# Signal Processing for Contactless Monitoring

Mohammad Saad Billah[1] , Md Atiqur Rahman Ahad[2], Upal Mahbub[3]

[1] Nauto, Inc., Palo Alto, California, USA
[2] University of Dhaka, Bangladesh
[3] Qualcomm Inc., San Diego, California, USA
http://www.springer.com/lncs

**Abstract.** Monitoring of human activities from a distance without actively interacting with the subjects to make a decision is a fascinating research domain given the associated challenges and prospects of building more robust artificial intelligence systems. In recent years, with the advancement of deep learning and high-performance computing systems, contactless human activity monitoring systems are becoming more and more realizable every day. However, when looked at closely, the basic building blocks for any such system is still strongly relying on the fundamentals of various signal processing techniques. The choices of a signal processing method depends on the type of signal, formulation of the problem and the choices of higher level machine learning components. In this chapter, a comprehensive review of the most popular signal processing methods used for contactless monitoring are provided highlighting their use across different activity signals and tasks.

**Keywords:** Signal processing; contactless monitoring; activity signals

## 1 Introduction

In recent times, contactless human monitoring has gain a lot of traction. Application areas of such systems include monitoring breathing pattern, respiratory rate and other vital signs [1–3], event recognition [4], human motion classification [5] and analysing crowded scenes [6].

The vast proliferation of contactless sensors has enabled contactless activity monitoring with different activity signals. Among these, audio-based, light-based, and radio-frequency based sensors are most widely used. Different types of sensors has different strengths and weaknesses. For example, radio frequency and proximity sensors provide cheap contactless monitoring but suffers from low accuracy and high environmental inferences. Light based sensors such as camera, depth sensors, and LIDARs has accuracy and resolution but also expensive and requires high computational power for processing. Table 1 lists different aspects of most popular signal sources for contactless monitoring.

Signal processing is one of the most important and fundamental block of contactless monitoring. From sensing the physical world to making a decision e.g. recognizing, modeling, understanding etc. - signal processing techniques are used in every step in between.

**Table 1.** Different aspects of most popular signal sources for contactless monitoring.

| Sensor | Advantages/Disadvantages | Notable Use-Cases |
|---|---|---|
| Audio-based<br>– Speech<br>– Acoustic<br>– Ultrasound | **Advantages**<br>– Moderate to High accuracy<br>– Moderate to low cost<br>– Ultrasound is precise for determining distances, highly sensitive to motion and has Long operating range<br><br>**Disadvantages**<br>– Easily influenced by other audio signal/noise<br>– Prone to false detection<br>– Range limited<br>– Privacy issues<br>– Ultrasound is unidirectional, sensitive to temperature and angle of target and performance drops at very close proximity | – Intelligent personal assistants (IPAs) [7]<br>– Audio-based context/scene recognition [8, 9]<br>– Human activity recognition [10]<br>– Heart and respiration rate monitoring [2]<br>– Office and indoor activity analysis [11, 12]<br>– Rehabilitation support [13] |
| Radio Frequency-based<br>– RF<br>– WiFi | **Advantages**<br>– Low cost<br>– Simple computation<br><br>**Disadvantages**<br>– Environmental Inference | – Measuring vital signs [3]<br>– Indoor/outdoor localization and tracking |
| Light-based<br>– Infrared Sensors<br>– Thermal Imaging Sensors<br>– 2D Cameras<br>– Depth Sensors and Hybrid Sensors | **Advantages**<br>– High Accuracy<br><br>**Disadvantages**<br>– High cost<br>– Privacy issues<br>– Influenced by illumination, pose, occlusion and noise | – Activity recognition from thermal videos [14]<br>– Facial expression analysis<br>– Action recognition for robotics and HCI<br>– Crowded scene analysis, anomaly detection<br>– Pose prediction<br>– Pedestrian detection for autonomous driving<br>– Technology for assisted living such as fall detection<br>– 3D body shape, face and hand modeling for augmented and virtual reality applications<br>– Precise tracking of face and eye for autonomous driving scenarios<br>– Liveliness detection for anti-spoofing of authentication systems |
| Other Sensors<br>– Passive Infrared Sensors (PIR)<br>– Proximity Sensor | **Advantages**<br>– Low cost<br>– Simple computation<br><br>**Disadvantages**<br>– Low Accuracy<br>– Limited usability | – Activity recognition and tracking [15]<br>– Collision avoidance technology for blind people and wheelchairs<br>– Motion-based automatic control of switches for smart home systems |

Figure 1 shows typical lifetime of a signal in contactless monitoring that starts from activity signal acquisition by sensing the world using different contactless sensors such as microphone, camera, Lidar, infra-red, ultrasonic sensors etc. Different sampling and windowing techniques are used to acquire discrete signals from the continuous real world. The signal then goes through different pre-processing steps such as denoising and other filtering methods to enhance its quality. Features are then extracted from the signals to be used by different activity analysis algorithms. This chapter briefly discusses the signal processing steps and their applications.

The chapter is organized as follows: Section gives an overview of different sampling and windowing techniques. Section 3 discusses time and frequency

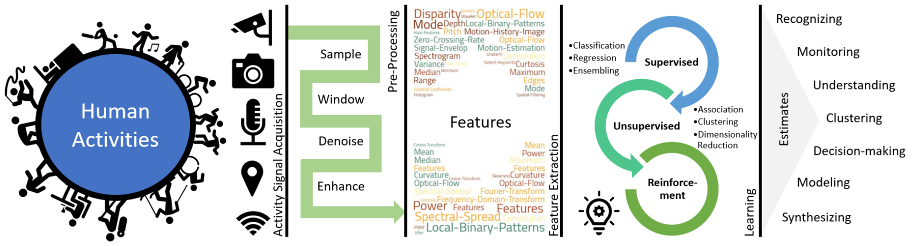


**Fig. 1.** Any contactless human activity analysis system usually follows this pipeline from left to right.

domain processing techniques and their applications. In section 4, some widely used feature descriptor and their extraction techniques have been described. Different dimensionality reduction methods and their applications have been discussed in Section 5. Conclusions are drawn in Section 6. Activity analysis algorithms are beyond the scope of this chapter and therefore, not discussed.

## 2 Activity Signals Sampling and Windowing Techniques

Signal sampling and windowing are two important steps of signal processing that is applied during or right after signal acquisition and plays an important role in the performance of the system. This section discusses applications of signal sampling and windowing methods and the impact of windowing in activity analysis.

### 2.1 Applications of Signal Sampling

Sampling is the process of converting a continuous time signal from the real, analog world to a discrete time signal in the digital domain. The value of the analog signal is measured at certain time intervals to read 'Samples' for the digital domain. Analog signals are continuous in both amplitude and time, while the sampled digital signals are discrete in both. If a continuous signal is sampled at a frequency $f_s$, the frequency components of the analog signal are repeated at the sample rate resulting in the discrete frequency response repeated at origin, $\pm f_s$, $\pm 2f_s$, and so on. According to Nyquist-Shannon sampling theory [16], Sampling needs to be at least at Nyquist rate (2 x the maximum frequency of a signal $f_{max}$) or more for exact reproduction. Sampling below Nyquist rate ($f_{Ny}$) causes information loss and aliasing. Unwanted components are introduced in the reconstructed signals during aliasing when signal frequencies overlap due to low sampling rate, while some frequencies of the original signal gets lost in the process. Results of sampling a simple sine wave at different rates are shown in Fig. 2. In many real-life applications, noises represent the highest frequency component of a signal and aliasing of those frequencies are undesired. Hence, low pass filtering is performed before sampling to prevent aliasing of the noise components.

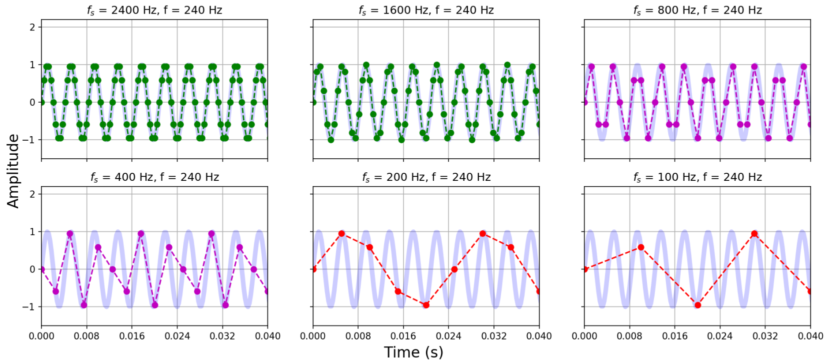


**Fig. 2.** Effect of sampling frequency. A 240Hz sine wave is sampled using sampling frequencies of 2400Hz, 1600Hz, 800Hz, 400Hz, 200Hz and 100Hz (From top-left to bottom-right image). The aliasing cases are shown in the bottom row where $f_s < 2 \times f_{max}$. Clearly the sampled signals in the bottom figures has unwanted components due to aliasing

While "temporal aliasing" occurs in signals sampled in the time domain (such as audio signals), it can also occurs for spatially sampled signals, such as an image - a phenomenon referred to as "spatial aliasing". Spatial sampling can cause jaggies on the edges as commonly seen on low resolution versions of an image (example shown in Fig. 3). Other artifacts of aliasing includes wagon wheel effect[1] for temporal sampling, temporal strobing when sampling in space-time, Moiré effect [17] when sampling texture coordinates and sparkling highlights.

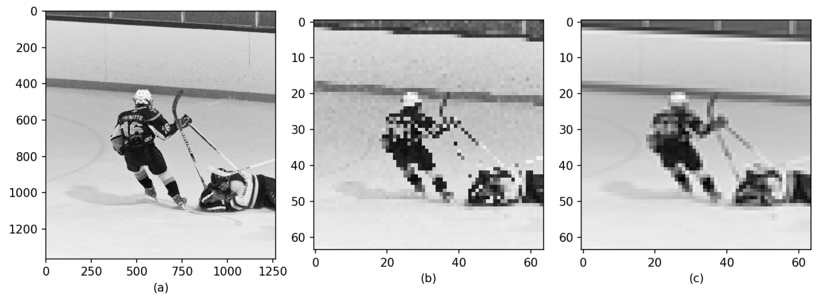


**Fig. 3.** (a) Original 1365 × 1365 pixel image obtained and modified from the Open Image Dataset V6 [18]. (b) Image down-sampled to 64 × 64 pixels by sampling every fourth sample and applying a box filter. The jagged patterns and high dimensional noise introduced by aliasing and the box filtering are clearly visible. (c) Down-sampled to 64 × 64 pixels using an anti-aliasing Lanczos filter [19].

[1] `https://michaelbach.de/ot/mot-wagonWheel/index.html`

For spatio-temporal data like videos collected for surveillance, the temporal sampling rate needs to be high to prevent strobing effect as well as to ensure no critical information is lost due to lower sampling rate, which will otherwise defeat the purpose of a surveillance system. In Fig. 4 the histograms for average frame per second rate of video surveillance systems for two different years are shown[2]. According to IPVM statistics [3] the average frame for video surveillance systems increased from $6-8$fps in 2011 to $\approx$ 10fps in 2016 statistics and then to 15fps in 2019. It is understandable that commercial video surveillance systems are inclining towards higher frame rates to ensure high-quality seamless video streams for the customers. While a increased frame rate can lead to higher bandwidth requirements for such a system, depending on the compression methods used, bandwidth does not increase linearly with frame rate[4].

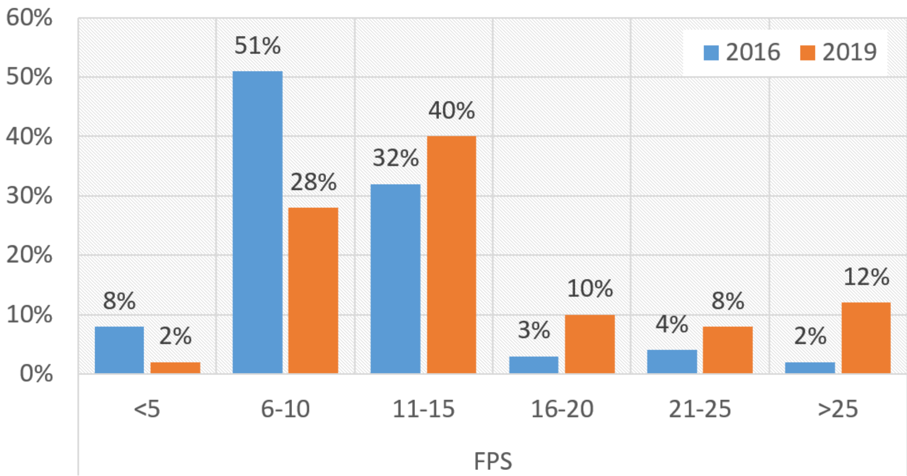


**Fig. 4.** Histogram of average FPS for video surveillance. There is a clear trend of increasing FPS over the years.

Table 2 lists the specifications of some cameras used in the industry. As listed in the table, it can be seen that the newer models have more spatio-temporal sampling rate. Table 3 lists the specifications of LIDAR sensors.

### 2.2 Impact of Signal Windowing on Activity Analysis

Windowing plays vital role in the activity analysis performance. Given the type of signal and application, the method and duration of windowing can vary widely. For example, in [20], the authors demonstrated that the size of the window plays

[2] Based on `https://ipvm.com/reports/frame-rate-surveillance-guide`
[3] `https://ipvm.com/reports/avg-frame-rate-2019`
[4] `https://ipvm.com/reports/frame-rate-surveillance-guide`

**Table 2.** Specifications of some surveillance cameras used in the industry

| Model | Release | Resolution | FPS |
|---|---|---|---|
| Axis M3004 | 2012 | 1.0 MP | 30 |
| Sony SNC-EM600 | 2013 | 1.3 MP | 30 |
| Reolink RLC-423 | 2015 | 5 MP | 25 |
| Reolink RLC-410 | 2017 | 5 MP | 25 |
| Hanwha (Samsung) PNO-9080R | 2016 | 12 MP | 20 |

**Table 3.** Specifications of some LIDER sensors used in the industry

| Model | Release | Range | Resolution | Scan rate | Accuracy | Weight |
|---|---|---|---|---|---|---|
| Velodyne HDL 64 | 2007 | 120m | 0.08/0.4 | 2.2M | 2cm | 12.7 kg |
| Velodyne Puck Ultra | 2016 | 200m | 0.1/0.33 | 1.2M | 3cm | 0.925 kg |
| Quanergy M8 | 2016 | 150 | 0.03 | 1.26M | 3cm | 0.900 kg |

a significant role in determining speech intelligibility and the optimum hamming window duration for speech reconstruction from short-term magnitude spectrum is 15-32 ms. When choosing a window for a 1-D signal, the following factors can be considered:

- width of the main lobe,
- spectral leakage from the attenuation of the side lobes, and
- rate of attenuation of the side-lobes.

In Fig. 5, the five time domain window functions, namely, rectangle, bartlett, hamming, hanning and blackman [21, 22], with their respective frequency domain responses are shown. The values of the window functions at the $n$-th sample for a window length of $N$ where $0 \leq n \leq N$ are defined as follows:

$$\text{Rectangular, } w[n] = 1, \tag{1}$$

$$\text{Bartlett, } w[n] = 1 - \frac{n - N/2}{N/2}, \tag{2}$$

$$\text{Hamming, } w[n] = 0.54 - 0.46cos(\frac{2\pi n}{N}), \tag{3}$$

$$\text{Hanning, } w[n] = 0.5 - 0.5cos(\frac{2\pi n}{N}), \tag{4}$$

$$\text{Blackman, } w[n] = 0.42 - 0.5cos(\frac{2\pi n}{N}) + 0.08cos(\frac{4\pi n}{N}). \tag{5}$$

As can be seen in Fig. 5, the rectangle window has the narrowest main lobe but higher side lobe strength, while the other windows have wider main lobe but lower side lobes. Hence, a rectangular window would be a better choice to separate two signals with similar frequency and strength but worse choice for identifying two signals with different frequencies and strength due to the spectral leakage and lower rate of attenuation of the side lobes [23, 22]. The 1-D signal windowing techniques are extended to 2-D spatial windows, also known as kernels. The choice of a kernel depends on the type of the image processing task. A simple example would be Gaussian kernels that are widely used for image smoothing and de-noising [24]. An isotropic 2D Gaussian kernel of unit magnitude has the following form:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{6}$$

where, $x$ and $y$ are the pixel indexes from the center and $\sigma$ is the standard deviation.

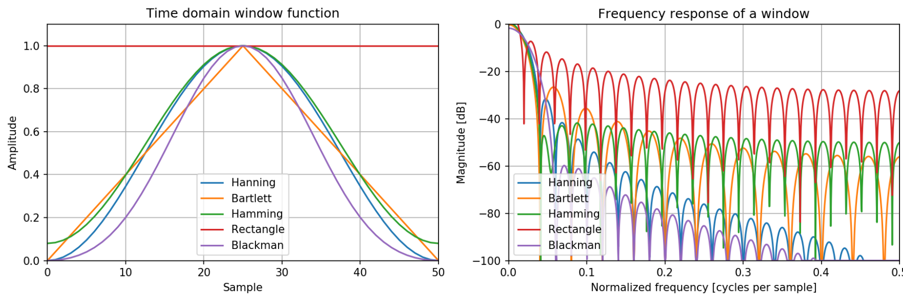


**Fig. 5.** Time (left) and Frequency (right) domain responses of five different window functions.

Temporal windowing, a.k.a temporal segmentation is an integral part of action recognition systems for real-time applications. Sliding windows are the most common windowing techniques for such scenarios[25]. However, based on specific use-cases the length of the temporal window might or might not change dynamically. Also, the temporal overlap between consecutive windows are considered. Also, the size of the windows can be dynamically expanded or shrunk based on activity inference in some system[25]. In a macro-level view, the design choices are as follows:

1. Fixed-length window
   - Non-overlapping windows
     - No dynamic shrinking and/or expansion
     - Dynamic shrinking and/or expansion
   - Overlapping windows
     - No dynamic shrinking and/or expansion
     - Dynamic shrinking and/or expansion
2. Dynamic-length window
   - Non-overlapping windows
     - No dynamic shrinking and/or expansion
     - Dynamic shrinking and/or expansion

- Overlapping windows
  - No dynamic shrinking and/or expansion
  - Dynamic shrinking and/or expansion

When training machine learning systems, windowing plays an implicit yet vital role for most application when creating mini-batches. In a recent work, the authors proposed a framework that uses a sliding-window data scheduler to achieve state-of-the-art performance for instance classification task [26]. More examples use cases of windowing associated with deep learning include object localization [27, 28], autonomous navigation [29], window slicing and pooling techniques in deep neural networks [30] and modeling temporal patterns [31].

Now that we have established the importance of signal sampling and windowing techniques on acquiring the sensor data in a convenient way for digital processing, we move forward to discuss how time and frequency domain signal processing approaches are being utilized to extract or meaningful information from those data in the next section.

# 3 Time and Frequency Domain Processing for Contactless Monitoring

Time and frequency domain techniques are applied to activity signals to analyze and enhance the signal. Different frequency domain transform techniques are frequently used in activity analysis. Time and frequency domain filtering is another important and widely used technique used for signal enhancement. This section first discusses the applications of frequency domain transforms. The latter part of the section provides a brief introduction to filtering and some notable use cases.

## 3.1 Applications of Frequency Domain Transforms

Frequency domain transforms are commonly applied to activity signals to analyze and leverage the periodicity information for decision making purposes. A very practical use-case is Remote photoplethysmography (rPPG) for monitoring heart-rate from surveillance videos [32, 33]. For example, in [34] the authors extracted the pixels of interest from the face images in consecutive video frames, took the average pixel values for each of the RGB channels, filtered-out low-frequency components and investigated the frequency-domain representation to find the frequency with maximum power which is a close approximation of the heart-rate. The most popular frequency domain representation for such applications is the power spectral density (PSD) which is a measure of signal power at different frequencies. For speech analysis, such concentration of acoustic energy around a particular frequency, known as formants, have been used for a wide range of applications including automatic speech recognition [35], voice activity detection [36] and speech enhancement [37].

When dealing with 1D temporal signals such as speech or ultrasound, one of the most popular analysis tools is the short-time Fourier Transfrom (STFT) which is frame-level frequency domain representation [38–40]. A visual extension of STFT is a Spectrogram (also known as sonographs/voicegrams/voiceprints), which is commonly plotted as a frequency vs time series 2D image where the pixel intensities represent the magnitude of the frequency component [41]. Spectrograms are calculated for short-time, overlapped, sliding windows of $T$ time samples $\mathbf{x} = (x_1, x_2, ..., x_T)$ where the temporal duration of the window is chosen to be small(typically 25 to 35 ms) to ensure that the speech within that frame will be stationary. The value of the Spectrogram at the $k$'th frequency bin is defined as

$$\mathrm{Spec}_k(x) = |\sum_{t=1}^{T} e^{ikt} x_t|^2 = (\sum_{t=1}^{T} cos(kt) x_t)^2 + (\sum_{t=1}^{T} sin(kt) x_t)^2. \tag{7}$$

Spectrograms are convenient for visualizing the effects of speech enhancement as an be seen in Fig.6 obtained with permission from [42]. In [42], the authors addressed the problem of acoustic echo cancellation from speech under noisy condition. Apart from the spectrogram to visualize the results, the authors also applied spectral subtraction [43] for noise reduction which involves transforming the noisy signal into frequency domain using Fast Fourier Transform (FFT) [44] on the short-term windows of the discretized speech signal and subtracting frequency-domain estimate of noise spectrum (usually obtained and updated from speech pauses) before reverting the signal to time domain samples using inverse FFT (IFFT).

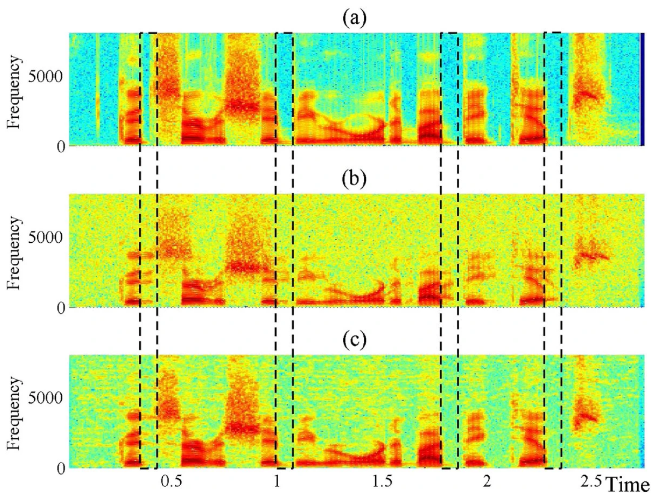


**Fig. 6.** Spectrogram for (a) original, (b) echo and noise corrupted, and (c) enhance signal - reproduced with permission from [42].

Typically, spectrograms uses linear frequency scaling. Mel-frequency scales are developed inspired by the properties of human auditory system to follow a quasi-logarithmic spacing. Mel-frequency filters are non-uniformly spaced in frequency domain with more filters in the low frequency region compared to higher frequency regions. Cepstral coefficients obtained for Mel-spectrum are popularly known as MFCC (Mel-frequency Cepstral Coefficients) features, which can be considered as “biologically inspired” speech features [45–47].

Following are notable use-cases of different variations of frequency-domain transforms in the contact-less human activity analysis domain:

– **Wavelet transform**[48]: Data compression such as JPEG2000 image compression standard [49]; video-based human activity recognition [50, 51]; Doppler range control radar sensor-based fall detection [52]; WiFi signal-based human activity recognition [53]; audio compression [54].
– **Discrete Cosine Transform**[55]: 3D motion analysis [56]; audio compression [54];
– **Laplace Transform**[57–59]: Non-articulatory sound recognition [60];
– **Z-transform**[57, 58]: Speech recognition[61]; Speech modeling and analysis [62]; Pole-zero representation for linear physical system for analysis and filter design [63].

### 3.2 Time and Frequency Domain Filtering

A filter is a function or operator that modifies a signal by performing mathematical operations to enhance or reduce certain aspects of the signal. If $n$-dimensional signal is represented as an $n$-dimensional function, then mathematically, a linearly-filtered 2D-signal can be represented as

$$g(x, y) = \sum_{m,n}^{W} f(x + m, y + n)h(m, n). \tag{8}$$

Here, $h$ is known as the filter kernel and $h(m, n)$ is known as a kernel weight or filter coefficients.

Original Image

5x5 Box Filter Kernel

| | | | | |
|---|---|---|---|---|
| 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |
| 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |
| 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |
| 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |
| 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |

Filtered Image

**Fig. 7.** Left - original image, middle - 5x5 box filter kernel, right - filtered image.

A simple filter kernel is the moving average or box filter that computes the average over a neighborhood or window. Fig. 7 shows an example application of such box filter which is also a form of low-pass/blurring filter. Applications of signal filtering include enhancement such as denoising and resizing, information extraction such as texture and edge extraction, pattern detection such as template matching etc. Fig. 8 shows such examples where filtering is used to extract vertical and horizontal edges.

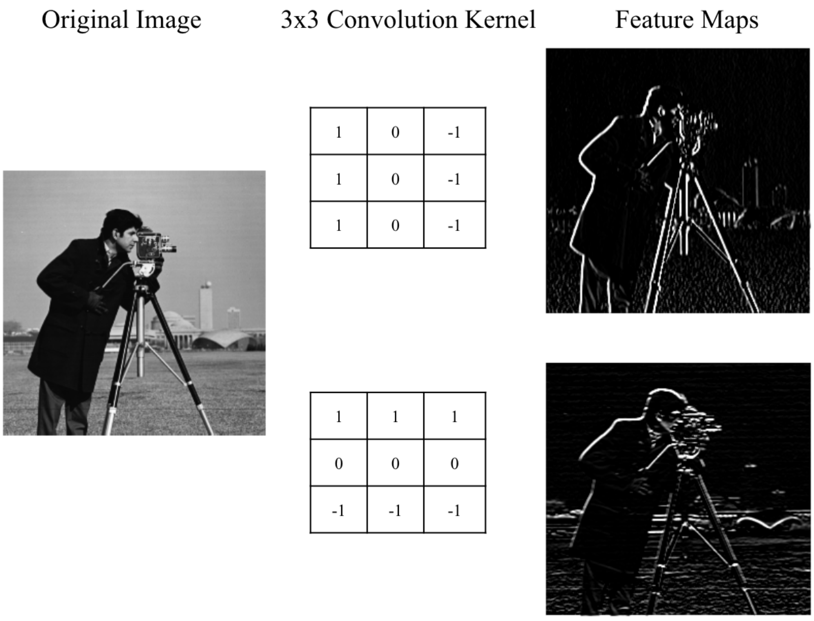


**Fig. 8.** Left - original image. Middle:top - kernel that emphasizes vertical edges, bottom - kernel that emphasizes horizontal edges. Right - output feature map corresponding to the kernel on the left.

An extension to the basic filters are adaptive filters whose coefficients change based on an objective or cost function (eqn. 8). These filters are used to modify input signals such a way so that its output is a good estimate of a desired signal. Examples include Least Mean Square (LMS) adaptive filters, Recursive Least Square (RLS) adaptive filters, adaptive Wiener filters, adaptive anisotropic filters etc. Adaptive filtering has applications active noise control [64–66], echo cancellation [67], biomedical signal enhancement [68], tracking [69], equalization of communications channels etc.

Some notable use cases of filtering such as contrast stretching and histogram equalization, denoising, and convolutional filters are briefly discussed next.

**Contrast Stretching and Histogram Equalization** In a poorly contrasted image, a large number of pixels occupy only a small portion of the available range of intensities. The problem can efficiently be handled by histogram modification and thereby reassigning each pixel with a new intensity value so that the dynamic range of gray levels is increased Contrast Stretching and Histogram Equalization are such two contrast enhancement technique.

The idea behind contrast stretching is to increase the dynamic range of the gray levels in the image being processed [24]. Contrast stretching is a simple image enhancement technique that attempts to improve the contrast in an image by 'stretching' the range of intensity values it contains to span a desired range of values, e.g. the the full range of pixel values that the image type concerned allows.

Histogram Equalization is a method that increases the contrast of an image by increasing the dynamic range of intensity given to pixels with the most probable intensity values. The histogram equalization is a basic procedure that allow to obtain a processed image with a specified intensity distribution. Sometimes, the distribution of the intensities of a scene is known to be not uniform. The goal of histogram equalization is to map the luminance of each pixel to a new value such that the output image has approximately uniform distribution of gray levels. In order to find the appropriate mapping, the cumulative distribution function (CDF) of the pixel values of the original image is matched with a uniform CDF [70].

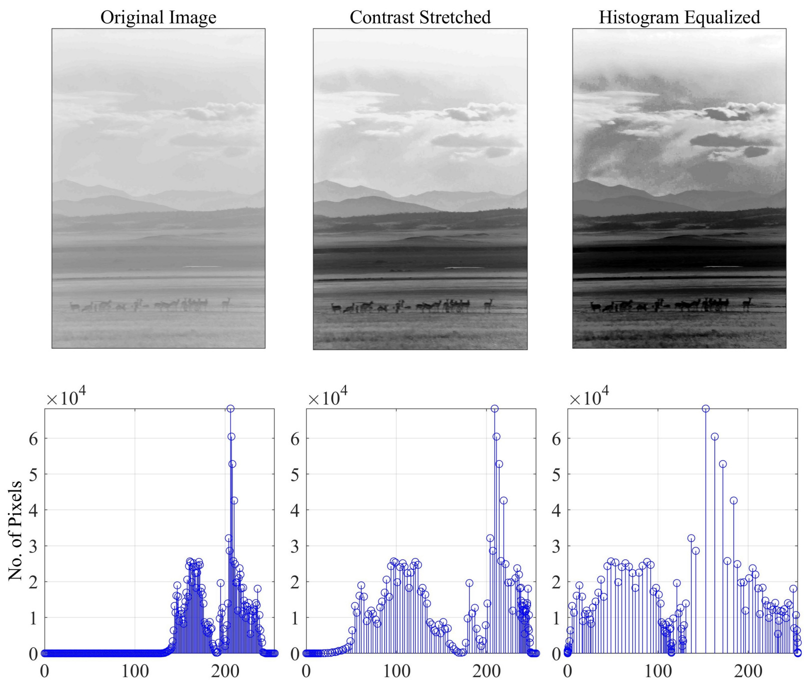


**Fig. 9.** Top: left - original image, middle - image enhanced by contrast stretching, right - enhanced by histogram equalisation. Bottom: histogram of pixel values for the corresponding top row image.

**Denoising** Denoising is the process of removing noise from a signal. Noise reduction techniques exist for both 1D signals such as speech and 2D signals such as images. Denoising is generally a pre-processing step used before extracting features from a signal. If we have a signal $\mathbf{x}$ that is corrupted with noise $\eta$ as

$$\bar{f}(x,y) = f(x,y) + \eta(x,y)$$

then a denoising filter $h$ is a filter designed to estimate $f$ such that

$$f(x,y) = \sum_{m,n}^{W} \bar{f}(x+m, y+n)h(m,n) \tag{9}$$

For example, median filter is a denoising filter that perform very well on images containing binary noise such as salt and pepper noise. The median filter considers each pixel in the image in turn and looks at its nearby neighbors to make sure that it is representative of its surroundings by replacing it with the median of those values. It is a non-linear filter and its output is the following-

$$f(x,y) = \text{median}\left(\bar{f}(x+m, y+n), (m,n) \in W\right) \tag{10}$$

In general, the median filter allows a great deal of high spatial frequency detail to pass while remaining very effective at removing noise on images where less than half of the pixels in a smoothing neighborhood have been effected. One of the major problems with the median filter is that it is relatively expensive and slow to compute since finding the median requires sorting all the values in the neighborhood into numerical order. A common enhancement technique is to utilize the relative sorting information from the previous neighborhood window to the next.

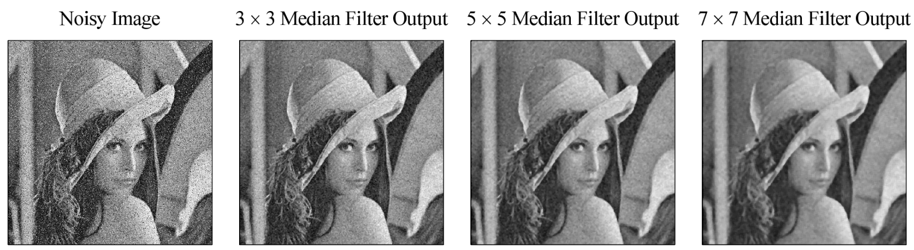


**Fig. 10.** Left to right: Median filter of sizes $3 \times 3$, $5 \times 5$ and $7 \times 7$, respectively, are applied on a noisy image (left-most) for denoising.

**Convolutional Filters** Another application where this kind of filtering is central is the convolutional neural network or CNN [71]. Convolutional neural networks use multiple filters in parallel where each kernel extracts specific feature of the input. The convolutional layers are not only applied to the input, but they are also applied to the output of other layers. The output of these layers are called feature maps as they contain valuable information extracted from the input that helps the network perform its task. Unlike traditional computer vision, where the kernels are generally hand-crafted, CNN learns the weight of the kernels during the training of the network. For example, in [72] a wonderful demo for visualizing the output of each convolution layer for a convolutional neural network trained to perform handwritten digit classification is presented. The input (a handwritten digit '4'), intermediate convolutional and fully connected layer output features as well as final predicted class for a convolutional neural network trained on the MNIST dataset [73] is shwon in Fig. 11. The network used is the famous LeNet-5 proposed in [74].

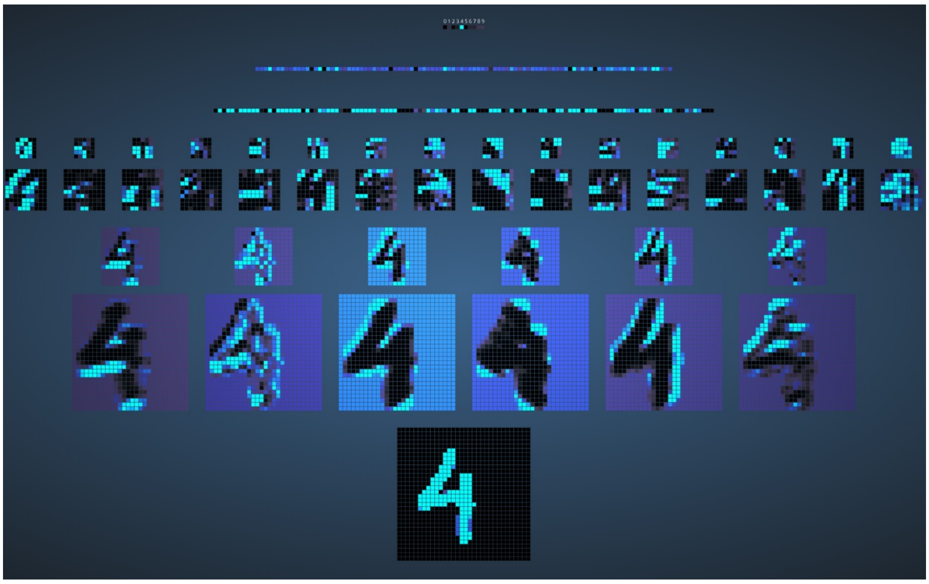

**Fig. 11.** Input, intermediate features and classification output (bottom to top) of a CNN produced using the web tool provided by [72]

It can be observed that the output of the lower level convolution layers (second and third rows from bottom) are visually interpretable such as edges and corners of the input image, whereas the visual information are abstracted out in the higher level features producted by the fully connected layers (third and second rows from the top) in an effort to compress and convert the data in the output classification domain.

The time and frequency domain filtering techniques discussed in this section are heavily utilized for signal pre-processing as well as meaningful feature extraction. In the next two sections, we discuss the low and high level feature feature extraction methods that are direct application of different signal processing methods.

## 4 Feature Extraction

A feature vector or descriptor encodes a signal such a way that allows it to be compared with another signal. A local descriptor describes or encodes a path within the signal. Multiple local descriptors are used to encode or compare signals. Local descriptors are used in application like activity recognition. A global descriptor describes the whole image. Global descriptors are generally used for applications like activity detection, and classification etc.

### 4.1 Local Descriptors

Local descriptors describe a feature on the basis of unique patterns present in the neighborhood of the feature location. Some feature descriptor algorithm has its own feature detector. However, individual detectors can also paired with different descriptors. For convenience, this section is organized in two subsections. Section 4.1 discusses the time/spatial domain features and Section 4.1 discusses the frequency domain features.

**Time/Spatial Domain Features.** Time or spatial domain features are the features that extracted from the time or spatial domain representation of the signal. Some of the widely used low-level features and their applications is briefly discussed next. They are generally easy to define and extract and has weaker requirements for invariant extraction [75]. Latter part of the section discusses some of the widely used high level local feature decriptors.

Zero Crossing Rate (ZCR) is a time domain feature that measures the noisiness of the a signal. It is the rate of sign-changes of the signal. For the $i-th$ frame of lengh $N$ with samples $x_i(n)$ where $n = 0, 1, \ldots, (N-1)$, the ZCR is defined as

$$Z_i = \frac{1}{2N} \sum_{n=1}^{N-1} |sgn[x(n)] - sgn[x(n-1)]|, \tag{11}$$

where, $sgn[x]$ is the sign function defined as

$$sgn[x] = \begin{cases} -1 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0. \end{cases} \tag{12}$$

Kim et al. in [76] proposed a new model for speech recognition in noisy environments that uses ZCR. It is also used in speech-music discrimination [77], music genre classification [78], and several other applications.

The signal envelope of an oscillating signal is the smooth curve outlining its extremes. Speech signal envelope and its change are used in speech recognition applications [79].

The short term energy of a signal is another simple time domain feature. If a signal window contains $N$ samples, then the short-term power is computed according to the equation:

$$E = \frac{1}{N} \sum_{n=1}^{N} |x_i(n)|^2 \tag{13}$$

The short-term power exhibit high variation over successive speech window i.e. power envelope rapidly alternates between high and low power states. Therefore, an alternative statistic, which is independent of the signal intensity, is the standard deviation by mean value ratio is also used. Signal power based features are used in speech activity detection applications [80].

Edge is an important feature used computer vision. An edge in an image is a local change in the image intensity. Edges in an image are associated with discontinuity in the image intensity which generally corresponds to discontinuities in depth, variations in material properties or scene illumination etc. Canny, Sobel, Prewitt are some examples of edge detectors.

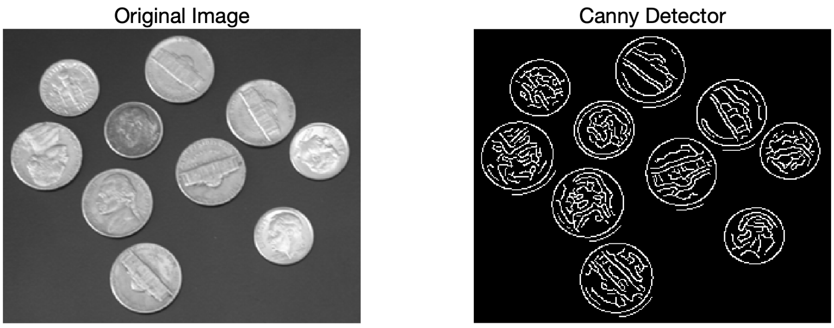


**Fig. 12.** Example of edge features (Canny).

Corner features are frequently used in motion detection, video tracking, and object recognition. A corner is defined as the intersection of two edges. In the region around a corner, image gradient has two or more dominant directions. Corners are easily recognizable in an image when looking through a small window and shifting the window in any direction give a large change in intensity. The Shi-Tomasi detector[81] and the Harris detector [82] are examples of two popular corner detectors

Among the high level local descriptors, Scale Invariant Feature Transform (SIFT)[83] is one of the most popular feature descriptor for images. SIFT has scale in-variance property. The feature extracted by the SIFT algorithm is called feature descriptor which consists of a normalized 128-dimensional vector and it describes a feature point in terms of location, scale, orientation. SIFT feature is used in activity analysis such behaviour detection [84], activity recognition [85] etc.

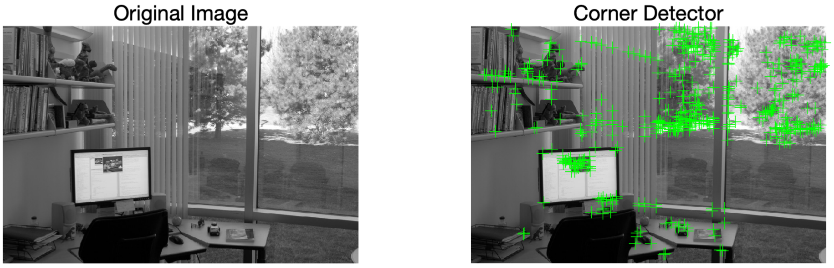


**Fig. 13.** Example of Haris corner features.

Another edge/gradient-based feature detector inspired by the SIFT is speeded up robust features (SURF) [86]. The main interest of the SURF approach lies in its fast computation of operators using box filters, thus enabling real-time applications such as tracking and object recognition [87–90].

Despite their good performance, both SIFT and SURF are quite memory intensive (512 bytes and 256 bytes respectively per feature point) which makes them infeasible for resource-constrained applications. Binary Robust Independent Elementary Features (BRIEF) provides a shortcut to find binary string from the floating point feature descriptors [91]. One important point is that BRIEF is a feature descriptor, it doesn't provide any method to find the features, so a feature detector like SIFT, SURF, or FAST[92] has to be used to locate the keypoints. Gunduz et al. extracted crowd dynamics using BRIEF features in [93].

An efficient alternative to SIFT and SURF that provides better performance than BRIEF is Oriented FAST and Rotated BRIEF (ORB) descriptor. BRIEF performs poorly with rotation, so ORB steers BRIEF and according to the orientation of keypoints. ORB features has been used in activity forecasting [94] and motion detection [95] applications.

Binary Robust Invariant Scalable Keypoints (BRISK) [96], Fast Retina Keypoint (FREAK) [97], KAZE [98], and Accelerated-KAZE (AKAZE) [99], are some other widely used feature descriptors. Fig. 14 shows the performance of different low level feature detectors.

**Frequency Domain Features** The spectral centroid and the spectral spread are two measures of spectral position and shape of a signal. The spectral centroid is the center of gravity of the spectrum and the spectral spread is the second central moment of the spectrum. These features are useful in audio analysis tasks such as audio brightness prediction [100], audio timbre measurement [101] etc.

Spectral Entropy is another frequency domain feature. To compute spectral entropy, first the signal spectrum is divided into $L$ sub-bands, the energy $e_l$ of each sub-band is then normalized by the total spectral energy, and the entropy is finally computed as

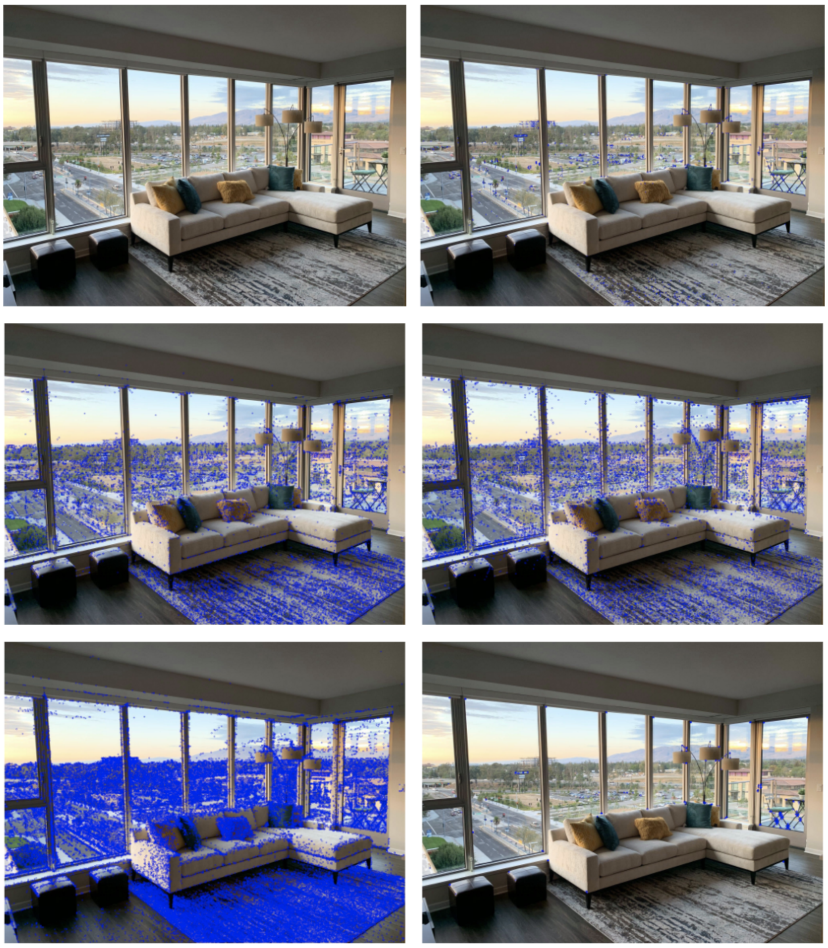

**Fig. 14.** Low level feature detection, from top left to bottom right, original, Shi-Tomasi, SIFT, SURF, FAST, and ORB.

$$H = -\sum_{l=0}^{L-1} \frac{e_l}{\sum_{l=0}^{L-1} e_l} \log \left( \frac{e_l}{\sum_{l=0}^{L-1} e_l} \right) \tag{14}$$

Standard deviation of sequences of spectral entropy is used to classify sound classes [102, **?**]. Other applications include music fingerprinting [103], encoding [104], signal monitoring [105] etc. A variant of spectral entropy called chromatic entropy has been also used in order to efficiently discriminate between speech and music.

Other examples of 1-D low level features include spectral flux [106], spectral rollof, etc. Frequency domain techniques can be used in images in the same way as one dimensional speech signals. However, images do not have their information encoded in frequency domain, which makes this techniques much less useful to understand information encoded in images [107].

### 4.2 Global Descriptors

Among different available global descriptors, Motion History Image (MHI) and its variants are very widely explored for various human action recognition and applications for a longer period [108–111]. MHI template or image can incorporate the entire motion information of a motion sequence or video in a compact manner [110]. It has been a very useful template, especially when a single person's action or motion information is needed to extract. So, from a video of many frames, finally we can create just a single image called MHI. A binarized image from MHI or based on MHI is called Motion Energy Image (MEI) [110]. The MEI retains the entire motion area or locations where there were any motion information in the entire video sequence.

Figure 15's (top row) depicts five Motion History Images for an action for for the first 10 frames (as shown in the 1st column), until 15 frames, until 34 frames, until 36 frames, and until the end of 46 frames from the beginning [108]. The respective Motion Energy Images are demonstrated in Figure 15 (bottom row) for the same action. These are computed from a gesture from the Kaggle Gesture Challenge 'ChaLearn' Database.

The MHI and MEI pairs have been explored for many action recognition and analysis. The MHI provides the *history* of the motion information and direction or flow of the motion. On the other hand, the MEI retains the motion region or area – thereby, it provides the *energy* or the points of motion areas. The MHI is a grayscale image, whereas the MEI is a binary image. However, smarter silhouette sequence can allow us to get better MHI template. The MHI also provides the temporal changes and directions of the motion. For example, if a video has sitting to standing sequences, the produced MHI can give a final image where past or initial motion information becomes less-brighter than the later or final motion regions with brighter pixel values. From these, we can assume that

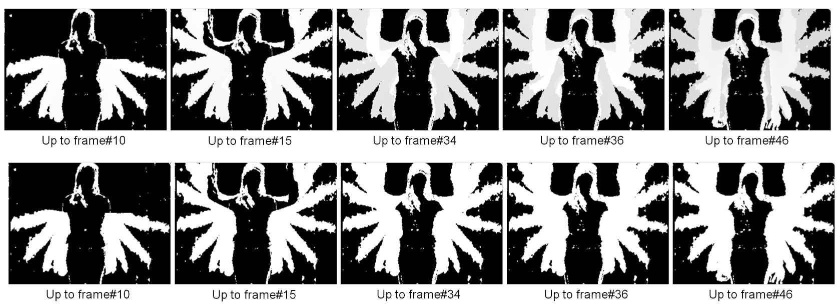


**Fig. 15.** Examples of the computation of the MHI (top row) and the MEI (bottom row) images for a gesture at different temporal states from the beginning of the action [108].

the motion has lower to upper direction. The MHI representation is less sensitive to shadows, silhouette noises or minor missing parts.

The MHI can be used for Action recognition and analysis, Gait recognition, Gesture recognition, Video analysis, Surveillance, Face-based depressive symptomatology [112] analysis, Fall detection [113], Visualization of the hypoperfusion (decreased blood flow) in a mouse brain [114], Depth image-based action recognition and removal of self-occlussion [115], Body movement trajectory recognition [116], Biospeckle assessment of growing bacteria [117], and Emotion recognition [118].

It has been also explored for gaming and other interactive applications and in real-time, as the computational cost is really minimal.

There have been a number of variants at the top of the MHI. For example, Average Motion Energy (AME) [119], Mean Motion Shape (MMS) [119], Motion-shape Model, modified-MHI, Silhouette History Image (SHI) [120], Silhouette Energy Image (SEI) [120], Hierarchical Motion History Histogram (HMHH) [121], Directional Motion History Image (DMHI) [122, 123], Multi-level Motion History Image (MMHI) [124], Edge MHI [125], Hierarchical Filtered Motion (HFM) [126], Landmark MHI [127], Gabor MHI [112], Enhanced-MHI [113], Local Enhanced MHI (LEMHI) [118, 128], etc. are exploited for human action recognition.

For gait recognition with the MHI/MEI, Dominant Energy Image (DEI) [129], Motion Energy Histogram (MEH) [130], Gait Moment Energy Image (GMI) [131], Moment Deviation Image (MDI) [131] are explored along with the most-widely explored approach for gait recognition called Gait Energy Image (GEI) [132]. Till-to-date, the GEI becomes the unparalleled leader for gait recognition methods. Motion Color Image (MCI) [133], Volume Motion Template (VMT) [134], Silhouette History Image (SHI), Silhouette Energy Image (SEI), etc. are exploited for gesture recognition. Motion History Volume (MHV) [135, 136] and Motion Energy Volume (MEV) are explored to detect unusual behavior for the application of video surveillance. Volumetric Motion History Image (VMHI) [137, 138] is another model similar to the VMT [134], or the MHV [136] as 3D model of the MHI template for other applications.

MHI and its variants have also seen some applications in deep learning domain. A recent work explored the MHI in deep learning [139] for gesture recognition. They fed the MHI into a 2D CNN based VGGNet, in parallel with 3D DenseNet model to recognize some gestures. Depressive symptomatology are assessed by using a variant of the MHI called Gabor MHI [112] and they explored deep learning in their method. In another approach, the MHI is used with ResNet classifier to detect the early-start intention of cyclists [140]. For emotion recognition, a Local Enhanced MHI (LEMHI) is fed into a CNN network in [118, 128]. However, in the future, the MHI or its variants can be explored more along with deep learning approaches by the researchers. Convolutional Neural Networks (CNN) are also successfully being used to generate global descriptors. Autoencoder networks [141] learns a compact representation/descriptor of the input data which is used in dimensionality reduction [142], clustering [143] etc. State-of-the-Art classifier models such as ResNet [144], InceptionNet [145], RetinaNet [146] etc. are also used in global descriptor learning [147] and has demonstrated superior performance over traditional embeddings [148].

## 5 Dimensionality Reduction Methods

Working with data in high-dimensional spaces in not always suitable due to high computational requirement and the sparsity of raw data. Dimensionality reduction is the transformation of data from a high dimensional space to low dimensional space while retaining some of its meaningful properties.

If we have data with dimensionality $D$ lying on the space $S$ that has intrinsic dimesionality $d$, where $d < D$ and often $d << D$. Here, intrinsic dimensionality means that the data is lying on or near a manifold with dimensionality $d$ that is embedded in the $D$-dimensional space without making any assumptions on the structure of this manifold. Dimensionality reduction techniques transform data with dimensionality $D$ into a new data with dimensionality $d$, while retaining as much information as possible. The problem of dimensionality reduction can then be formalized as follows:

Given sample $\{x\}_{n=1}^{N} \subset S$, find a space $T$ of dimension $d$, a dimensionality reduction mapping $\boldsymbol{F}$, and a smooth, non-singular reconstruction mapping $\boldsymbol{f}$, such that $d < D$ is as small as possible and the reconstruction error of the sample is small [149].

Unsupervised dimensionality reduction techniques can be subdivided into convex and non-convex techniques. Among the convex techniques that perform decomposition of full matrices also known as full spectral techniques, Principal Components Analysis (PCA) is the most popular. It is a linear technique for dimensionality reduction, which means that it performs dimensionality reduction by embedding the data into a linear subspace of lower dimensionality. Fig. 16 shows an application of PCA on a MNIST sample image [150]. Kernel PCA (KPCA) is the reformulation of traditional linear PCA in a high-dimensional space that is constructed using a kernel function. Unfortunately, it is unclear how the kernel function $\kappa$ should be selected. Maximum Variance Unfolding (MVU) is a technique that attempts to resolve this problem by learning the kernel matrix. Some other full spectral techniques are Diffusion Maps [151, 152], Isomaps [153], etc.

The other type of techniques optimizes a non-convex objective function. One such technique is Sammon mapping that adapts the classical PCA cost function by weighting the contribution of each pair to the cost function by the inverse of their pairwise distance in the high dimensional space. Another technique is Multilayer autoencoders that are feed-forward neural networks that are trained to minimize the mean squared error between the input and the output of the network (ideally, the input and the output are equal). Locally Linear Coordination (LLC) [154] computes a number of locally linear models and subsequently performs a global alignment of the linear models.

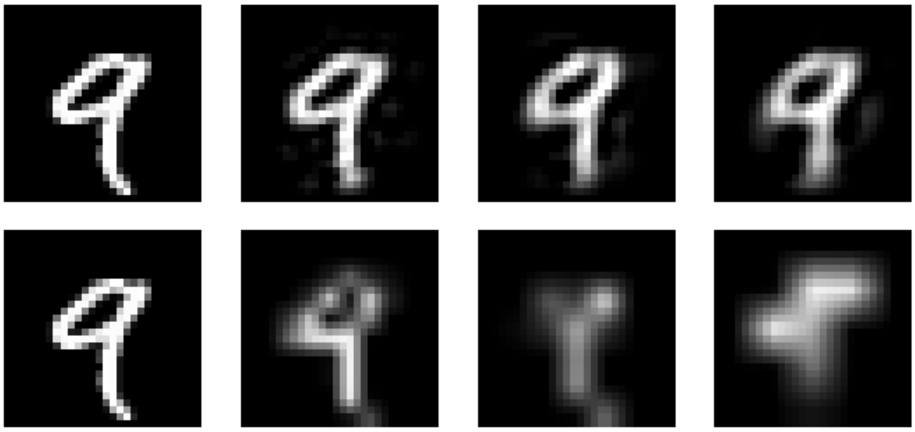

**Fig. 16.** Comparison between dimensionality reduction (top) and downsampling (bottom). Top: Original and reconstructed image after PCA compression. The original image has 784 components (left). The compression is performed by keeping 87, 43, and 26 principle components respectively. Bottom: Original and images down-sampled with same number of components.

Outside of this class of unsupervised dimensionality reduction only techniques, there are techniques that combine dimensionality reduction technique with clustering such as self-organizing maps[155] and their probabilistic extension (GTM). There are techniques of supervised nature such as Linear Discriminant Analysis (LDA) [156], Generalized Discriminant Analysis (GDA) [157], and Neighborhood Components Analysis (NCA) [158], and recently proposed metric learners [159, 160]. Techniques for Independent Component Analysis (ICA) are mainly designed for blind-source separation [161].

## 6 Conclusion

In this chapter, we shed light on the inherent yet inevitable usage of signal processing techniques in contactless, automatic human activity monitoring frameworks. Since the early days of contactless monitoring till the recent advancements and, in fact, for all future research and developments, signal processing has been and will be an integral part of any such system. The chapter covers a wide range of activity signals and associated application areas - highlighting different signal processing methods that are being utilized for different purposes. Interestingly, it is imperceptible to determine the boundaries between signal processing and machine learning, since the underlying mechanism of most,if not all, machine learning techniques are essentially signal processing approaches. Starting from describing a generic contactless monitoring pipeline, the chapter covers the basic idea, usage areas and trends of numerous signal processing methods that are closely associated with different parts of the pipeline. Hence, we believe that this chapter will provide researchers proper guidance for designing efficient contacless monitoring systems for human activities and to determine task-appropriate signal processing approaches for different components of such system.

## References

1. Carlo Massaroni, Daniel Simões Lopes, Daniela Lo Presti, Emiliano Schena, and Sergio Silvestri. Contactless monitoring of breathing patterns and respiratory rate at the pit of the neck: A single camera approach. *J. Sensors*, 2018:4567213:1–4567213:13, 2018.
2. T. Wang, D. Zhang, L. Wang, Y. Zheng, T. Gu, B. Dorizzi, and X. Zhou. Contactless respiration monitoring using ultrasound signal with off-the-shelf audio devices. *IEEE Internet of Things Journal*, 6(2):2959–2973, April 2019.
3. Ming-Chun Huang, Jason J. Liu, Wenyao Xu, Changzhan Gu, Changzhi Li, and Majid Sarrafzadeh. A self-calibrating radar sensor system for measuring vital signs. *IEEE Transactions on Biomedical Circuits and Systems*, 10:352–363, 2016.
4. Mohamad Forouzanfar, Mohamed Mabrouk, Sreeraman Rajan, Miodrag Bolic, Hilmi R. Dajani, and Voicu Groza. Event recognition for contactless activity monitoring using phase-modulated continuous wave radar. *IEEE Transactions on Biomedical Engineering*, 64:479–491, 2016.
5. Jingli Li, Son Lam Phung, F. H. C. Tivive, and A. Bouzerdoum. Automatic classification of human motions using doppler radar. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, June 2012.
6. T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan. Crowded scene analysis: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(3):367–386, March 2015.
7. Mateusz Dubiel, Martin Halvey, and Leif Azzopardi. A survey investigating usage of virtual personal assistants. *ArXiv*, abs/1807.04606, 2018.
8. A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi. Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):321–329, Jan 2006.
9. B. Clarkson and A. Pentland. Extracting context from environmental audio. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No.98EX215)*, pages 154–155, Oct 1998.
10. J. A. Stork, L. Spinello, J. Silva, and K. O. Arras. Audio-based human activity recognition using non-markovian ensemble voting. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 509–514, Sep. 2012.
11. H. Griffith, F. Hajiaghajani, and S. Biswas. Office activity classification using first-reflection ultrasonic echolocation. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4451–4454, July 2017.
12. S. Biswas, B. Harrington, F. Hajiaghajani, and R. Wang. Contact-less indoor activity analysis using first-reflection echolocation. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2016.
13. H. Griffith and S. Biswas. Home-based upper extremity rehabilitation support using a contactless ultrasonic sensor. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 853–856, July 2017.

14. G. Batchuluun, D. T. Nguyen, T. D. Pham, C. Park, and K. R. Park. Action recognition from thermal videos. *IEEE Access*, 7:103893–103917, 2019.
15. Xiaomu Luo, Qiuju Guan, Huoyuan Tan, Liwen Gao, Zhengfei Wang, and Xiaoyan Luo. Simultaneous indoor tracking and activity recognition using pyroelectric infrared sensors. *Sensors*, 17(8):1738, Jul 2017.
16. C. E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.
17. G. Oster and Y. Nishijima. Moiré patterns. *Scientific American*, 208(5):54–63, 1963.
18. Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.
19. Claude E. Duchon. Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology*, 18(8):1016–1022, 1979.
20. K. Paliwal and K. Wojcicki. Effect of analysis window duration on speech intelligibility. *IEEE Signal Processing Letters*, 15:785–788, 2008.
21. R. B. Blackman and J. W. Tukey. The measurement of power spectra from the point of view of communications engineering — part i. *The Bell System Technical Journal*, 37(1):185–282, 1958.
22. F. J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.
23. Steven W. Smith. *The Scientist and Engineer's Guide to Digital Signal Processing.* California Technical Publishing, USA, 1997.
24. Rafael C Gonzalez, Richard Eugene Woods, and Steven L Eddins. *Digital image processing using MATLAB.* Pearson Education India, 2004.
25. Liming Chen and Chris D. Nugent. *Time-Window Based Data Segmentation.* Springer International Publishing, Cham, 2019.
26. Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, and Han Hu. Parametric instance classification for unsupervised visual feature learning, 2020.
27. Z. Zhao, P. Zheng, S. Xu, and X. Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3212–3232, 2019.
28. Abdulkader Helwan and Dilber Uzun Ozsahin. Sliding window based machine learning system for the left ventricle localization in mr cardiac images. *Applied Computational Intelligence and Soft Computing*, 2017:3048181, Jun 2017.
29. H. U. Unlu, N. Patel, P. Krishnamurthy, and F. Khorrami. Sliding-window temporal attention based deep learning system for robust sensor modality fusion for ugv navigation. *IEEE Robotics and Automation Letters*, 4(4):4216–4223, 2019.
30. Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, Jul 2019.
31. Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval.* ACM, jun 2018.

32. Fokko Wieringa, Frits Mastik, and Anton F. W. van der Steen. Contactless multiple wavelength photoplethysmographic imaging: A first step toward "spo2 camera" technology. *Annals of Biomedical Engineering*, 33:1034–1041, 2005.
33. Wenjin Wang, Sander Stuijk, and Gerard de Haan. Exploiting spatial redundancy of image sensor for motion robust rppg. *IEEE Transactions on Biomedical Engineering*, 62:415–425, 2015.
34. Koen M. van der Kooij and Marnix Naber. An open-source remote heart rate imaging method with practical apparatus and algorithms. *Behavior Research Methods*, 51(5):2106–2119, Oct 2019.
35. David J. Broad. Formants in automatic speech recognition. *International Journal of Man-Machine Studies*, 4(4):411 – 424, 1972.
36. I. Yoo, H. Lim, and D. Yook. Formant-based robust voice activity detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2238–2245, 2015.
37. Qifang Zhao, Tetsuya Shimamura, Junichi Takahashi, and Jouji Suzuki. Improvement of noise robustness for formant frequency extraction based on linear predictive analysis. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, 85(9):1–14, 2002.
38. J. Allen. Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3):235–238, 1977.
39. J. Allen. Applications of the short time fourier transform to speech processing and spectral analysis. In *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, pages 1012–1015, 1982.
40. Albert Bregman. Auditory scene analysis: The perceptual organization of sound. *Journal of The Acoustical Society of America - J ACOUST SOC AMER*, 95, 01 1990.
41. John R. Deller, John G. Proakis, and John H. Hansen. *Discrete Time Processing of Speech Signals*. Prentice Hall PTR, USA, 1st edition, 1993.
42. Upal Mahbub, Shaikh Anowarul Fattah, Wei-Ping Zhu, and M. Omair Ahmad. Single-channel acoustic echo cancellation in noise based on gradient-based adaptive filtering. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):20, May 2014.
43. S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, 1979.
44. James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301, Apr 1965.
45. S. Chu, S. Narayanan, and C. . J. Kuo. Environmental sound recognition with time–frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1142–1158, 2009.
46. Wenwu Wang and Wenwu Wang. *Machine Audition: Principles, Algorithms and Systems*. IGI Global, USA, 1st edition, 2010.
47. Jonathan T. Foote. Content-based retrieval of music and audio. In C.-C. Jay Kuo, Shih-Fu Chang, and Venkat N. Gudivada, editors, *Multimedia Storage and Archiving Systems II*, volume 3229, pages 138 – 147. International Society for Optics and Photonics, SPIE, 1997.
48. Ali N. Akansu and Richard A. Haddad. *Multiresolution Signal Decomposition: Transforms, Subbands, and Wavelets*. Academic Press, Inc., USA, 1992.
49. David Taubman and Michael Marcellin. *JPEG2000 Image Compression Fundamentals, Standards and Practice*. Springer Publishing Company, Incorporated, 2013.

50. Muhammad Siddiqi, Rahman Ali, Md. Rana, Een-Kee Hong, Eun Kim, and Sungyoung Lee. Video-based human activity recognition using multilevel wavelet decomposition and stepwise linear discriminant analysis. *Sensors*, 14(4):6370–6392, Apr 2014.
51. D.K. Vishwakarma, Prachi Rawat, and Rajiv Kapoor. Human activity recognition using gabor wavelet transform and ridgelet transform. *Procedia Computer Science*, 57:630 – 636, 2015. 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015).
52. B. Y. Su, K. C. Ho, M. J. Rantz, and M. Skubic. Doppler radar fall activity detection using the wavelet transform. *IEEE Transactions on Biomedical Engineering*, 62(3):865–875, 2015.
53. Hui Yuan, Xiaolong Yang, Ailin He, Zhaoyu Li, Zhenya Zhang, and Zengshan Tian. Features extraction and analysis for device-free human activity recognition based on channel statement information in b5g wireless communications. *EURASIP Journal on Wireless Communications and Networking*, 2020(1):36, Feb 2020.
54. Jithin James and V. J. Thomas. Audio compression using dct and dwt techniques. *Journal of Information Engineering and Applications*, 4:119–124, 2014.
55. K. R. Rao and P. Yip. *Discrete Cosine Transform: Algorithms, Advantages, Applications*. Academic Press Professional, Inc., USA, 1990.
56. Nikola Božinović and Janusz Konrad. Motion analysis in 3d dct domain and its application to video coding. *Signal Processing: Image Communication*, 20(6):510 – 528, 2005. Special Issue on Advanced Aspects of Motion Estimation.
57. Paul A. Lynn. *The Laplace Transform and the z-transform*, pages 225–272. Macmillan Education UK, London, 1986.
58. Charles L Phillips, John Parr, and Eve Riskin. *Signals, Systems, and Transforms*. Prentice Hall Press, USA, 4th edition, 2007.
59. David Vernon Widder. *Laplace transform (PMS-6)*. Princeton university press, 2015.
60. Francisco Carlos M. Souza, Alinne C. Correa Souza, Carolina Y. V. Watanabe, Patricia Pupin Mandrá, and Alessandra Alaniz Macedo. An analysis of visual speech features for recognition of non-articulatory sounds using machine learning. *International Journal of Computer Applications*, 177(16):1–9, Nov 2019.
61. Z. Jiang, H. Huang, S. Yang, S. Lu, and Z. Hao. Acoustic feature comparison of mfcc and czt-based cepstrum for speech recognition. In *2009 Fifth International Conference on Natural Computation*, volume 1, pages 55–59, 2009.
62. Baris Bozkurt. *Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals*. PhD thesis, Faculté Polytechnique de Mons, University of Mons, September 2005.
63. Shlomo Engelberg. *Discrete-time Systems and the Z-transform*. Springer London, London, 2008.
64. Ali A Milani, Issa MS Panahi, and Philipos C Loizou. A new delayless subband adaptive filtering algorithm for active noise control systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5):1038–1045, 2009.
65. R Leahy, Zhenyu Zhou, and Yung-Chih Hsu. Adaptive filtering of stable processes for active attenuation of impulsive noise. In *1995 international conference on acoustics, speech, and signal processing*, volume 5, pages 2983–2986. IEEE, 1995.
66. Sascha Spors and Herbert Buchner. Efficient massive multichannel active noise control using wave-domain adaptive filtering. In *2008 3rd International Sympo-*

*sium on Communications, Control and Signal Processing*, pages 1480–1485. IEEE, 2008.
67. Upal Mahbub, Shaikh Anowarul Fattah, Wei-Ping Zhu, and M Omair Ahmad. Single-channel acoustic echo cancellation in noise based on gradient-based adaptive filtering. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):20, 2014.
68. Carl-Fredrik Westin, Lars Wigström, Tomas Loock, Lars Sjöqvist, Ron Kikinis, and Hans Knutsson. Three-dimensional adaptive filtering in magnetic resonance angiography. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 14(1):63–71, 2001.
69. Raymond H Kwong and Edward W Johnston. A variable step size lms algorithm. *IEEE Transactions on signal processing*, 40(7):1633–1642, 1992.
70. Anil K Jain. *Fundamentals of digital image processing.* Englewood Cliffs, NJ: Prentice Hall,, 1989.
71. Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113, 1997.
72. Adam W Harley. An interactive node-link visualization of convolutional neural networks. In *ISVC*, pages 867–877, 2015.
73. Yann LeCun and Corinna Cortes. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/, 2010.
74. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
75. Amit Sethi. Interaction between modules in learning systems for vision applications, 2006. AAI3223715.
76. Doh-Suk Kim, Soo-Young Lee, and Rhee Man Kil. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. *IEEE Transactions on speech and audio processing*, 7(1):55–69, 1999.
77. Costas Panagiotakis and Georgios Tziritas. A speech/music discriminator based on rms and zero-crossings. *IEEE Transactions on multimedia*, 7(1):155–166, 2005.
78. Fabien Gouyon, François Pachet, Olivier Delerue, et al. On the use of zero-crossing rate for an application of classification of percussive sounds. In *Proceedings of the COST G-6 conference on Digital Audio Effects (DAFX-00), Verona, Italy*, page 26, 2000.
79. Pamela Souza, Eric Hoover, and Frederick Gallun. Application of the envelope difference index to spectrally sparse speech. *Journal of Speech, Language, and Hearing Research*, 2012.
80. Timo Matheja, Markus Buck, and Tobias Wolff. Enhanced speaker activity detection for distributed microphones by exploitation of signal power ratio patterns. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2501–2504. IEEE, 2012.
81. Jianbo Shi et al. Good features to track. In *1994 Proceedings of IEEE conference on computer vision and pattern recognition*, pages 593–600. IEEE, 1994.
82. Christopher G. Harris and Mike Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, 1988.
83. David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.
84. Shivali Choudhary, Nitish Ojha, and Vrijendra Singh. Real-time crowd behavior detection using sift feature extraction technique in video sequences. In *2017 In-*

*ternational Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 936–940. IEEE, 2017.
85. Jianxin Wu, Adebola Osuntogun, Tanzeem Choudhury, Matthai Philipose, and James M Rehg. A scalable approach to activity recognition based on object use. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007.
86. Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.
87. Xinghao Jiang, Tanfeng Sun, Bing Feng, and Chengming Jiang. A space-time surf descriptor and its application to action recognition with video words. In *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, volume 3, pages 1911–1915. IEEE, 2011.
88. Jun-Wei Hsieh, Li-Chih Chen, and Duan-Yu Chen. Symmetrical surf and its applications to vehicle detection and vehicle make and model recognition. *IEEE Transactions on intelligent transportation systems*, 15(1):6–20, 2014.
89. Jin Zhao, Sichao Zhu, and Xinming Huang. Real-time traffic sign detection using surf features on fpga. In *2013 IEEE high performance extreme computing conference (HPEC)*, pages 1–6. IEEE, 2013.
90. Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.
91. Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *European conference on computer vision*, pages 778–792. Springer, 2010.
92. Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *European conference on computer vision*, pages 430–443. Springer, 2006.
93. Ayşe Elvan Gündüz, Alptekin Temizel, and Tuğba Taşkaya Temizel. Feature detection and tracking for extraction of crowd dynamics. In *2013 21st Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2013.
94. Nicholas Rhinehart and Kris M Kitani. First-person activity forecasting with online inverse reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3696–3705, 2017.
95. R Ramya and B Sudhakara. Motion detection in moving background using orb feature matching and affine transform. *the International Journal of Innovative Technology and Research (IJITR)*, pages 162–164, 2015.
96. Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. Brisk: Binary robust invariant scalable keypoints. In *2011 International conference on computer vision*, pages 2548–2555. Ieee, 2011.
97. Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst. Freak: Fast retina keypoint. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 510–517. Ieee, 2012.
98. Pablo Fernández Alcantarilla, Adrien Bartoli, and Andrew J Davison. Kaze features. In *European Conference on Computer Vision*, pages 214–227. Springer, 2012.
99. Pablo F Alcantarilla and T Solutions. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell*, 34(7):1281–1298, 2011.

100. John M Grey and John W Gordon. Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America*, 63(5):1493–1500, 1978.
101. Emery Schubert, Joe Wolfe, and Alex Tarnopolsky. Spectral centroid and timbre in complex, multiple instrumental textures. In *Proceedings of the international conference on music perception and cognition, North Western University, Illinois*, pages 112–116. sn, 2004.
102. Ng Chee Han, Sithi V Muniandy, and Jedol Dayou. Acoustic classification of australian anurans based on hybrid spectral-entropy approach. *Applied Acoustics*, 72(9):639–645, 2011.
103. Ya-Duo Liu, Wei Li, Xiao-Qiang Li, Zhu-Rong Wang, and Rui Feng. A robust compressed-domain music fingerprinting technique based on mdct spectral entropy. *Dianzi Xuebao(Acta Electronica Sinica)*, 38(5):1172–1176, 2010.
104. Karlheinz Brandenburg, Jürgen Herre, James D Johnston, Yannick Mahieux, and Ernst F Schroeder. Aspec-adaptive spectral entropy coding of high quality music signals. In *Audio Engineering Society Convention 90*. Audio Engineering Society, 1991.
105. Antonio Camarena-Ibarrola, Edgar Chávez, and Eric Sadit Tellez. Robust radio broadcast monitoring using a multi-band spectral entropy signature. In *Iberoamerican Congress on Pattern Recognition*, pages 587–594. Springer, 2009.
106. Simon Dixon. Onset detection revisited. In *Proceedings of the 9th International Conference on Digital Audio Effects*, volume 120, pages 133–137. Citeseer, 2006.
107. Steven W Smith et al. *The scientist and engineer's guide to digital signal processing*. California Technical Pub. San Diego, 1997.
108. Md Atiqur Rahman Ahad. *Motion History Images for Action Recognition and Understanding*. Springer, 2013.
109. Md Atiqur Rahman Ahad. *Computer vision and action recognition*. Atlantic Press, Amsterdam, available in Springer, 2011.
110. A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.
111. Md Atiqur Rahman Ahad, Joo Kooi Tan, Hyoungseop Kim, and Seiji Ishikawa. Motion history image: its variants and applications. *Machine Vision and Applications*, 23:255–281, 2012.
112. Anastasia Pampouchidou, Matthew Pediaditis, Anna Maridaki, Muhammad Awais, Calliope-Marina Vazakopoulou, Stelios Sfakianakis, Manolis Tsiknakis, Panagiotis Simos, Kostas Marias, Fan Yang, and Fabrice Meriaudeau. Quantitative comparison of motion history image variants for video-based depression assessment. *EURASIP Journal on Image and Video Processing*, 2017(1):64, Sep 2017.
113. Suad Albawendi, Kofi Appiah, Heather Powell, and Ahmad Lotfi. Video based fall detection with enhanced motion history images. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '16, New York, NY, USA, 2016. Association for Computing Machinery.
114. Mohammad Zaheer Ansari and Asad Mujeeb. Application of motion history image (mhi) on dynamic fluorescent imaging for monitoring cerebral ischemia induced by occlusion of middle cerebral artery (mca) in mouse brain. *Biomedical spectroscopy and imaging*, 6:135–142, 2017.

115. Shou-Jen Lin, Mei-Hsuan Chao, Chao-Yang Lee, and Chu-Sing Yang. Human action recognition using motion history image based temporal segmentation. *Int. J. Pattern Recognit. Artif. Intell.*, 30:1655017:1–1655017:31, 2016.
116. Zheng Chang, Xiaojuan Ban, Qing Shen, and Jing Guo. Research on three-dimensional motion history image model and extreme learning machine for human body movement trajectory recognition. *Mathematical Problems in Engineering*, 2015:528190, May 2015.
117. Mohammad Zaheer Ansari, Evelio E. Ramírez-Miquet, Isabel Otero, Dania Rodríguez, and Juan G. Darias. Real time and online dynamic speckle assessment of growing bacteria using the method of motion history image. *Journal of Biomedical Optics*, 21(6):1 – 6, 2016.
118. Haowen Wang, Guoxiang Zhou, Min Hu, and Xiaohua Wang. Video emotion recognition using local enhanced motion history image and cnn-rnn networks. In Jie Zhou, Yunhong Wang, Zhenan Sun, Zhenhong Jia, Jianjiang Feng, Shiguang Shan, Kurban Ubul, and Zhenhua Guo, editors, *Biometric Recognition*, pages 109–119, Cham, 2018. Springer International Publishing.
119. Liang Wang and D. Suter. Informative shape representations for human action recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 2, pages 1266–1269, 2006.
120. Mohiuddin Ahmad, Irine Parvin, and Seong-Whan Lee. Silhouette history and energy image information for human movement recognition. *Journal of Multimedia*, 5:12–21, 2010.
121. H. Meng, N. Pears, and C. Bailey. A human action recognition system for embedded computer vision application. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2007.
122. Md. Atiqur Rahman Ahad, J. K. Tan, H. S. Kim, and S. Ishikawa. Temporal motion recognition and segmentation approach. *International Journal of Imaging Systems and Technology*, 19(2):91–99, 2009.
123. Md Atiqur Rahman Ahad, Joo Kooi Tan, Hyoungseop Kim, and Seiji Ishikawa. Analysis of motion self-occlusion problem due to motion overwriting for human activity recognition. *Journal of Multimedia*, 5:36–46, 2010.
124. Maja Pantic, Ioannis Patras, and Michel F. Valstar. Learning spatio-temporal models of facial expressions. In *International conference on measuring behaviour*, pages 7–10, 2005.
125. Md Atiqur Rahman Ahad, Joo Kooi Tan, Hyoungseop Kim, and Seiji Ishikawa. Approaches for global-based action representations for games and action understanding. *Face and Gesture 2011*, pages 753–758, 2011.
126. Y. Tian, L. Cao, Z. Liu, and Z. Zhang. Hierarchical filtered motion for action recognition in crowded videos. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(3):313–323, 2012.
127. Anastasia Pampouchidou, Olympia Simantiraki, Amir Fazlollahi, Matthew Pediaditis, Dimitris Manousos, Alexandros Roniotis, Georgios Giannakakis, Fabrice Meriaudeau, Panagiotis Simos, Kostas Marias, Fan Yang, and Manolis Tsiknakis. Depression assessment by fusing high and low level features from audio, video, and text. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, page 27–34, New York, NY, USA, 2016. Association for Computing Machinery.
128. Min Hu, Haowen Wang, Xiaohua Wang, Juan Yang, and Ronggui Wang. Video facial emotion recognition based on local enhanced motion history image and cnn-ctslstm networks. *Journal of Visual Communication and Image Representation*, 59:176 – 185, 2019.

129. Changhong Chen, Jimin Liang, Heng Zhao, Haihong Hu, and Jie Tian. Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognit. Lett.*, 30:977–984, 2009.
130. Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 441–444, 2006.
131. Q. Ma, S. Wang, D. Nie, and J. Qiu. Recognizing humans based on gait moment image. In *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007)*, volume 2, pages 606–610, 2007.
132. J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2006.
133. R. Singh, B. Seth, and U. Desai. A real-time framework for vision based human robot interaction. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5831–5836, 2006.
134. Myung-Cheol Roh, Ho-Keun Shin, Sang-Woong Lee, and Seong-Whan Lee. Volume motion template for view-invariant gesture recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 2, pages 1229–1232, 2006.
135. C. Canton-Ferrer, J. R. Casas, and M. Pardàs. Human model and motion based 3d action recognition in multiple view scenarios. In *2006 14th European Signal Processing Conference*, pages 1–5, 2006.
136. István Petrás, Csaba Beleznai, Yigithan Dedeoglu, Montse Pardàs, Lisa Kovacs, Zoltán Szlávik, László Havasi, Tamás Szirányi, B. Ugur Töreyin, Ugur Güdükbay, A. Enis Çetin, and Cristian Canton-Ferrer. Flexible test-bed for unusual behavior detection. In *ACM Conference on Image and Videwo Retreieval (CIVR)*, pages 105–108, 2007.
137. A. B. Albu, T. Beugeling, N. Virji-Babul, and C. Beach. Analysis of irregularities in human actions with volumetric motion history images. In *2007 IEEE Workshop on Motion and Video Computing (WMVC'07)*, pages 16–16, 2007.
138. Alexandra Branzan Albu and Trevor Beugeling. A three-dimensional spatiotemporal template for interactive human motion analysis. *Journal of Multimedia*, 2(4):45–54, 2007.
139. Erhu Zhang, Botao Xue, Fangzhou Cao, Jinghong Duan, Guangfeng Lin, and Yifei Lei. Fusion of 2d cnn and 3d densenet for dynamic gesture recognition. *Electronics*, 8(12):1511, 2019.
140. Stefan Zernetsch, Viktor Kress, Bernhard Sick, and Konrad Doll. Early start intention detection of cyclists using motion history images and a deep residual network. *CoRR*, abs/1803.02242, 2018.
141. Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991.
142. Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.
143. Xifeng Guo, Xinwang Liu, En Zhu, and Jianping Yin. Deep clustering with convolutional autoencoders. In *International conference on neural information processing*, pages 373–382. Springer, 2017.
144. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

145. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
146. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
147. Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. *arXiv preprint arXiv:1811.12649*, 2018.
148. Andrew Zhai, Hao-Yu Wu, Eric Tzeng, Dong Huk Park, and Charles Rosenberg. Learning a unified embedding for visual search at pinterest. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2412–2420, 2019.
149. Miguel Angel Carreira-Perpinan. *Continuous latent variable models for dimensionality reduction and sequential data reconstruction.* PhD thesis, Citeseer, 2001.
150. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
151. Stephane Lafon and Ann B Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE transactions on pattern analysis and machine intelligence*, 28(9):1393–1403, 2006.
152. Boaz Nadler, Stéphane Lafon, Ronald R Coifman, and Ioannis G Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.
153. Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.
154. Yee W Teh and Sam T Roweis. Automatic alignment of local representations. In *Advances in neural information processing systems*, pages 865–872, 2003.
155. Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.
156. Yuting Su, Yang Li, and Anan Liu. Open-view human action recognition based on linear discriminant analysis. *Multimedia Tools and Applications*, 78:767–782, 2018.
157. G. Baudat and Fatiha Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12:2385–2404, 2000.
158. Hany Ferdinando and Esko Alasaarela. Enhancement of emotion recogniton using feature fusion and the neighborhood components analysis. In *ICPRAM*, 2018.
159. Du Tran and Alexander Sorokin. Human activity recognition with metric learning. In *ECCV*, 2008.
160. Mubarak G. Abdu-Aguye and Walid Gomaa. Robust human activity recognition based on deep metric learning. In *ICINCO*, 2019.
161. P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation.* Academic Press, Oxford, 2010.