INTELLIGENCE IN MODERN COMMUNICATION SYSTEMS

C. Kotropoulos, and I. Pitas

Department of Informatics, Aristotle University of Thessaloniki
Box 451, Thessaloniki 540 06, Greece
Phone: +30-31-998225, Fax: +30-31-996304
Email: {costas,pitas}@zeus.csd.auth.gr

ABSTRACT

Modern communication systems encompass noise suppression algorithms to compensate for recording and transmission errors as well as multimedia data representation modules offering both expressive and discriminating capabilities. Fusion of decisions derived by mono-modal detectors and features extracted separately for each data type should be supported. In addition, modern terminals should be equipped with human- centered interfaces and should authenticate the user on the basis of biometric information. Moreover, the design of copyright protection procedures is needed. In the paper, state-of-the-art techniques are reviewed and applications are described. The potentialities of neural networks to provide the desired intelligence are highlighted as well.

1. INTRODUCTION

The recent advances in hardware, software and digital signal processing allow for the integration of different data streams, such as text, images, graphics, speech, audio, video, animation, handwriting and data files in a single framework, motivating the convergence of traditionally separated technologies, namely, digital signal processing (e.g., discrete-time speech and audio processing), digital image processing, computer vision, computer graphics and telecommunications to a unified discipline that is now called multimedia signal processing [1]. We may think of this new field as the place where signal processing and telecommunications meet computer vision, computer graphics and document processing. The former disciplines have been traditionally explored in electrical engineering curricula whereas the latter ones have grown within computer science societies. The new discipline aims at providing seamless, easy-to-use, high quality, affordable multimedia communications between people and machines anywhere and anytime [2]. Accordingly, multimedia signal processing is the integration and the interaction of the aforementioned different media types that creates challenging research opportunities. The scope of multimedia signal processing goes beyond signal compression and coding. The purpose of this tutorial is to briefly address these research opportunities which constitute the intelligence of modern communication systems.

When we think of advanced telecommunications, we talk about the Internet and cellular phones. Because of the ability to interact with Web-based services via a standard graphical user interface associated with the PC and the availability of personal digital assistants able to access and interact with the network via wired and wireless connections, the meaning of ubiquitous access has become less clear in advanced telecommunications systems [2]. Ubiquitous access to broad-band services seems to require a broad-band access terminal. One potential solution might be a portable PC with wireless broad-band access. Another way of providing ubiquitous access is via voice access to web services, e.g., voice based intelligent agents that access web information servers to retrieve information about the traffic, the weather, etc. Therefore, new smart terminals are needed to facilitate the creation, display, access, indexing, browsing and searching of multimedia content in a convenient and easy-to-use manner. To achieve modern communications systems the goal of ubiquitous access, a number of technological issues must be addressed, including:

1. Efficient data representation techniques in which the emphasis is shifted from bits to the actual content (i.e., the objects) in order to facilitate the organization, storage and retrieval of information as well as the integration of natural content with computer generated one [3]. Having segmented the objects in the multimedia signal, feature extraction and feature selection is performed. Data-embedding and watermarking algorithms that attempt to establish the ownership (i.e., copyright protection) and important descriptive or reference information in a given multimedia signal should be considered as parts of the representation [4, 5].

2. Basic techniques for accessing the multimedia signals by providing tools that match the user to the machine which deal with (i) the extension of the graphical user interface associated with a standard PC to a spoken language interface when mouse and keyboard

or a touch screen are not available [2], (ii) the design of bimodal interfaces that integrate both audio and visual information, e.g., lip reading, speech-driven face animation [6], (iii) the extension of bimodal interfaces to multimodal human-centered ones by fusing information from multiple sources, e.g., hand gestures, facial expressions, eye movements, brain activity [7, 12, 13, 14], (iv) the development of agents that monitor the multimedia sessions and provide assistance in all transactions, (v) the establishment of facial communications between cloned and virtual faces [9], and, (vi) the biometric authentication of users accessing multimedia services [10, 11].

3. Basic compression and coding algorithms for the various media that constitute the multimedia signal [2, 15].

4. Basic noise suppression algorithms, including (i) robust signal detectors for wireless communications to cope with multiple access interference [16], (ii) error concealment techniques to cope with random bit errors in variable length coding that desynchronize the coded information [17], (iii) robust prediction algorithms in the presence of transmission errors [18], (iii) linear and nonlinear adaptive filtering algorithms that suppress the cumulative noise effect due to transmission channel and quantization errors [19, 20, 21], and, (iv) possibly postprocessing algorithms, such as speech-assisted video warping algorithms (e.g., lip synchronization) to interpolate the video signal when the decoder operates at a lower frame rate due to bandwidth constraints [6].

5. Basic techniques for searching in order to find the multimedia sources that provide the desired information via text requests, image matching methods and speech queries.

6. Basic techniques for browsing individual multimedia documents and digital libraries in order to take advantage of human intelligence to find the desired material via text browsing, indexed image browsing and voice browsing.

Among these issues, the tutorial focuses to content representation, access techniques, noise suppression algorithms and content-based searching in multimedia databases. The major contribution of our research team to these issues is described. For the remaining topics, the interested reader may refer to [1] as well at to the other tutorials that will be presented in this conference. The potentialities of neural networks to provide solutions in these problems are briefly reviewed. An abstract data processing model that can address the majority of the problems in the design of communication systems is commented.

2. CONTENT REPRESENTATION

The vast majority of today's natural content is captured using traditional cameras and microphones. The acquisition process fails to capture a significant part of the inherent structure of visual and aural information, such as 3-D geometrical structure unless special hardware equipment is being used. For a long time, the emphasis in representation has been on compression, that is, to describe a signal with as fewer bits as possible. Current needs, such as content-based retrieval and browsing, imply object-oriented design approaches. This is evident in MPEG-4 that is centered around a basic unit of content, the so-called audio-visual object. Moreover, in MPEG-4, an effort is made to integrate natural and synthetic content, enabling synthetic-hybrid coding.

Human face and speech are particular objects that received much attention because of their prominent role in communications. An important technical problem is to design real-time object segmentation tools for both visual and audio content. Face-like regions can be segmented in a scene by exploiting the distinct skin-like color that corresponds to a particular sector of the HSV color space and their elliptical shape [22, 39](Figure 1). When only grey level information is available, face

Figure 1: Fitting an ellipse to a frontal face.

detection can be based on a hierarchical knowledge-based pattern recognition system that uses multiresolution images, known as mosaic images [23, 24] (Figure 2). Both approaches employ a number of assumptions, including the mirror-symmetry of the face, the biometric analogies (i.e., a prescribed ratio of face height over face width), the fact that key facial features, such as eyebrows, eyes, nostrils, etc., are related to image local minima identifiable in the vertical image profile whereas face cheeks and the chin can be modeled by simple curves that can be detected by Hough transform. The initial ellipse model can be refined subsequently by employing snakes [25]. The eye pupils can be accurately detected by minimizing a matching error weighted by a 2-D circularly symmetric kernel and further compensation of head rotation can be based on their accurate positions [26]. The major benefit of the aforementioned approaches is their simplicity and their low computational requirements. However,

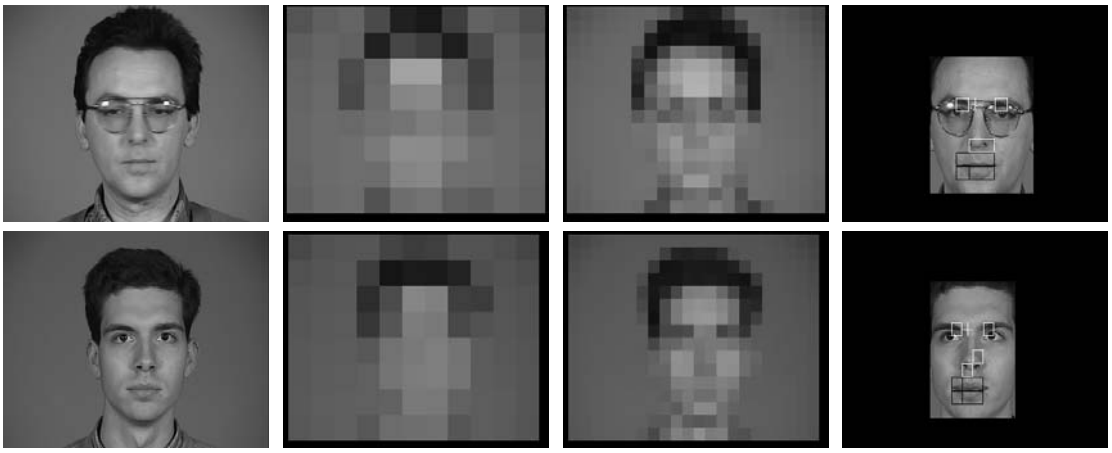

Figure 2: Two frontal face images from M2VTS database. Their quartet and octet images are shown in the second and third column, respectively. The outcome of face detection algorithm is depicted in the last column.

they lack a concrete theoretical background. Example-based approaches based on Support Vector Machines alleviate this drawback [27]. Snakes can be used to track the face contour and block matching techniques can be used to track facial features in a video sequence [22]. In audio processing, two important problems are silence detection and voiced-unvoiced discrimination. Silence audio frames are audio frames of background noise with low energy level with respect to voice segments and can be discarded by thresholding [28]. Alternatively, the average magnitude and zero-crossing rate can be exploited for end point detection, i.e., to determine the beginning and the end of words, phrases or sentences [29, 30]. Voiced-unvoiced discrimination can be based on the fact that unvoiced sounds exhibit significant high frequency content in contrast to voiced ones. It is seen that the current state-of-the-art techniques cannot recognize high-level objects, but low-level "salient" image regions or audio segments that can be used to partially characterize the content (Figure 3).

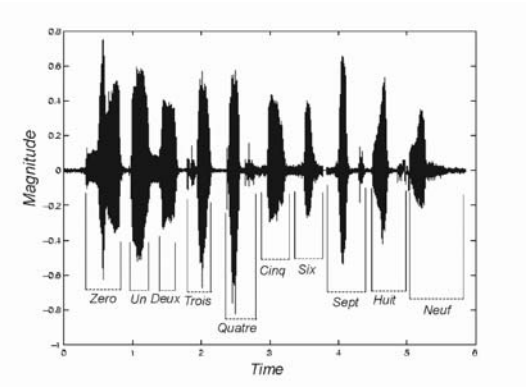

Figure 3: Segmentation of a speech signal.

The next step is to describe the object properties by extracting a feature vector and by selecting the most expressive or discriminating features according to the task that is undertaken. The simplest case is to avoid the feature extraction by using raw information (i.e., image pixel values or speech samples) and to apply subsequently a feature selection procedure. Alternatively, the signal is preprocessed so that additional information (i.e., a parametric representation) is extracted. For example, the responses of a set of 2-D Gabor filters tuned to different orientations and scales [31] or the output of multiscale morphological dilation-erosion at several scales [32] can be employed to form a local descriptor in face analysis (Figure 4). In speech analysis, pre-emphasized voiced frames are windowed by a Hamming window and subsequently linear prediction analysis is applied to derive the linear prediction coefficients (LPC) [33]. Linear prediction coefficients are converted to cepstral coefficients (LPCC) or to mel-frequency cepstral coefficients (MFCC). Either the raw or the parametric information can be described more efficiently, if the feature vectors are projected onto an appropriate subspace. In image compression (e.g., JPEG, MPEG), the Discrete Cosine Transform basis functions are used due to their information packing ability, that is to compact the signal energy in the low-frequency coefficients [34]. In the general case, when the objective is to approximate the signal, the Karhunen-Loeve transform or principal component analysis (PCA) can yield the basis vectors needed to define such a subspace. Since the feature vectors produced by PCA have shown good performance in image reconstruction and compression tasks are called most expressive features [35]. When the objective is to discriminate among objects (i.e., object recognition, information retrieval, etc.) there is no guarantee that the most expressive features are necessarily good. It is well known that optimality in discrimination among all possible linear combinations of features is achieved by employing linear discriminant analysis (LDA). Category information

Figure 4: Dilated and eroded images with a scaled hemisphere for scales 1–9.

labels are necessary in this case. The feature vectors produced after the LDA projection are called most discriminating features [35]. PCA and LDA are parametric techniques closely related to matching pursuit filters [36].

Copyright protection is an integral part of content representation. The importance of this topic is manifested by the fact that JPEG2000 and MPEG4 will include protection mechanisms for intellectual property rights. Watermarks [4, 5] modify slightly the digital signal to embed non-perceptible encoded copyright information. Two main classes of algorithms have been proposed that either use the original signal during the detection phase or not. Detectors that do not imply the availability of the original signal can detect watermarks in images that have been extensively modified in various ways. However, these detectors cannot be combined with web-crawling and automatic watermark searching in a digital library. Watermark embedding can be done either in the time (spatial) domain or in an appropriate transform domain, e.g., DCT, wavelet, Fourier, etc. The imposed changes usually take into account the properties of the human auditory/visual system (perceptual masking) in order to design watermarks that are imperceptible.

3. ACCESS TECHNIQUES TO MULTIMEDIA SERVICES

In this section, we briefly review the state-of-the-art in access techniques to multimedia services with emphasis to biometric person authentication.

Let us first consider the spoken language interfaces [2]. Speech is the most intuitive and most natural communicating modality for most people. The argument for speech interfaces is further strengthened by the ubiquity of both the telephone and microphones attached to PCs. To design powerful speech interfaces, one should be aware of the strengths and limitations of speech synthesis, speech recognition, spoken language understanding, and user interface technologies. Spoken language interfaces replaced the key-pressed commands of the older interactive voice response systems with single-word voice commands. They depend heavily on text-to-speech systems. A well-known text-to-phoneme system is the Festival Speech System developed at the University of Edinburgh, U.K. Synthetic voice can be produced by the system MBROLA developed at the Faculté Polytechnique de Mons, Belgium. MBROLA is a publicly available speech synthesizer based on diphone concatenation techniques that is used to generate speech from a phonetic transcription. Early 80's speech recognition systems were able to recognize 40 to 50 words pronounced in isolation for a single speaker. They were based on pattern matching using dynamic time warping. Improvements have been made towards three directions: from isolated to continuous speech, from speaker-dependent to speaker-independent and the increase in the vocabulary size. The driving force that allowed for a large improvement on the three directions simultaneously has been the use of Hidden Markov Models (HMMs). They have also increased the robustness of speech recognition systems. HMMs allow to use phone models instead of word models, and to reconstruct words including the various pronunciations reflecting the regional variants of foreign accents for the same language from the phone models. Using context-dependent phones or even word-dependent phones helps taking into account the coarticulation effect while solves the problem of a priori segmentation. A similar statistical approach was also used for language modeling, using bigrams or trigrams whose probabilities are estimated from a large text corpus corresponding to the application task at hand. Although N-gram language models have been very successful in speech recognition, some recognition errors can only be corrected by using a larger context than the immediately adjacent few words. Probabilistic linguistics-based models able to capture longer-range effects have improved speech recognition accuracy. The state-of-

the-art performance of speech recognition systems indicate that a very good performance can be achieved for constrained tasks (e.g., digit strings) but the error rate increases rapidly for unconstrained conversational speech. Current issues investigated worldwide are: speaker adaptation, environmental robustness, task independence, spontaneous speech and real-time operation. It has been found that adaptation to different variabilities, such as microphone channel, background noise, vocal tract length, etc., tends to be additive in terms of increased accuracy.

The interaction of audio and video is one of the most interesting interactions between different media. For multimedia applications that involve person to person conversation (i.e., video telephony, video conferencing) , such interaction is particularly significant, because human speech is bimodal in nature. Parallel to the basic unit of acoustic speech, i.e., the phoneme, in the visual domain, we have the notion of viseme, i.e., the basic unit of mouth movements that constitutes a visibly distinguishable unit of speech. The acoustic and visual components of the speech signal are not purely redundant; they are complementary as well. Research topics along the direction of audio-visual integration include automatic lip reading, speech-driven face animation, lip tracking, joint audio-video coding, and bimodal person authentication [6]. Automated lip reading systems are based on an appropriate combination of audio and visual speech recognizers. Visual speech recognizers are based on HMMs or time delayed neural networks where either a number of features of binary mouth images, such as height, width, perimeter, along with their derivatives or active shape models or the aforementioned geometric parameters combined with the wavelet transform of the mouth images are fed as inputs. Another interesting topic is to produce visual speech from auditory speech, i.e., to generate speech-driven talking heads. Wireframe models are used, such as the CANDIDE model developed at Linköping University, Sweden. To synthesize various facial expressions, the facial action coding system is used to generate the required trajectories of the vertices. Computer graphics techniques, such as texture mapping are employed to add photorealism. HMMs or vector quantization techniques are used to find speech-to-viseme mappings and to drive the talking heads.

Besides automatic speech recognition that has been a topic of research for long time, automatic hand gesture recognition, analysis of facial expressions, head and body movement, eye tracking, force sensing, or electroencephalogram are recently gained interest as potential modalities for human-centered interfaces [7, 12, 13, 14]. Some modalities like speech and lip movements, are more closely tied than others, such as speech and hand gestures. The fusion of such different modalities can be explored at different levels of integration. Depending on the chosen level of integration the actual fusion can then be performed by exploiting distributed detection and data fusion algorithms [8]. Briefly, three distinct levels of integration can be distinguished: data fusion, feature fusion and decision fusion. Data-level fusion is characterized by the highest level of information detail among the three types mentioned. It implies the concatenation of data possessing high level of synchronization before feature extraction. However, it is susceptible to noise, modality failures, etc. Feature fusion is commonly found in the integration of modalities and implies the concatenation of features provided by the different modalities. It retains less detailed information than data fusion, but it is also less sensitive to noise. When feature fusion aims at providing a combined fusion output, Kalman filters are usually exploited. When feature fusion is integrated with the decision maker, multilayer perceptrons or HMMs are employed. The type of fusion most commonly used in both human-centered interface design and biometric person authentication, described subsequently, is decision-level fusion which is more robust to individual modality failure.

However, a truly human-centered design must move beyond usability or user-friendliness and the immediate interactions between person and machine. Interfaces should be activity centered, context-bound and problem-driven. Towards this goal tools for assisting human operators to monitor and control complex environments by filtering data, and for devising intelligent planning and control mechanisms are of the highest priority.

A number of biometrics have been proposed, researched, and evaluated for authentication applications including, voice, face, fingerprints, iris, infrared facial and hand vein thermograms, ear, gait, keystroke dynamics, DNA, signature and acoustic emissions, odor, retinal scan, hand and finger geometry [11]. In this tutorial, we briefly review voice and face biometric modalities. A voice signal available for authentication is typically degraded by the microphone, communications channel, and digitizer characteristics. We have already described audio feature extraction. The matching strategy may typically employ approaches based on HMMs, vector quantization or dynamic time warping. Text-dependent speaker verification authenticates the identity of a subject based on a fixed predetermined phrase (Figure 5). Text-independent speaker authentication is more difficult and verifies the speaker identity independent of the phrase. Language-independent speaker authentication verifies the speaker identity irrespective of the language of the uttered phrase and is even more challenging. Voice capture is unobtrusive and voice print is an acceptable biometric in almost all societies. Over a telephone channel, voice authentication is the only feasible biometric. However, voice is a behavioral biometric and is affected by person's health, stress, emotions, etc. A comprehensive survey

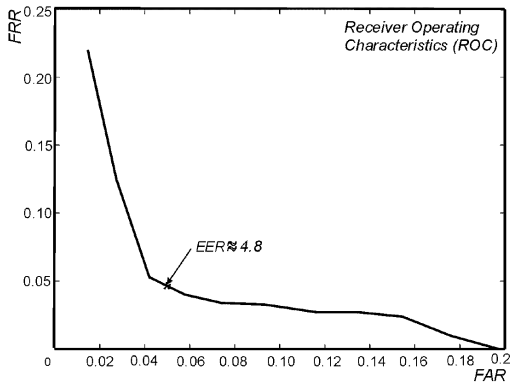

Figure 5: Receiver operating characteristic of a speaker verification system based on vector quantization.

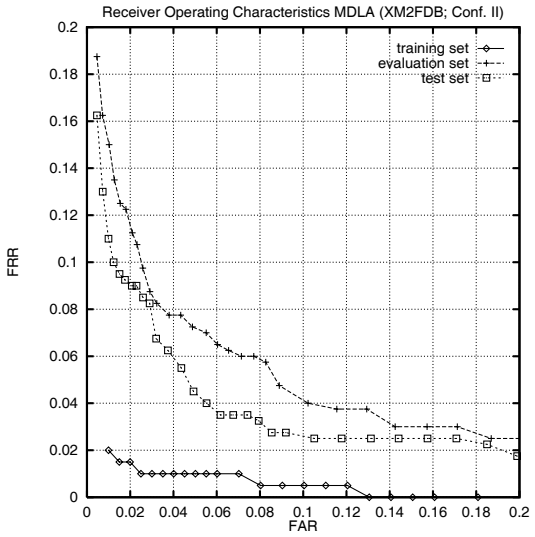

Figure 6: Receiver Operating Characteristic of morphological elastic graph matching on the training, evaluation and test sets on the extended M2VTS database.

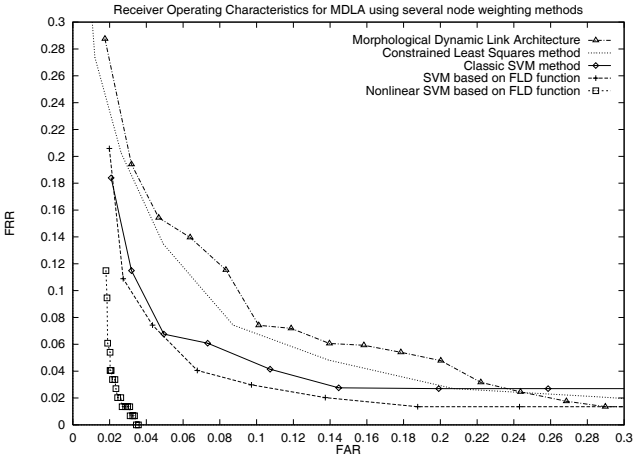

Figure 7: Receiver Operating Characteristics of MDLA with discriminatory analysis based on Support Vector Machines.

of human and machine recognition techniques can be found in [37]. Several approaches in developing face recognition systems exist. A first approach is based on user defined face-specific features, e.g., models of eyebrows, eyes, nose etc., and the derivation of geometrical attributes, such as eyebrow thickness, nose width, a number of radii describing the chin shape, etc. Another approach is based on templates and attempts to match well-defined portions of the face. A third approach employs linear projections of face images (treated as 1-D vectors) using either PCA or LDA [35, 38]. There are techniques stemming from neural network community like the dynamic link architecture (DLA), a general object recognition technique that represents an object by projecting its image onto a rectangular elastic grid where a Gabor wavelet bank response is measured at each node [31]. Recently, a variant of dynamic link architecture based on multiscale dilation-erosion, the so-called morphological dynamic link architecture (MDLA), was proposed and tested for face authentication [32, 39]. It has been tested on multimedia databases collected under controlled conditions ranging from 37 to 295 persons (Figure 6) and small galleries recorded during real-world tests, such as access-control to buildings and cash dispenser services or access-control to tele-services via INTERNET in a typical office environment. The compensation for image variations attributed to variable recording conditions, i.e., changes in illumination, face size differences and varying face position, is also addressed [40]. A novel approach to discriminant analysis that reformulates Fisher's Linear Discriminant ratio to a quadratic optimization problem subject to a set of appropriate inequality constraints is proposed in [41]. The method combines statistical pattern recognition and Support Vector Machines. Both linear and nonlinear Support Vector Machines are constructed to yield the optimal separating hyperplanes and the optimal polynomial decision surfaces, respectively (Figure 7).

Using each single modality, however, has certainly limitations in both security and robustness. Face verification suffers from variation in lighting conditions, scale differences, varying face position, and variable facial expressions. The use of voice only for verification is not reliable, because speech verification is vulnerable to operating conditions (e.g., context, training resources, etc.) at a greater degree than the face modality. Bayesian conciliation theory and AND/OR/median rules that fuse audio- and video-based authentication modalities (i.e., frontal face, profile or fingerprints) are included in [10, 11]. Fuzzy K-means, fuzzy vector quantization and median radial basis function neural networks are also proven efficient decision fusion managers [42].

4. NOISE SUPPRESSION ALGORITHMS

In transmission systems, two basic cases can be distinguished, depending on whether transmission is baseband or if it uses a carrier frequency [19]. The baseband case occurs when the bandwidth of the information message a and that of the channel response overlap significantly. The signal observed at the receiver can be modeled by $x = \mathcal{C}(a) + b$, where b denotes the noise and all involved quantities are real-valued. When the bandwidth of the message a and of the channel $\mathcal{C}$ do not overlap, the message must be frequency-shifted by the use of a carrier frequency f_0 lying in the channel bandwidth. Let us consider a quadrature amplitude modulation (QAM) coherent transmission system. It can easily be shown that the aforementioned model still holds with the exception that x and a are complex-valued signals [19]. Channel equalization, echo cancellation and prediction can be described by a common linear noisy model which relates the two observed sequences of signals $x(n)$ and $a(n)$, according to $a(n) = \mathcal{F}(x(n)) + m(n)$ with $x(n) = a(n-1)$ in the specific case of prediction. We call $x(n)$ and $a(n)$ input and reference. If $\mathcal{F}$ is assumed to be a linear filter, an adaptive estimate of the noise $m(n)$, $\hat{m}(n) = a(n) - \mathcal{H}_{n-1}(x(n))$, can be obtained to control adaptively $\mathcal{H}_{n-1}$ so that $\hat{m}(n)$ becomes as small as possible.

The most widely known adaptive filters are linear ones having the form of either finite impulse response or lattice filters. Such filters may not be suitable for applications where the transmission channel is nonlinear or where the noise is impulsive. A multitude of nonlinear techniques have proven to be successful alternatives to the linear techniques [43]. One of the best known nonlinear filter classes is based on order statistics. L-filters, whose output is defined as a linear combination of the order statistics, is a prominent example of order statistics filters. Let us suppose that the observed signal $x(n)$ can be expressed as a sum of a noise-free signal $a(n)$ plus zero-mean additive white model. Adaptive L-filters estimate the coefficient vector $\mathbf{c}(n)$ of the L-filter output $y(n) = \mathbf{c}^T \mathbf{x}_r(n)$ so that the mean-squared error between the filter output and the noise-free signal is minimized [20, 21]. Additional filtering tasks that depend on order statistics are described subsequently.

The design of detection schemes for spread spectrum multiple access (SSMA) networks in which security restrictions do not permit the distribution of all users' signaling parameters is a formidable task, as the multiple access interference has a non-Gaussian distribution with an exact shape that depends on the received power of each of the active users in the network. Variations in the users' power due to roaming, the advent or departure of users and/or imperfect power control cause the statistics of the multiple access interference to assume vastly different characteristics ranging from near-Gaussian, over multi-modal, to heavy-tailed [16]. While many of the classical robust detectors, including minimax and non-parametric detectors, can offer an acceptable performance over a limited class of possible noise statistics [44], they become inefficient when the uncertainty about the noise distribution is large, as is the case, in spread spectrum networks. A rank order diversity detector based on L-filters to test the polarity of the transmitted signal is proposed in [16]. The L-filter weights are adjusted to the prevailing noise characteristics based on the second order moments of the order statistics of noise. Asymptotically, as the number of samples goes to infinity and under mild assumptions on the noise distribution, the probability of error of the rank order diversity detector is less than or equal to that of the linear detector with the equality holding only for Gaussian noise.

Differential pulse code modulation (DPCM) is an efficient and computationally feasible method for the coding of video signals. However, in the presence of channel errors, the performance of a DPCM coder will deteriorate significantly. Median operators have been proposed for DPCM coders. The median operator can reject outliers and median-based predictors are, in general, less sensitive to channel errors than linear predictors. The performance of weighted median, multistage median and FIR median hybrid predictors in the presence of channel errors is analyzed in [18].

Error control and concealment in video communication is becoming increasingly important because of the growing interest in video delivery over unreliable channels, such as wireless networks and the Internet [17]. Forward error concealment includes methods that add redundancy at the source end and enhance error resilience of the coded bit streams. Error concealment by postprocessing refers to operations at the decoder to recover the damaged areas based on characteristics of image and video signals. Interactive error concealment is based on a dialogue between the source and the destination. We are interested in error concealment by postprocessing which is closely related to filtering. Two types of errors are observed at the receiver. Bitstream errors which are caused by direct loses of part or the entire compressed bitstream of a coded macroblock and result in the loss of slicing information, and propagation errors which corrupt P- and B- frames due to the additional use of motion-compensated information during the decoding. A split-match error concealment has been used as spatial domain interpolation method for I-frames and forward-backward temporal block matching has been exploited for concealment in the I- and B-frames [45] (Figure 8). Vector rational interpolation of erroneously received motion fields of an MPEG-2 coded video-stream for error concealment purposes is proposed in [46]. Rational functions, i.e., the ratio of two polynomials, have been extensively used for image filtering, image restora-

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

Figure 8: (a) First erroneous frame of Flower Garden sequence (PER=0.1). (b) First frame concealed by spatial diffusion. (c) Frame 14 (B-frame) of the same sequence. (d) Frame 14 concealed by forward-backward block matching. (Adapted from [45].)

tion and image interpolation, because they are universal approximators. They are good extrapolators able to be trained using a linear algorithm and requiring lower degree terms than Volterra expansions.

5. CONTENT-BASED INFORMATION RETRIEVAL

Highly sophisticated methods have evolved for textual material based on both direct text matching techniques and associative matching that are related to the semantic interpretation of the requested text terms for the match over the past few decades [2]. For other multimedia modalities, such as speech, audio, image and video, advanced matching techniques are neither well known nor well developed. For example, automatic image indexing based on simple low-level features such as color, texture and shape has been to a large degree successful [3]. Such low-level features however, do not provide complete solutions for most users. Beyond simple queries for specific forms of information, users would like capabilities to extract information at higher levels of abstraction and track the evolution of concepts at higher semantic levels. More specifically, document imaging requires zero-error optical character recognition error for all document types, fonts, character sizes, special symbols and languages. Audio and speech indexing requires perfect speech and audio understanding in all acoustic environments and for all speakers and languages. Image and video indexing requires the solution of the "general" vision problem, that is to locate and identify all objects (and ultimately their relationships) in complex scenes. Among the critical research issues are [3]: multimedia content analysis and feature extraction; efficient indexing techniques and query optimization; integration of multimedia; automatic recognition of semantic content, etc. In this tutorial, we briefly describe a content-based video parsing application [28, 30]. Content-based video parsing involves temporal video segmentation into elementary units and content extraction of those units based on visual and/or audio semantic primitives. Audio-video interaction serves two purposes: (a) to enhance the content findings of one source by exploiting the knowledge offered by other sources, and, (b) to offer a more detailed content description about the same video instances by combining the semantic labels of all data sources using fusion rules. The problem addressed in these works is to estimate the speaker presence likelihood in every face shot. Speaker recognition based only on audio information is prone to recognition errors. These errors can be reduced by utilizing the related visual information, e.g., presence of a talking person in the scene. The presence likelihood of a particular speaker can be estimated by the relative frequency of speech frames assigned to this speaker over the total number of speech frames in a sequence of successive frames that correspond to a single camera start and end session. The speaker that exhibits the higher presence likelihood is the winner. All speech frames in face shot are indexed with the winner's identity. The interaction of audio and visual semantic labels (e.g., speech, silence, speaker identity, face presence, face absence, talking face presence, etc.) can be exploited by the user to issue queries in the system (Figure 9).

6. POTENTIALITIES OF NEURAL NETWORKS

Neural networks (NNs) [47] can offer solutions to many of the problems addressed in this tutorial. They offer unsupervised clustering and supervised learning mechanisms for recognition of objects that are deformed or defined in an incomplete way. They are powerful pattern classifiers. Moreover, temporal neural models specifically designed to deal with temporal signals are appropriate for multimedia processing. NNs have successfully applied to facial expression and emotion categorization, face recognition, lip-reading analysis, multimodal representation and information retrieval. In the following, a brief exposition of NN potentialities is given. The interested reader may refer to [48].

An efficient representation of a vast amount of multimedia data can often be achieved by adaptive data

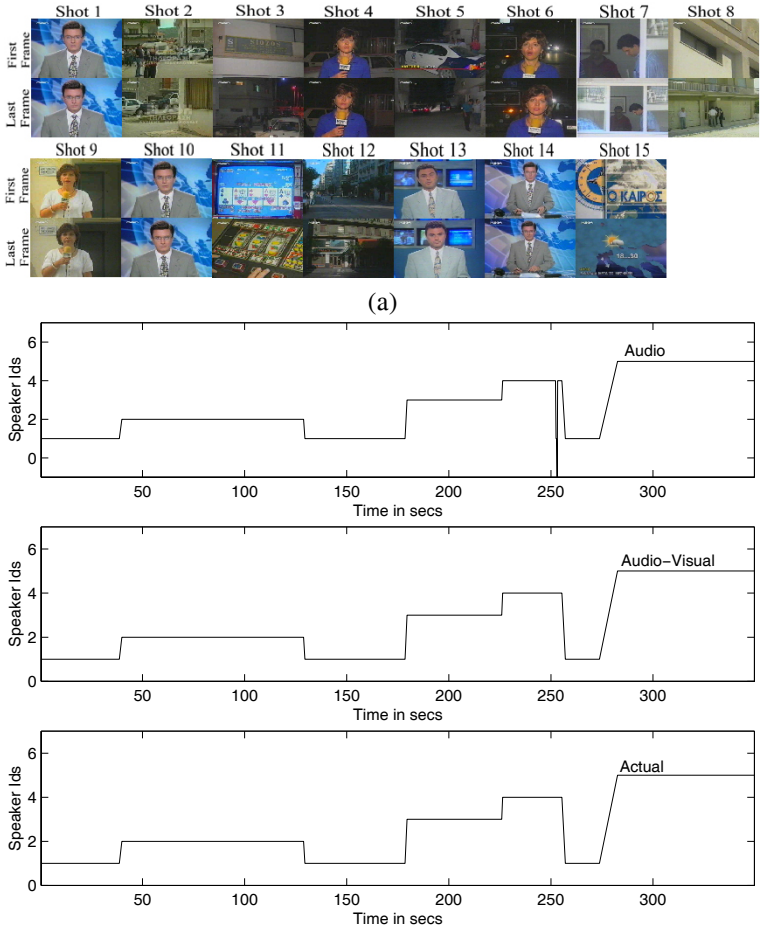

Figure 9: (a) First frames of detected shots. (b) Time-dependent mapping functions: speaker labels from audio, after audio-visual refinement and actual ones. (Adapted from [30].)

clustering or model representation mechanisms which are the most promising properties of unsupervised neural networks, such as the self-organizing feature map (SOFM) and the principal component neural networks. SOFM combines the advantage of statistical data clustering (such as vector quantization) and local continuity constraints (as imposed in the active contour model search). The Expectation Maximization (EM) algorithm is a well-established iterative method for maximum likelihood estimation and for clustering a mixture of Gaussian distributions. Radial basis function (RBF) NNs constitute a neural implementation of the EM algorithm. As already has been said, PCA provides an effective way to find representative (as opposed to discriminating) components of a large set of multivariate data. The basic learning rules for extracting principal components follow the Hebbian rule and the Oja rule. A PCA NN (e.g. the APEX lateral network [49]) can be viewed as unsupervised neural network followed by the Hebbian-type learning. Independent component analysis minimizes the output kyrtosis and it can be seen as a generalization of PCA. The APEX lateral network can be used to extract the independent components of the signal as well.

The multilayer perceptron is one of the most popular neural networks for detection and classification and fusion. Alternatively, RBF NNs could be used for the same purpose. An interesting feature of RBF NNs is that they adopt a one-class-in-one-network structure, where one subnet is designated to one class only. Same modular structure with RBF NNs have the decision-based neural networks. Their probabilistic variant is appropriate for data fusion.

Temporal neural networks allow forward and backward connections between a pair of neurons and sometimes feedback connections from a neuron to itself. The ordinary feedforward NNs are static in the sense that a given input can produce only one set of output values rather than a sequence of data. In contrast, temporal neural networks they employ a recurrent structure, i.e., they have loops due to feedback. The so-called time-delayed NN (TDNN) has feedforward connections enhanced with a carefully chosen set of feedback connections and shift-invariance properties that

are required in speech recognition. Other fully recurrent NN architectures are the real-time recurrent learning networks and the backpropagation through time networks. However, they are much more costly than TDNNs.

7. ABSTRACT DATA PROCESSING MODEL

An abstract data processing model is outlined in this section. Such a model should be adaptive, that is, it should be trained by following an example-based learning procedure to cope with the difficulty in describing explicitly and completely the object properties and the object interrelationships. Such a model should possess expressive and discriminating capabilities. Expressive capabilities are needed for efficient data representation, that is, for data compression and data abstraction. Data compression is of fundamental importance in transmission and storage. Data abstraction can be used for efficient content description. However, for information retrieval, the feature vectors need to be not only compact, but to possess discriminating properties. Immediate examples are the algorithms that apply LDA after having performed PCA [35, 38, 32]. It is worth noting that many well-established algorithms encompass an expressive step (i.e., an approximation step) and a discriminating step (i.e., a detection step) in their description. For example, vector quantization [50] is composed of an optimal encoder for a given decoder and an optimal decoder for a given encoder. The best encoder for a given codebook satisfies the nearest neighbor condition aiming at data partitioning (i.e., a discriminating rule), whereas the best decoder for a fixed partition satisfies the centroid condition (i.e., an expressive rule). Adaptive subspace self-organizing maps [51], radial basis function NN [52], the expectation-maximization algorithm are additional examples that follow the same paradigm. Sometimes, an additional property is desirable, namely, the simplicity. For example, the adaptive channel equalization procedure should be performed at the same time as transmission. In that respect, the Least Mean Squares algorithm satisfies the argument on simplicity whereas the Recursive Least Squares algorithm contradicts with the same argument [19]. From the very beginning of the neural network research the goal was to demonstrate problem solving without explicit programming. The neurons and the networks were supposed to learn from examples and store this knowledge in a distributed way among the connection weights. In principle, everything the neural network does, should be accomplished by a large number of simple local computations using the available input and output signals without involving heavy numerical algorithms [53]. However, this is not a prerequisite, as V. Vapnik argues in his monograph [54]. The majority of learning algorithms minimize the so-called empirical risk and do not possess guaranteed generalization properties. Support Vector Machines (SVMs) is a state-of-the-art pattern recognition technique whose foundations are stemming from statistical learning theory [54]. However, the scope of SVMs is beyond pattern recognition, because they can handle also the other two learning problems, i.e., regression estimation and density estimation, making them prime candidates for the abstract data processing model. Indeed, SVMs are based on guaranteed risk bounds of statistical learning theory, i.e., the so-called structural risk minimization principle. They can implement a set of functions that approximate best the supervisor's response with an expected risk bounded by the sum of the empirical risk and the Vapnik-Chervonenkis (VC) confidence, a bound on the generalization ability of the learning machine, that depends on the so-called VC dimension of the set of functions implemented by the machine.

8. CONCLUSIONS

In this tutorial, a broad overview of the problems arisen in modern communications systems has been attempted. State-of-the-art solutions in content representation, access to multimedia services, noise suppression and content based retrieval have been outlined. The potentialities of neural networks have been described and an abstract data processing model has been proposed.

9. REFERENCES

[1] T. Chen, J.R. Liu, and A.M. Tekalp, Eds., Special issue on Multimedia Signal Processing, Parts I-II, *Proceedings of the IEEE*, vol. 86, nos. 5-6, May-June 1998.

[2] R.V. Cox, B.G. Haskell, Y. Lecun, B. Shahraray, and L. Rabiner, "On the Applications of Multimedia Processing to Communications," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 755–824, May 1998.

[3] S.-F. Chang, A. Eleftheriadis, and R. McClintock, "Next-Generation Content Representation, Creation and Searching for New-Media Applications in Education," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 884–904, May 1998.

[4] M.D. Swanson, M. Kobayachi, and A.H. Tewfik, "Multimedia Data-Embedding and Watermarking Technologies," *Proceedings of the IEEE*, vol. 86, no. 6, pp. 1064–1087, June 1998.

[5] G. Voyatzis, N. Nikolaidis, and I. Pitas, "Copyright protection of multimedia documents: From theory to application," in *Proc. of the IEEE Int. Conf. on Multimedia Computing and Systems*, to appear 1999.

[6] T. Chen, and R.R. Rao, "Audio-Visual Integration in Multimodal Communications," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 837–852, May 1998.

[7] R. Sharma, V.I. Pavlović, and T.S. Huang, "Toward Multimodal Human-Computer Interface," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 853–869, May 1998.

[8] P. K. Varshney, *Distributed Detection and Data Fusion*. N.Y.: Springer-Verlag, 1997.

[9] N.M. Thalmann, P. Karla, and M. Escher, "Face to Virtual Face," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 870–883, May 1998.

[10] J. Bigün, G. Chollet, and G. Borgefors (Eds.), *Audio- and Video-based Biometric Person Authentication*, Lecture Notes in Computer Science, vol. 1206, Berlin: Springer-Verlag, 1997.

[11] A. Jain, R. Bolle, and S. Pankanti (Eds.), *Biometrics: Personal Identification in Networked Society*. Norwell, MA: Kluwer Academic, 1999.

[12] M. Bischel (Ed.), *Proceedings of the First Int. Workshop on Automatic Face- and Gesture- Recognition*, Zurich, Switzerland, 1995.

[13] *Proceedings of the Second Int. Workshop on Automatic Face and Gesture Recognition*, IEEE Computer Society Press, Killington, Vermont, 1996.

[14] *Proceedings of the Third Int. Workshop on Automatic Face and Gesture Recognition*, Nara, Japan, 1998.

[15] T. Ebrahimi, and M. Kunt, "Visual Data Compression for Multimedia Applications," *Proceedings of the IEEE*, vol. 86, no. 6, pp. 1109–1125, June 1998.

[16] A. Flaig, and G.R. Arce, "Rank order diversity detectors for wireless communications," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Arizona, Phoenix 1999.

[17] Y. Wang, and Q.-F. Zhu, "Error Control and Concealment for Video Communication: A Review," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 974–997, May 1998.

[18] X. Song, T. Viero, and Y. Neuvo, "Interframe DPCM with Robust Median-Based Predictors for Transmission of Image Sequences Over Noisy Channels," *IEEE Trans. on Image Processing*, vol. 5, no. 1, pp. 16–32, January 1996.

[19] O. Macchi, *Adaptive Processing. The Least Mean Squares Approach with Applications in Transmission*. Chichester, U.K.: J. Wiley & Sons, 1996.

[20] C. Kotropoulos, and I. Pitas, "Adaptive Nonlinear Filters for Digital Signal/Image Processing," in *Control and Dynamic Systems* (C. Leondes, Ed.), vol. 67, pp. 263–317, 1994.

[21] C. Kotropoulos, and I. Pitas, "Adaptive LMS L-filters for Noise Suppression in Images," *IEEE Trans. on Image Processing*, vol. 5, no. 12, pp. 1596–1609, December 1996.

[22] K. Sobottka, and I. Pitas, "A Novel Method for Automatic Face Segmentation, Facial Feature Extraction and Tracking," *Signal Processing: Image Communication*, vol. 12, pp. 263–281, 1998.

[23] G. Yang, and T.S. Huang, "Human Face Detection in a Complex Background," *Pattern Recognition*, vol. 27, no. 1, pp. 53–63, 1994.

[24] C. Kotropoulos, and I. Pitas, "Rule-based face detection in frontal views," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 97)*, vol. IV, pp. 2537–2540, 1997.

[25] A. Nikolaidis, and I. Pitas, "Facial feature extraction and determination of pose," in *Proc. of the 1998 NOBLESSE Workshop on Nonlinear Model Based Image Analysis*, 1998.

[26] S. Tsekeridou, and I. Pitas, "Facial feature extraction in frontal views using biometric analogies," in *Proc. of the IX European Signal Processing Conference*, vol. I, pp. 315–318, 1998.

[27] E. Osuna, R. Freund, and F. Girosi, "Training Support Vector Machines: an application to face detection," in *Proc. of the Computer Vison and Pattern Recognition Conference*, 1997.

[28] S. Tsekeridou, and I. Pitas, "Speaker dependent video indexing based on audio-visual interaction," in *Proc. of the IEEE Int. Conf. on Image Processing*, 1998.

[29] L. Rabiner, and R.W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, N.J.: Prentice Hall, 1978.

[30] S. Tsekeridou, and I. Pitas, "Audio-visual content analysis for content-based video indexing," in *Proc. of the IEEE Int. Conf. on Multimedia Computing and Systems*, 1999.

[31] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C. v.d. Malsburg, R.P. Würtz, and W. Konen, "Distortion Invariant Object Recognition in the Dynamic Link Architecture," *IEEE Trans. on Computers*, vol. 42, no. 3, pp. 300–311, March 1993.

[32] C. Kotropoulos, A. Tefas, and I. Pitas, "Frontal face authentication using variants of dynamic link matching based on mathematical morphology," in *Proc. of the IEEE Int. Conf. on Image Processing*, vol. I, pp. 122-126, 1998.

[33] J.R. Deller, Jr., J. G. Proakis, and J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*. N.Y.: Macmillan, 1993.

[34] R.C. Gonzalez and R.E. Woods, *Digital Image Processing*. Reading, MA: Addison-Wesley, 1992.

[35] D.L. Swets, and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. on Pattern Analysis and Machine Intelligence* vol. 18, no. 8, pp. 831–837, August 1996.

[36] P. Jonathon Phillips, "Matching Pursuit Filters Applied to Face Identification," *IEEE Trans. on Image Processing*, vol. 7, no. 8, pp. 1150–1164, August 1998.

[37] R. Chellapa, C.L. Wilson, and S. Sirohey, "Human and Machine Recognition of Faces: A survey," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–740, May 1995.

[38] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. on Pattern Analysis and Machine Intelligence* vol. 19, no. 7, pp. 711–720, July 1997.

[39] C. Kotropoulos, A. Tefas, and I. Pitas, "Morphological Elastic Graph Matching applied to frontal face authentication under well-controlled and real conditions," *Pattern Recognition*, accepted for publication, 1999.

[40] A. Tefas, Y. Menguy, C. Kotropoulos, G. Richard, I. Pitas, and P. Lockwood, "Compensating for variable recording conditions in frontal face authentication algorithm," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1999.

[41] A. Tefas, C. Kotropoulos, and I. Pitas, "Enhancing the Performance of Elastic Graph Matching for Face Authentication by using Support Vector Machines," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, submiteed 1999.

[42] V. Chatzis, A.G. Bors, and I. Pitas, "Decision level fusion by clustering algorithms for person authentication," in *Proc. of the IX European Signal Processing Conference*, vol. III, pp. 1309–1312, 1998.

[43] I. Pitas, and A.N. Venetsanopoulos, *Nonlinear Digital Filters: Principles and Applications*. Dordrecht, Holland: Kluwer Academic, 1990.

[44] S.A. Kassam, and H.V. Poor, "Robust Techniques for Signal Processing: A Survey," *Proceedings of the IEEE*, vol. 73, no. 3, pp. 433–481, 1985.

[45] S. Tsekeridou, and I. Pitas, "MPEG-2 Error Concealment Based on Block Matching Principles," *IEEE Trans. on Circuits and Systems for Video Technology*, accepted for publication, 1999.

[46] S. Tsekeridou, F. Alaya Cheikh, M. Gabbouj, and I. Pitas, "Motion field estimation by vector rational interpolation for error concealment purposes," in *IEEE Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing*, 1999.

[47] S. Haykin, *Neural Networks: A Comprehensive foundation*. N.Y.: Macmillan College Publishing Company, 1994.

[48] S.-Y. Kung and J.-N. Hwang, "Neural Networks for Intelligent Multimedia Processing," *Proceedings of the IEEE*, vol. 86, no. 6, pp. 1244-1272, June 1998.

[49] K.I. Diamantaras, and S.-Y. Kung, *Principal Component Neural Networks: Theory and Applications*. N.Y.: J. Wiley & Sons, 1996.

[50] A. Gerso, and R.M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer Academic Publishers, 1992.

[51] T. Kohonen, E. Oja, O. Simula, A. Visa, and J. Kangas, "Engineering Applications of the Self-Organizing Map," *Proceedings of the IEEE*, vol. 84, no. 10, pp. 1358–1384, October 1996.

[52] T. Poggio, and F. Girosi, "Networks for Approximation and Learning," *Proceedings of the IEEE*, vol. 78, pp. 1481–1497, 1990.

[53] L. Holmström, P. Koistinen, J. Laaksonen, and E. Oja, "Neural and Statistical Classifiers-Taxonomy and Two Case Studies," *IEEE Trans. on Neural Networks*, vol. 8, no. 1, pp. 5–17, January 1997.

[54] V. Vapnik, *The Nature of Statistical Learning Theory*. N.Y.: Springer Verlag, 1995.