in monocular image sequences, Computer Vision and Image Understanding, 74 (3) (1999) 174-192.
40- J.M. Rehg, T. Kanade, Model-based tracking of self-occluding articulated objects. Proc. of 5th Intl. Conf. on Computer Vision. Cambridge, 1995, pp. 612-617.
41. I.A. Kakadiaris, D. Metaxas, Model-based estimation of 3-D human motion with occlusion based on active multi-viewpoint selection. Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. San Francisco, 1996, pp. 81-87.
42- N. Goddard, Incremental model-based discrimination of articulated movement from motion features. Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects. Austin, 1994, pp. 89-94.
43- J. Badenas, J. Sanchiz, F. Pla, Motion-based segmentation and region tracking in image sequences, Pattern Recognition, 34 (2001) 661-670.
44- A. Baumberg, D. Hogg, An efficient method for contour tracking using active shape models. Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects. Austin, 1994, pp. 194-199.
45- M. Isard, A. Blake, Contour tracking by stochastic propagation of conditional density. Proc. of European Conference on Computer Vision. 1996, pp. 343-356.
46- N. Paragios, R. Deriche, Geodesic active contours and level sets for the detection and tracking of moving objects, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22 (3) (2000) 266-280.
47- R. Cutler, L.S. Davis, Robust real-time periodic motion detection, analysis, and applications, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22 (8) (2000) 781-796.
48- M. Oren, C. Papageorgiou, P. Sinha, and E. Osuna. Pedestrian detection using wavelet templates, Proc. of IEEE CS Conf. Computer vision and Pattern Recognition. 1997, pp. 193-199.
49- C. Stauffer, Automatic hierachical classification using time-base co-occurrences. Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition. 1999, pp. 333-339.
50- S.A. Niyogi, E.H. Adelson, Analyzing and recognizing walking figures in XYT. Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. 1994, pp. 469-474.
51- H. Fujiyoshi, A.J. Lipton, Real-time human motion analysis by image skeletonization. Proc. of IEEE Workshop on

Applications of Computer Vision. 1998, pp. 15-21.
52- I-C. Chang, C-L. Huang, Ribbon-based motion analysis of human body movements. Proc. of Intl. Conf. on Pattern Recognition. Vienna, 1996, pp. 436-440.
53- A. Geurtz. Model-based Shape Estimation. PhD thesis, Department of Electrical Engineering, Polytechnic Institute of Lausanne, 1993.
54-A.J. Lipton, Local application of optic flow to analyse rigid versus non-rigid motion. In the website http://www.eecs.lehigh.edu/FRAME/Lipton/iccvframe.html.
55- A. Selinger, L. Wixson, Classifying moving objects as rigid or non-rigid without correspondences. Proc. of DAPRA Image Understanding Workshop, Vol. 1, 1998, pp. 341-358
56-169. A. Shio and J. Sklansky. Segmentation of people in motion. In IEEE Workshop on Visual Motion, pages 325-332, 1991.
57- A. Iketani et al., Detecting persons on changing background. Proc. of International Conference on Pattern Recognition, Volume 1: 74-76, 1998.
58- C. Cedras, M. Shah, Motion-based recognition: a survey, Image and Vision Computing, 13 (2) (1995) 129-155.
59- A.M. Elgammal, L.S. Davis, Probabilistic framework for segmenting people under occlusion. Proc. of International Conference on Computer Vision, 2001.
60- I.A. Karaulova, P.M. Hall, A.D. Marshall, A hierarchical model of dynamics for tracking people with a single video camera. British Machine Vision Conference. 2000, pp. 352-361.
61- Y. Guo, G. Xu, S. Tsuji, Tracking human body motion based on a stick figure model, Visual communication and Image Representation, 1994, 5: 1-9.
62- C.R. Wren, B.P. Clarkson, A. Pentland, Understanding purposeful human motion. Proc. of Intl. Conf. on Automatic Face and Gesture Recognition. France, March 2000.
63- Y. Luo, F.J. Perales, J. Villanueva, An automatic rotoscopy system for human motion based on a biomechanic graphical model, Comput. Graphics 16 (4) (1992) 355-362.
64- C. Yaniz, J. Rocha, F. Perales, 3D region graph for reconstruction of human motion. Proc. of Workshop on Perception of Human Motion at ECCV, 1998.
65- C. Vogler, D. Metaxas, ASL recognition based on a coupling between HMMs and 3D motion analysis. Proc. of International Conference on Computer Vision. 1998, pp. 363-369.
66- M.K. Leung, Y.H. Yang, First sight: a human body outline labeling system, IEEE Trans. on Pattern Analysis and Machine Intelligence, 17 (4) (1995) 359-377.
67- I-C. Chang, C-L. Huang, Ribbon-based motion analysis of human body movements. Proc. of Intl. Conf. on Pattern Recognition. Vienna, 1996, pp. 436-440.
68- S. Ju, M. Black, Y. Yaccob, Cardboard people: a parameterized model of articulated image motion. Proc. of IEEE Intl. Conf. on Automatic Face and gesture Recognition. 1996, pp. 38-44.
69- C. Hu et al., Extraction of parametric human model for posture recognition using generic algorithm. Proc. of the Fourth Intl. Conf. on Automatic Face and Gesture Recognition. France, March 2000.
70- M. Kass, A. Witkin, D. Terzopoulos, Snakes: Active contour models, International Journal of Computer Vision, Vol. 2, No. 3, pp. 321-331, 1988
71- D. Meyer, J. Denzler, H. Niemann, Model based extraction of articulated objects in image sequences. Proc. of the Fourth IEEE International Conference on Image Processing, 1997.
72- Y. Zhong, A.K. Jain, M. P. Dubuisson-Jolly, Object tracking using deformable templates, IEEE Trans. on Pattern Analysis and Machine Intelligence, 22 (5) (2001) 544-549.
73- S. Wachter, H-H. Nagel, Tracking persons in monocular image sequences, Computer Vision and Image Understanding, 74 (3) (1999) 174-192.
74- J.M. Rehg, T. Kanade, Model-based tracking of self-occluding articulated objects. Proc. of 5th Intl. Conf. on Computer Vision. Cambridge, 1995, pp. 612-617.
75- I.A. Kakadiaris, D. Metaxas, Model-based estimation of 3-D human motion with occlusion based on active multi-viewpoint selection. Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. San Francisco, 1996, pp. 81-87.
76- C. Bregler, J. Malik, Tracking people with twists and exponential maps. Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition, 1998.
77- O. Munkelt et al., A model driven 3D image interpretation system applied to person detection in video images. Proc. of International Conference on Pattern Recognition, 1998.
78- Q. Delamarre, O. Faugeras, 3D articulated models and multi-view tracking with silhouettes. Proc. of International Conference on Computer Vision. Greece, September 1999.
79- D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence , 15(6):580-591, 1993.
80- I. Biedermann. Human image understanding. *Computer Vision, Graphics, and Image Processing*, 32:29-73, 1985
81- D. Terzopoulos, Andrew Witkin, and Michael Kass. Constraints on deformable models:
Recovering 3d shape and nonrigid motion. Artificial Intelligence, 36,1, 1988.
82- Anil K. Jain**,**Chitra Dorai, 3D object recognition: Representation and matching, Statistics and Computing, 10 (2): 167-182, April 2000

83- A. Pentland. Automatic extraction of deformable models. International Journal of Computer Vision, 4:107-126, 1990.
84- A. Pentland, B. Horowitz, *Recovery of non-rigid motion and structure*, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (1991) 730 -- 742.
85-Rein-Lien Hsu, Mohamed Abdel-Mottaleb, Anil K. Jain, Face Detection in Color Images, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 707–711, May 2003.
86- E. Saber and A.M. Tekalp, "Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions," *Pattern Recognition Letters*, vol. 19, no. 8, pp. 669–680, 1998.
87-Q. Chen, H. Wu, and M. Yachida. Face detection by fuzzy pattern matching. In *Proc. 5th Int. Conf. On Comp. Vision*, pages 591-596, MIT, Cambridge, MA., 1995.
88- Y. Dai, Y. Nakano, and H. Miyao. Extraction of facial images from a complex background using SGLD matrices. In *Proc. Int. Conf. on Pattern Recognition*, volume A, pages 137-141, Jerusalem, 1994. IEEE Comp. Soc. Press.
89- A. Lanitis, C. J. Taylor, and T.F. Cootes. An automatic face identification system using flexible appearnace models. *Image and Vision Computing*, 13(5):393-401, 1995.
90-K. K. Sung and T. Poggio. Example-based learning for view-based human face detection. Technical Report A.I. Memo 1521, CBLC Paper 112, MIT, Dec. 1994.
91-G. Yang and T. S. Huang. Human face detection in a complex background. *Pattern Recognition*, 27(1):53-63, 1994.
92- P. Perona. Steerable-scalable kernels for edge detection and junction analysis. In G. Sandini, editor, *Proc. 2nd European Conf. on Comp. Vision*, pages 3-18, Italy, 1992. Springer-Verlag.
93- W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. On Patt. Analy. And Machine Intell.*, 13(9):891-906, 1991.
94- T. Leung, M.Burl, and P. Perona. Finding faces in cluttered scenes using labelled randim graph matching. In *Proc. 5th Int. Conf. on Comp. Vision*, pages 637-644, MIT, Cambridge, MA., 1995.
95- K. C. Yow and R.Cipolla. finding initial estimates of human face location. In *Proc. 2nd Asian Conf. on Comp. Vision*, volume 3, pages 514-518, Singapore, 1995.
96- Y. Rui, T. S. Huang, and S. Chang. Image retrieval: Current techniques, promising directions, and open issues. Journal of Visual Communication and Image Representation, 10(4):39, 62, 1999.
97- H. Schneiderman and T.Kanade. A statistical approach to 3D object recognition applied to faces and cars. In IEEE Conference on Computer Vision and Pattern Recognition, page 100, 2000.
98- M.M. Fleck, D.A. Forsyth, and C.Bregler. Finding naked people. In 4th European Conference on Computer vision, volume 2, pages 591-602, 1996.
99- S. Ornager, View a picture. Theoretical image analysis and empirical studies on idexing and retrieval, Swedish Library Research, vol.2, pp.31-41, 1996.
100- M. Markkula, E. Sormunen, End-user searching challenges indexing pictures in the digital newspaper photo archive, Information retrieval, vol.1, pp.259-285, 2000.
101- C.G.M. Snoek and M. Worring, Multimodal Video Indexing: A Review of the State-of-the-art, Multimedia Tools and Applications, 2005 (in press).
102- I. A. Erdem, M. E. Erdem, Volkan Atalay, A. E. Cetin, Vision-based continuous Graffiti-like text entry system, Optical Engineering, 43(03) p. 553-558, March 2004.
103- A. Hauptmann et.al., Informedia at TRECVID 2003: Analyzing and Searching Broadcast News Video Proceedings of (VIDEO) TREC 2003 (Twelfth Text Retrieval Conference), Gaithersburg, MD, November 17-21, 2003
104- C.-Y. Lin et.al., IBM Research TRECVID-2003 Video Retrieval System, Proceedings of (VIDEO) TREC 2003 (Twelfth Text Retrieval Conference), Gaithersburg, MD, November 17-21, 2003
105- O. Maron and A. L. Ratan Multiple-Instance Learning for Natural Scene Classification, International Conference on Machine Learning (ICML-98), 1998.
106- J. Jeon, V. Lavrenko and R. Manmatha, Automatic Image Annotation and Retrieval using cross-media relevance models, ACM SIGIR, 2003.
107- J. Li, J. Z. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 25, no. 10, 14 pp., 2003.
108- Y. Mori, H. Takahashi, R. Oka, Image-to-word transformation based on dividing and vector quantizing imageswith words, MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management October 30th, 1999 Orlando, Florida, in conjonction with ACM Multimedia Conference 1999
109- P. Duygulu, K. Barnard, N.d. Freitas, and D. A. Forsyth, Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, In Seventh European Conference on Computer Vision (ECCV), volume 4, pages 97-112, Copenhagen, Denmark, May 27 - June 2 2002.
110- K. Barnard, P. Duygulu, N. de Freitas, D. A. Forsyth, D. Blei, and M. Jordan, Matching words and pictures, Journal of Machine Learning Research, 3:1107-1135, 2003.
111- P. Duygulu, H. Wactlar, Associating video frames with text, Multimedia Information Retrieval Workshop, in conjuction with the 26th annual ACM SIGIR conference on Information Retrieval, August 1, 2003, Toronto, Canada
112- P. Duygulu, Alex Hauptmann, What's news, what's not? Associating News videos with words, The 3rd International Conference on Image and Video Retrieval (CIVR 2004) Ireland, July 21-23, 2004.

113- J.-Y. Pan, H.J. Yang, C. Faloutsos, P. Duygulu, Automatic Multimedia Crossmodal Correlation Discovery, The Tenth ACMSIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2004) Seattle, WA, August 22-25, 2004.
114- T. Miller, A.C. Berg, J. Edwards, M. Maire, R. White, Y.W. Teh, E. Learned- Miller, D.A. Forsyth, Faces and names in the news, International Conference on Computer Vision and pattern recognition, 2004.
115- A. Bobick, A. Wilson, A state-based technique for the summarization and recognition of gesture. Proc. of Intl. Conf. on Computer Vision. Cambridge, 1995, pp. 382-388.
116- K. Takahashi, S. Seki et al., Recognition of dexterous manipulations from time varying images. Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects. Austin, 1994, pp. 23-28.
117- L. Rabinier, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of IEEE 77 (2) (1989) 257-285.
118- T. Starner, A. Pentland, Real-time American Sign Language recognition from video using hidden Markov models. Proc. of Intl. Symp. on Computer Vision. 1995, pp. 265-270.
119- J. Yamato, J. Ohya, K. Ishii, Recognizing human action in time-sequential images using hidden Markov model. Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. 1992, pp. 379-385.
120- M. Brand, N. Oliver, A. Pentland, Coupled hidden Markov models for complex action recognition. Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition. 1997, pp. 994-999.
121- Ozer, O.F.; Ozun, O.; Tuzel, C.O.; Atalay, V.; Cetin, A.E.;Vision-based single-stroke character recognition for wearable computing, IEEE Intelligent Systems,Volume: 16 , Issue: 3 , Pages:33 – 37, May-June 2001
122- A. M. Bagci, R. Ansari, A. Khokhar, A. E. Cetin**,** Eye Tracking Using Markov Models, Int. Conf. Pattern Recognition (ICPR), August 23-26,2004.
123- J. Nam, A. E. Cetin, A. Tewfik, "Speaker Identification and Analysis for Hierarchical Video Shot .Classification, Proc. of ICASSP, Munich, 1997, vol. 4, pp. 2597-2600.
124- Bala, E.; Cetin, A.E.; Computationally Efficient Wavelet Affine Invariant Functions for Shape Recognition IEEE Trans. on Pattern Analysis and Machine Intelligence,Volume: 26 , Issue: 8 , Pages:1095 - 1099Aug. 2004
125- H.Ramoser, T.Schlögl, C.Beleznai, M.Winter, H.Bischof; "Shape-based Detection of Humans for Video Surveillance Applications"; Proc. IEEE Int. Conf. on Image Processing; Vol. III; pp. 1013-1016; Barcelona, Spain; 2003
126- C.Beleznai, B.Frühstück, H.Bischof; "Human Detection in Groups Using a Fast Mean Shift Procedure"; Proc. Int. Conf. On Image Processing ICIP, 2004.
127- T. Schlögl, B.Wachmann, H.Bischof, W.Kropatsch; "People Counting in Complex Scenarios"; Proc. Workshop of the AAPR/ÖAGM; pp. 159-166; Graz, Austria; 2002
128- T. Szirányi, J. Zerubia, L. Czúni, D. Geldreich, Z. Kato: ' Image Segmentation Using Markov Random Field Model in Fully Parallel Cellular Network Architectures,' Real-Time Imaging, Vol. 6, No. 3, pp. 195-211, 2000
129- A. Licsár, T. Szirányi, "Hand Gesture Recognition in Camera-Projector System," International Workshop on Human-Computer Interaction, Lecture Notes on Computer Science, Vol. LNCS 3058, pp.83-93, 2004
130- M. Rosenblum, Y. Yacoob, L. Davis, Human emotion recognition from motion using a radial basis function network architecture. Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects. Austin, 1994, pp. 43-49.
131- C-T. Lin, H-W. Nein, W-C. Lin, A space-time delay neural network for motion recognition and its application to lipreading. International Journal of Neural Systems, 9 (4) (1999) 311-334. 159.
132- O. Chomat, J.L. Crowley, Recognizing motion using local appearance. International Symposium on Intelligent Robotic Systems, University of Edinburgh, 1998.