Human Behavior Analysis

- Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O., Machine recognition of human activities: A survey, *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11), 2008. (**required**)
- Vinciarelli, A., Pantic, M., Bourlard, H., Social signal processing: Survey of an emerging domain, *Image and Vision Computing*, 27(12), 1743-1759, 2009. (**optional**)
- Laptev, I., On space-time interest points, *International Journal of Computer Vision*, 64:2, 107-123, 2005. (**optional**)

Machine Recognition of Human Activities: A Survey

Pavan Turaga, *Student Member, IEEE*, Rama Chellappa, *Fellow, IEEE*, V. S. Subrahmanian, and Octavian Udrea

***Abstract*—The past decade has witnessed a rapid proliferation of video cameras in all walks of life and has resulted in a tremendous explosion of video content. Several applications such as content-based video annotation and retrieval, highlight extraction and video summarization require recognition of the activities occurring in the video. The analysis of human activities in videos is an area with increasingly important consequences from security and surveillance to entertainment and personal archiving. Several challenges at various levels of processing—robustness against errors in low-level processing, view and rate-invariant representations at midlevel processing and semantic representation of human activities at higher level processing—make this problem hard to solve. In this review paper, we present a comprehensive survey of efforts in the past couple of decades to address the problems of representation, recognition, and learning of human activities from video and related applications. We discuss the problem at two major levels of complexity: 1) "actions" and 2) "activities." "Actions" are characterized by simple motion patterns typically executed by a single human. "Activities" are more complex and involve coordinated actions among a small number of humans. We will discuss several approaches and classify them according to their ability to handle varying degrees of complexity as interpreted above. We begin with a discussion of approaches to model the simplest of action classes known as atomic or primitive actions that do not require sophisticated dynamical modeling. Then, methods to model actions with more complex dynamics are discussed. The discussion then leads naturally to methods for higher level representation of complex activities.**

***Index Terms*—Human activity analysis, image sequence analysis, machine vision, surveillance.**

I. Introduction

RECOGNIZING human activities from video is one of the most promising applications of computer vision. In recent years, this problem has caught the attention of researchers from industry, academia, security agencies, consumer agencies, and the general populace as well. One of the earliest investigations into the nature of human motion was conducted by the contemporary photographers E. J. Marey and E. Muybridge in the 1850s who photographed moving subjects and revealed several interesting and artistic aspects involved in human and animal locomotion. The classic moving light display (MLD) experiment of Johansson [1] provided a great impetus to the study and analysis of human motion perception in the field of neuroscience. This then paved the way for mathematical modeling of human action and automatic recognition, which naturally fall into the purview of computer vision and pattern recognition.

To state the problem in simple terms, given a sequence of images with one or more persons performing an activity, can a system be designed that can automatically recognize what activity is being or was performed? As simple as the question seems, the solution has been that much harder to find. In this survey paper, we review the major approaches that have been pursued over the last 20 years to address this problem.

Several related survey papers have appeared over the years. Most notable among them are the following. Aggarwal and Cai [2] discuss three important subproblems that together form a complete action recognition system—extraction of human body structure from images, tracking across frames, and action recognition. Cedras and Shah [3] present a survey on motion-based approaches to recognition as opposed to structure-based approaches. They argue that motion is a more important cue for action recognition than the structure of the human body. Gavrila [4] presented a survey focused mainly on tracking of hands and humans via 2-D or 3-D models and a discussion of action recognition techniques. More recently, Moeslund *et al.* [5] presented a survey of problems and approaches in human motion capture including human model initialization, tracking, pose estimation, and activity recognition. Since the mid 1990s, interest has shifted more toward recognizing actions from tracked motion or structure features and on recognizing complex activities in real-world settings. Hence, this survey will focus exclusively on approaches for recognition of action and activities from video and not on the lower level modules of detection and tracking, which is discussed at length in earlier surveys [2]–[6].

The terms "action" and "activity" are frequently used interchangeably in the vision literature. In the ensuing discussion, by "actions" we refer to simple motion patterns usually executed by a single person and typically lasting for short durations of time, on the order of tens of seconds. Examples of actions include bending, walking, swimming, etc. (e.g., Fig. 1). On the other hand, by "activities" we refer to the complex sequence of actions performed by several humans who could be interacting with each other in a constrained manner. They are typically characterized by much longer temporal durations, e.g., two persons shaking hands, a football team scoring a goal, or a coordinated bank attack by multiple robbers (Fig. 2). This is not a hard boundary and there is a significant "gray area" between these two extremes. For example, the gestures of a music conductor conducting an orchestra or the constrained dynamics of a group of humans (Fig. 3) is neither as simple as an "action" nor as complex as an "activity" according to the above interpretation. However, this simple categorization provides a starting point to organize the numerous approaches that have been pro-

Manuscript received February 25, 2008; revised June 19, 2008. First published September 26, 2008; current version published October 29, 2008. This work was supported in part by the U.S. Government VACE program. This paper was recommended by Associate Editor D. Xu

The authors are with the Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 USA (e-mail: pturaga@umiacs.umd.edu; rama@umiacs.umd.edu; vs@umiacs.umd.edu; udrea@umiacs.umd.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2008.2005594

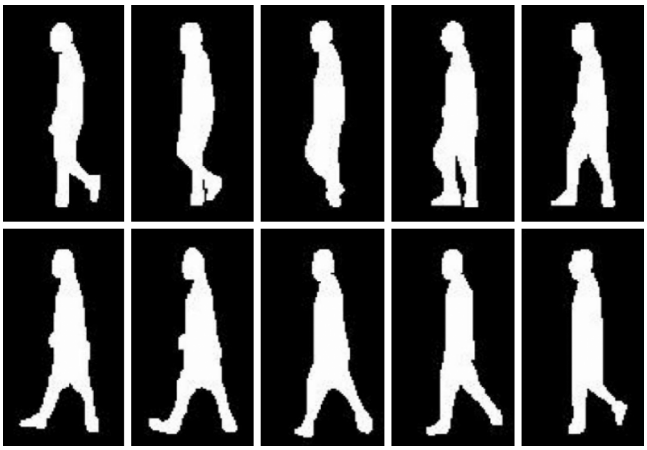


Fig. 1. Near-field video: Example of walking action. Figure taken from [7].

posed to solve the problem. A quick preview of the various approaches that fall under each of these categories is shown in Fig. 4. Real-life activity recognition systems typically follow a hierarchical approach. At the lower levels are modules such as background–foreground segmentation, tracking and object detection. At the midlevel are action–recognition modules. At the high level are the reasoning engines that encode the activity semantics based on the lower level action primitives. Thus, it is necessary to gain an understanding of both these problem domains to enable real-life deployment of systems.

The rest of this paper is organized as follows. First, we discuss a few motivating application domains in Section II. Section III provides an overview of methods for extraction of low-level image features. In Section IV, we discuss approaches for recognizing "actions." Then, in Section V, we discuss methods to represent and recognize higher level "activities." In Section VI, we discuss some open research issues for action and activity recognition and provide concluding remarks.

II. Applications

In this section, we present a few application areas that will highlight the potential impact of vision-based activity recognition systems.

1) Behavioral Biometrics: Biometrics involves study of approaches and algorithms for uniquely recognizing humans based on physical or behavioral cues. Traditional approaches are based on fingerprint, face, or iris and can be classified as physiological biometrics—i.e., they rely on physical attributes for recognition. These methods require cooperation from the subject for collection of the biometric. Recently, "behavioral biometrics" have been gaining popularity, where the premise is that behavior is as useful a cue to recognize humans as their physical attributes. The advantage of this approach is that subject cooperation is not necessary and it can proceed without interrupting or interfering with the subject's activity. Since observing behavior implies longer term observation of the subject, approaches for action recognition extend naturally to this task. Currently, the most promising example of behavioral biometrics is human gait [10].

2) Content-Based Video Analysis: Video has become a part of our everyday life. With video sharing websites experiencing relentless growth, it has become necessary to develop efficient indexing and storage schemes to improve user experience. This requires learning of patterns from raw video and summarizing a video based on its content. Content-based video summarization has been gaining renewed interest with corresponding advances in content-based image retrieval (CBIR) [11]. Summarization and retrieval of consumer content such as sports videos is one of the most commercially viable applications of this technology [12].

3) Security and Surveillance: Security and surveillance systems have traditionally relied on a network of video cameras monitored by a human operator who needs to be aware of the activity in the camera's field of view. With recent growth in the number of cameras and deployments, the efficiency and accuracy of human operators has been stretched. Hence, security agencies are seeking vision-based solutions to these tasks that can replace or assist a human operator. Automatic recognition of anomalies in a camera's field of view is one such problem that has attracted attention from vision researchers (cf., [9] and [13]). A related application involves searching for an activity of interest in a large database by learning patterns of activity from long videos [14], [15].

4) Interactive Applications and Environments: Understanding the interaction between a computer and a human remains one of the enduring challenges in designing human–computer interfaces. Visual cues are the most important mode of nonverbal communication. Effective utilization of this mode such as gestures and activity holds the promise of helping in creating computers that can better interact with humans. Similarly, interactive environments such as smart rooms [16] that can react to a user's gestures can benefit from vision-based methods. However, such technologies are still not mature enough to stand the "turing test" and thus continue to attract research interest.

5) Animation and Synthesis: The gaming and animation industry rely on synthesizing realistic humans and human motion. Motion synthesis finds wide use in the gaming industry where the requirement is to produce a large variety of motions with some compromise on the quality. The movie industry on the other hand has traditionally relied more on human animators to provide high-quality animation. However, this trend is fast changing [17]. With improvements in algorithms and hardware, much more realistic motion synthesis is now possible. A related application is learning in simulated environments. Examples of this include training of military soldiers, firefighters, and other rescue personnel in hazardous situations with simulated subjects.

III. General Overview

A generic action or activity recognition system can be viewed as proceeding from a sequence of images to a higher level interpretation in a series of steps. The major steps involved are the following:

1) input video or sequence of images;
2) extraction of concise low-level features;
3) midlevel action descriptions from low-level features;
4) high-level semantic interpretations from primitive actions.

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

Fig. 2. Medium-field video: Example video sequence of a simulated bank attack (courtesy [8]). (a) Person enters the bank. (b) Robber is identified to be an outsider. Robber is entering the bank safe. (c) A customer escapes. (d) Robber makes an exit.

Fig. 3. Far-field video: Modeling dynamics of groups of humans as a deforming shape. Figure taken from [9].

In this section, we will briefly discuss some relevant aspects of item 2, i.e., low-level feature extraction. Items 3 and 4 in the list will form the subject of discussion of Sections IV and V, respectively.

Videos consist of massive amounts of raw information in the form of spatio–temporal pixel intensity variations. However, most of this information is not directly relevant to the task of understanding and identifying the activity occurring in the video. A classic experiment by Johansson [1] demonstrated that humans can perceive gait patterns from point light sources placed at a few limb joints with no additional information. Extraneous factors such as the color of the clothes, illumination conditions, background clutter do not aid in the recognition task. We briefly describe a few popular low-level features and refer the readers to other sources for a more in-depth treatment as we progress.

A. Optical Flow

Optical flow is defined as the apparent motion of individual pixels on the image plane. Optical flow often serves as a good approximation of the true physical motion projected onto the image plane. Most methods to compute optical flow assume that the color/intensity of a pixel is invariant under the displacement from one video frame to the next. We refer the reader to [18] for a comprehensive survey and comparison of optical flow computation techniques. Optical flow provides a concise description of both the regions of the image undergoing motion and the velocity of motion. In practice, computation of optical flow is susceptible to noise and illumination changes. Applications include [19], which used optical flow to detect and track vehicles in an automated traffic surveillance application.

B. Point Trajectories

Trajectories of moving objects have popularly been used as features to infer the activity of the object (see Fig. 5). The image-plane trajectory itself is not very useful as it is sensitive to translations, rotations, and scale changes. Alternative representations such as trajectory velocities, trajectory speeds, spatio–temporal curvature, relative motion, etc., have been proposed that are invariant to some of these variabilities. A good survey of these approaches can be found in [3]. Extracting unambiguous point trajectories from video is complicated by several factors such as occlusions, noise, and background clutter. Accurate tracking algorithms need to be employed for obtaining motion trajectories [6].

C. Background Subtracted Blobs and Shape

Background subtraction is a popular method to isolate the moving parts of a scene by segmenting it into background and foreground (cf., [21]). As an example, from the sequence of background subtracted images shown in Fig. 1, the human's walking action can be easily perceived. The shape of the human silhouette plays a very important role in recognizing human actions, and it can be extracted from background subtraction blobs (see Fig. 6). Several methods based on global, boundary, and skeletal descriptors have been proposed to quantify shape. Global methods such as moments [22] consider the entire shape region to compute the shape descriptor. Boundary methods on the other hand consider only the shape contour as the defining characteristic of the shape. Such methods include chain codes [23] and landmark-based shape descriptors [24]. Skeletal methods represent a complex shape as a set of 1-D skeletal curves, for example, the medial axis transform [25]. Applications include shape-based dynamic modeling of the human silhouette as in [26] to perform gait recognition.

D. Filter Responses

There are several other features that can be broadly classified as based on spatio–temporal filter responses. In their work, Zhong *et al.* [13] process a video sequence using a spatial Gaussian and a derivative of Gaussian on the temporal axis. Due to the derivative operation on the temporal axis, the filter shows high responses at regions of motion. This response was then thresholded to yield a binary motion mask followed by aggregation into spatial histogram bins. Such a feature encodes motion and its corresponding spatial information compactly and is useful for far-field and medium-field surveillance videos. The notion of scale-space filtering has also been extended to videos by several researchers. Laptev *et al.* [27] propose a generalization of the Harris corner detector to videos using a set of spatio–temporal Gaussian derivative filters. Similarly, Dollar

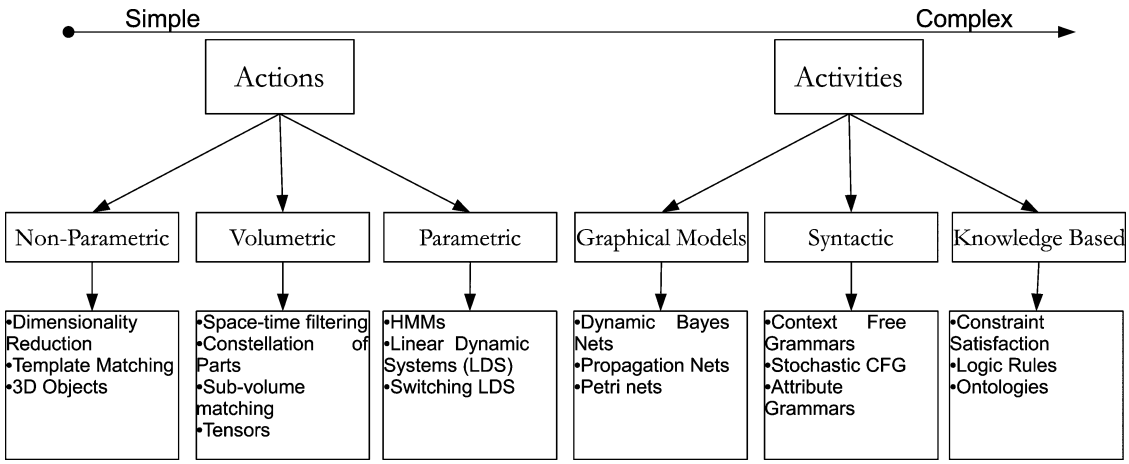


Fig. 4. Overview of approaches for action and activity recognition.

Fig. 5. Trajectories of a passenger and luggage cart. The wide difference in the trajectories is indicative of the difference in activities. Figure taken from [20].

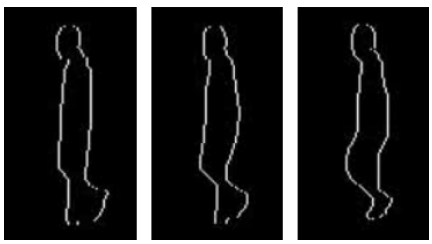


Fig. 6. Silhouettes extracted from the walking sequence shown in Fig. 1. Silhouettes encode sufficient information to recognize actions. Figure taken from [7].

et al. [28] extract distinctive periodic motion-based landmarks in a given video using a Gaussian kernel in space and a Gabor function in time. Because these approaches are based on simple convolution operations, they are fast and easy to implement. They are quite useful in scenarios with low-resolution or poor quality video where it is difficult to extract other features such as optical flow or silhouettes.

IV. Modeling and Recognizing Actions

Approaches for modeling actions can be categorized into three major classes—nonparametric, volumetric, and parametric time-series approaches. Nonparametric approaches typically extract a set of features from each frame of the video. The features are then matched to a stored template. Volumetric approaches on the other hand do not extract features on a frame-by-frame basis. Instead, they consider a video as a 3-D volume of pixel intensities and extend standard image features such as scale-space extrema, spatial filter responses, etc., to the 3-D case. Parametric time-series approaches specifically impose a model on the temporal dynamics of the motion. The particular parameters for a class of actions is then estimated from training data. Examples of parametric approaches include hidden Markov models (HMMs), linear dynamical systems (LDSs), etc. We will first discuss the nonparametric methods, then the volumetric approaches, and finally the parametric time-series methods.

A. Nonparametric Approaches for Action Recognition

1) 2-D Templates: One of the earliest attempts at action recognition without relying on 3-D structure estimation was proposed by Polana and Nelson [29]. First, they perform motion detection and tracking of humans in the scene. After tracking, a "cropped" sequence containing the human is constructed. Scale changes are compensated for by normalizing the size of the human. A periodicity index is computed for the given action and the algorithm proceeds to recognize the action if it is found to be sufficiently periodic. To perform recognition, the periodic sequence is segmented into individual cycles using the periodicity estimate and combined to get an average cycle. The average cycle is divided into a few temporal segments and flow-based features are computed for each spatial location in each segment. The flow features in each segment are averaged into a single frame. The average-flow frames within an activity cycle form the templates for each action class.

Bobick and Davis [30] proposed "temporal templates" as models for actions. In their approach, the first step involved is background subtraction, followed by an aggregation of a sequence of background subtracted blobs into a single static image. They propose two methods of aggregation—the first method gives equal weight to all images in the sequence, which gives rise to a representation called the "motion energy image"

Fig. 7. Temporal templates similar to [30]. Left: motion energy image of a sequence of a person raising both hands. Right: motion history image of the same action.

(MEI). The second method gives decaying weights to the images in the sequence with higher weight given to new frames and low weight to older frames. This leads to a representation called the "motion history image" (MHI) (for example, see Fig. 7). The MEI and MHI together comprise a template for a given action. From the templates, translation, rotation, and scale invariant Hu moments [22] are extracted that are then used for recognition. It was shown in [30] that MEI and MHI have sufficient discriminating ability for several simple action classes such as "sitting down," "bending," "crouching," and other aerobic postures. However, it was noted in [31] that MEI and MHI lose discriminative power for complex activities due to overwriting of the motion history and hence are unreliable for matching.

2) 3-D Object Models: Successful application of models and algorithms to object recognition problems led researchers in action recognition to propose alternate representations of actions as spatio–temporal objects. Syeda-Mahmood *et al.* proposed a representation of actions as generalized cylinders in the joint (x, y, t) space [32]. Yilmaz and Shah [33] represent actions as 3-D objects induced by stacking together tracked 2-D object contours. A sequence of 2-D contours in (x, y) space can be treated as an object in the joint (x, y, t) space. This representation encodes both the shape and motion characteristics of the human. From the (x, y, t) representation, concise descriptors of the object's surface are extracted corresponding to geometric features such as peaks, pits, valleys, and ridges. Because this approach is based on stacking together a sequence of silhouettes, accurate correspondence between points of successive silhouettes in the sequences needs to be established. Quasi-view invariance for this representation was shown theoretically by assuming an affine camera model. Similar to this approach, Gorelick *et al.* [34] proposed using background subtracted blobs instead of contours, which are then stacked together to create an (x, y, t) binary space–time (ST) volume (for example, see Fig. 8). Because this approach uses background subtracted blobs, the problem of establishing correspondence between points on contours in the sequence does not exist. From this ST volume, 3-D shape descriptors are extracted by solving a Poisson equation [34]. Because these approaches require careful segmentation of background and the foreground, they are limited in applicability to fixed camera settings.

3) Manifold Learning Methods: Most approaches in action recognition involve dealing with data in very high-dimensional spaces. Hence, these approaches often suffer from the "curse of dimensionality." The feature space becomes sparser in an exponential fashion with the dimension, thus requiring a larger

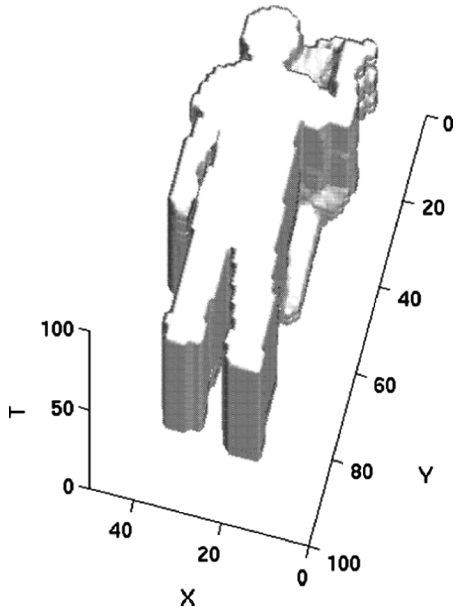


Fig. 8. The 3-D space–time object, similar to [34], obtained by stacking together binary background subtracted images of a person waving his hand.

number of samples to build efficient class-conditional models. Learning the manifold on which the data resides enables us to determine the inherent dimensionality of the data as opposed to the raw dimensionality. The inherent dimensionality contains fewer degrees of freedom and allows efficient models to be designed in the lower dimensional space. The simplest way to reduce dimensionality is via principal component analysis (PCA), which assumes that the data lies on a linear subspace. Except in very special cases, data does not lie on a linear subspace, thus requiring methods that can learn the intrinsic geometry of the manifold from a large number of samples. Nonlinear dimensionality reduction techniques allow for representation of data points based on their proximity to each other on nonlinear manifolds. Several methods for dimensionality reduction such as PCA, locally linear embedding (LLE) [35], Laplacian eigenmap [36], and Isomap [37] have been applied to reduce the high-dimensionality of video data in action–recognition tasks (cf., [38]–[40]). Specific recognition algorithms such as template matching, dynamical modeling, etc., can be performed more efficiently once the dimensionality of the data has been reduced.

B. Volumetric Approaches

1) Spatio–Temporal Filtering: These approaches are based on filtering a video volume using a large filter bank. The responses of the filter bank are further processed to derive action specific features. These approaches are inspired by the success of filter-based methods on other still image recognition tasks such as texture segmentation [41]. Further, spatio–temporal filter structures such as oriented Gaussian kernels and their derivatives [42] and oriented Gabor filter banks [43] have been hypothesized to describe the major spatio–temporal properties of cells in the visual cortex. Chomat *et al.* [44] model a segment of video as a (x, y, t) spatio–temporal volume and compute local appearance models at each pixel using a Gabor filter

bank at various orientation and spatial scales and a single temporal scale. A given action is recognized using a spatial average of the probabilities of individual pixels in a frame. Because actions are analyzed at a single temporal scale, this method is not applicable to variations in execution rate. As an extension to this approach, local histograms of normalized ST gradients at several temporal scales are extracted by Zelnik-Manor and Irani [45]. The sum of the chi-square metric between histograms is used to match an input video with a stored exemplar. Filtering with the Gaussian kernel in space and the derivative of the Gaussian on the temporal axis followed by thresholding of the responses and accumulation into spatial histograms was found to be a simple yet effective feature for actions in far-field settings [13].

Filtering approaches are fast and easy to implement due to efficient algorithms for convolution. In most applications, the appropriate bandwidth of the filters is not known *a priori*, thus a large filter bank at several spatial and temporal scales is required for effectively capturing the action dynamics. Moreover, the response generated by each filter has the same dimensions as the input volume, hence using large filter banks at several spatial and temporal scales is prohibitive.

2) Part-Based Approaches: Several approaches have been proposed that consider a video volume as a collection of local parts, where each part consists of some distinctive motion pattern. Laptev and Lindeberg [27] proposed a spatio–temporal generalization of the well-known Harris interest point detector, which is widely used in object recognition applications and applied it to modeling and recognizing actions in ST. This method is based on the 3-D generalization of scale-space representations. A given video is convolved with a 3-D Gaussian kernel at various spatial and temporal scales. Then, spatio–temporal gradients are computed at each level of the scale-space representation. These are then combined within a neighborhood of each point to yield stable estimates of the spatio–temporal second-moment matrix. Local features are then derived from these smoothed estimates of gradient moment matrices. In a similar approach, Dollar *et al.* [28] model a video sequence by the distribution of ST feature prototypes. The feature prototypes are obtained by k-means clustering of a large set of features—ST gradients—extracted at ST interest points from the training data. Neibles *et al.* [46] use a similar approach where they use a bag-of-words model to represent actions. The bag-of-words model is learned by extracting spatio–temporal interest points and clustering of the features. These interest points can be used in conjunction with machine learning approaches such as support vector machines (SVMs) [47] and graphical models [46]. Because the interest points are local in nature, longer term temporal correlations are ignored in these approaches. To address this issue, a method based on correlograms of prototype labels was presented in [48]. In a slightly different approach Nowozin *et al.* [49] consider a video as a sequence of sets, where each set consists of the parts found in a small temporally sliding window. These approaches do not directly model the global geometry of local parts instead considering them as a bag of features. Different actions may be composed of similar ST parts but may differ in their geometric relationships. Integrating global geometry into the part-based video representation was investigated by Boiman *et al.* [50] and Wong *et al.* [51]. This approach may be termed as a constellation of parts as opposed to the simpler bag-of-parts model. Computational complexity can be large for constellation models with a large number of parts, which is typically the case for human actions. Song *et al.* [52] addressed this issue by approximating the connections in the constellation via triangulation. Niebles *et al.* [53] proposed a hierarchical model where the higher level is a constellation of parts much smaller than the actual number of features. Each of the parts in the constellation consists of a bag of features at the lower level. This approach combines the advantages of both the bag of features and the constellation model and preserves computational efficiency at the same time.

In most of these approaches, the detection of the parts is usually based on linear operations such as filtering and spatio–temporal gradients, hence the descriptors are sensitive to changes in appearance, noise, occlusions, etc. It has also been noted that interest points are extremely sparse in smooth human actions and certain types of actions do not give rise to distinctive features [28], [46]. However, due to their local nature, they are more robust to nonstationary backgrounds.

3) Subvolume Matching: As opposed to part-based approaches, researchers have also investigated matching of videos by matching subvolumes between a video and a template. Shechtman *et al.* [54] present an approach derived from ST motion-based correlation to match actions with a template. The main difference of this approach from the part-based approaches is that it does not extract action descriptors from extrema in scale space, rather it looks for similarity between local ST patches based on how similar the motion is in the two patches. However, computing this correlation throughout a given video volume can be computationally intensive. Inspired by the success of Haar-type features or "box features" in object detection [55], Ke *et al.* [56] extended this framework to 3-D. In their approach, they define 3-D Haar-type features that are essentially outputs of 3-D filter banks with $+1$'s and -1's as the filter coefficients. These filter responses used in conjunction with boosting approaches result in robust performance. In another approach, Ke *et al.* [57] consider a video volume as a collection of subvolumes of arbitrary shape, where each subvolume is a spatially coherent region. The subvolumes are obtained by clustering the pixels based on appearance and spatial proximity. A given video is oversegmented into many subvolumes or "supervoxels." An action template is matched by searching among the oversegmented volumetric regions and finding the minimal set of regions that maximize overlap between their union and the template.

Subvolume matching approaches such as these are susceptible to changing backgrounds but are more robust to noise and occlusions. Another advantage is that these approaches can be extended to features such as optical flow as in [56] to achieve robustness to changes in appearance.

4) Tensor-Based Approaches: Tensors are generalizations of matrices to multiple dimensions. A 3-D ST volume can naturally be considered as a tensor with three independent dimensions. Vasilescu [58] proposed the modeling of human action, human identity, and joint angle trajectories by considering them as independent dimensions of a tensor. By decomposing the

overall data tensor into dominant modes (as a generalization of PCA), one can extract signatures corresponding to both the action and the identity of the person performing the action. Recently, Kim *et al.* [59] extended canonical correlation analysis to tensors to match videos directly to templates. In their approach, the dimensions of the tensor were simply the ST dimensions corresponding to (x, y, t). Similarly, Wolf *et al.* [60] extended low-rank SVM techniques to the space of tensors for action recognition.

Tensor-based approaches offer a direct method for holistic matching of videos without recourse to midlevel representations such as the previous ones. Moreover, they can incorporate other types of features such as optical flow, ST filter responses, etc., into the same framework by simply adding more independent dimensions to the tensor.

C. Parametric Methods

The previous section focused on representations and models that are well suited for simple actions. The parametric approaches that we will describe in this section are better suited for more complex actions that are temporally extended. Examples of such complex actions include the steps in a ballet dancing video, a juggler juggling a ball, and a music conductor conducting an orchestra using complex hand gestures.

1) Hidden Markov Models: One of the most popular state–space models is the hidden Markov model. In the discrete HMM formalism, the state space is considered to be a finite set of discrete points. The temporal evolution is modeled as a sequence of probabilistic jumps from one discrete state to the other. HMMs first found wide applicability in speech recognition applications in the early 1980s. An excellent source for a detailed explanation of HMMs and its associated three problems—inference, decoding, and learning—can be found in [61]. Beginning in the early 1990s, HMMs began to find wide applicability in computer vision systems. One of the earliest approaches to recognize human actions via HMMs was proposed by Yamato *et al.* [62] where they recognized tennis shots such as backhand stroke, backhand volley, forehand stroke, forehand volley, smash, etc., by modeling a sequence of background subtracted images as outputs of class-specific HMMs. Several successful gesture recognition systems such as in [63]–[65] make extensive use of HMMs by modeling a sequence of tracked features such as hand blobs as HMM outputs.

HMMs have also found applicability in modeling the temporal evolution of human gait patterns both for action recognition and biometrics (cf., [66] and [67]). All these approaches are based on the assumption that the feature sequence being modeled is a result of a single person performing an action. Hence, they are not effective in applications where there are multiple agents performing an action or interacting with each other. To address this issue, Brand *et al.* [68] proposed a coupled HMM to represent the dynamics of interacting targets. They demonstrate the superiority of their approach over conventional HMMs in recognizing two-handed gestures. Incorporating domain knowledge into the HMM formalism has been investigated by several researchers. Moore *et al.* [69] used HMMs in conjunction with object detection modules to exploit the relationship between actions and objects. Hongeng and Nevatia [70] incorporate *a priori* beliefs of state duration into the HMM framework and the resultant model is called hidden semi-Markov model (semi-HMMs). Cuntoor and Chellappa [71] have proposed a mixed-state HMM formalism to model nonstationary activities, where the state space is augmented with a discrete label for higher level behavior modeling.

HMMs are efficient for modeling time-sequence data and are useful both for their generative and discriminative capabilities. HMMs are well suited for tasks that require recursive probabilistic estimates [63] or when accurate start and end times for action units are unknown. However, their utility is restricted due to the simplifying assumptions that the model is based on. Most significantly the assumption of Markovian dynamics and the time-invariant nature of the model restricts the applicability of HMMs to relatively simple and *stationary* temporal patterns.

2) Linear Dynamical Systems: Linear dynamical systems are a more general form of HMMs where the state space is not constrained to be a finite set of symbols but can take on continuous values in $\mathbb{R}^k$ where k is the dimensionality of the state space. The simplest form of LDS is the first-order time-invariant Gauss–Markov processes, which is described by

$$x(t) = Ax(t-1) + w(t), \qquad w \sim N(0, Q) \tag{1}$$

$$y(t) = Cx(t) + v(t), \qquad v \sim N(0, R) \tag{2}$$

where $x \in \mathbb{R}^d$ is the d-dimensional state vector and $y \in \mathbb{R}^n$ is the n-dimensional observation vector with $d \ll n$. w and v are the process and observation noise, respectively, which are Gaussian distributed with zero-means and covariance matrices Q and R, respectively. The LDS can be interpreted as a continuous state–space generalization of HMMs with a Gaussian observation model. Several applications such as recognition of humans and actions based on gait [7], [72], [73], activity recognition [9], [74], and dynamic texture modeling and recognition [75], [76] have been proposed using LDSs.

Advances in system identification theory for learning LDS model parameters from data [77]–[79] and distance metrics on the LDS space [75], [80], [81] have made LDSs popular for learning and recognition of high-dimensional time-series data. More recently, in-depth study of the LDS space has enabled the application of machine learning tools on that space such as dynamic boosting [82], kernel methods [83], [84], and statistical modeling [85]. Newer methods to learn the model parameters [86] have made learning much more efficient than in the case of HMMs. Like HMMs, LDSs are also based on assumptions of Markovian dynamics and conditionally independent observations. Thus, as in the case of HMMs, the time-invariant model is not applicable to nonstationary actions.

3) Nonlinear Dynamical Systems: While time-invariant HMMs and LDSs are efficient modeling and learning tools, they are restricted to linear and stationary dynamics. Consider the following activity: a person bends down to pick up an object, then he walks to a nearby table and places the object on the table, and finally rests on a chair. This activity is composed of a sequence of short segments each of which can be modeled as an LDS. The entire process can be seen as switching between

LDSs. The most general form of the time-varying LDS is given by

$$x(t) = A(t)x(t-1) + w(t), \qquad w \sim N(0, Q) \tag{3}$$
$$y(t) = C(t)x(t) + v(t), \qquad v \sim N(0, R) \tag{4}$$

which looks similar to the LDS in (1) and (2), except that the model parameters A and C are allowed to vary with time. To tackle such complex dynamics, a popular approach is to model the process using switching linear dynamical systems (SLDSs) or jump linear systems (JLSs). An SLDS consists of a set of LDSs with a switching function that causes model parameters to change by switching between models. Bregler [87] presented a multilayered approach to recognize complex movements consisting of several levels of abstraction. The lowest level is a sequence of input images. The next level consists of "blob" hypotheses where each blob is a region of coherent motion. At the third level, blob tracks are grouped temporally. The final level consists of an HMM for representing the complex behavior. North *et al.* [88] augment the continuous state vector with a discrete state component to form a "mixed" state. The discrete component represents a mode of motion or more generally a "switch" state. Corresponding to each switch state, a Gaussian autoregressive model is used to represent the dynamics. A maximum-likelihood approach is used to learn the model parameters for each motion class. Pavlovic and Rehg [89], [90] model the nonlinearity in human motion in a similar framework, where the dynamics are modeled using LDS and the switching process is modeled using a probabilistic finite state machine.

Though the SLDS framework has greater modeling and descriptive power than HMMs and LDSs, learning and inference in SLDS are much more complicated, often requiring approximate methods [91]. In practice, determining the appropriate number of switching states is challenging and often requires large amounts of training data or extensive hand tuning.

V. Modeling and Recognizing Activities

Most activities of interest in applications such as surveillance and content-based indexing involve several actors, who interact not only with each other, but also with contextual entities. The approaches discussed so far are mostly concerned with modeling and recognizing actions of a single actor. Modeling a complex scene, the inherent structure and semantics of complex activities require higher level representation and reasoning methods.

A. Graphical Models

1) Belief Networks: A Bayesian network (BN) [92] is a graphical model that encodes complex conditional dependencies between a set of random variables that are encoded as local conditional probability densities (CPD). Dynamic belief networks (DBNs) are a generalization of the simpler BNs by incorporating temporal dependencies between random variables. DBNs encode more complex conditional dependence relations among several random variables as opposed to just one hidden variable as in a traditional HMM.

Huang *et al.* [19] used DBNs for vision-based traffic monitoring. Buxton and Gong [93] used BNs to capture the dependencies between scene layout and low-level image measurements for a traffic surveillance application. Remagnino *et al.* [94] present an approach using DBNs for scene description at two levels of abstraction—agent level descriptions and interagent interactions. Modeling two-person interactions such as pointing, punching, pushing, hugging, etc., was proposed by Park and Aggarwal [95] in a two-stage process. First, pose estimation is done via a BN and temporal evolution of pose is modeled by a DBN. Intille and Bobick [96] use BNs for multiagent interactions where the network structure is automatically generated from the temporal structure provided by a user. Usually the structure of the DBN is provided by a domain expert. However, this is difficult in real-life systems where there are a very large number of variables with complex interdependencies. To address this issue, Gong *et al.* [97] presented a DBN framework where the structure of the network is discovered automatically using Bayesian information criterion [98], [99].

DBNs have also been used to recognize actions using the contextual information of the objects involved. Moore *et al.* [69] conduct action recognition using belief networks based on scene context derived from other objects in the scene. Gupta *et al.* [100] present a BN for interpretation of human–object interactions that integrates information from perceptual tasks such as human motion analysis, manipulable object detection, and "object reaction" determination.

Though DBNs are more general than HMMs by considering dependencies between several random variables, the temporal model is usually Markovian as in the case of HMMs. Thus, only sequential activities can be handled by the basic DBN model. Development of efficient algorithms for learning and inference in graphical models (cf., [101]) have made them popular tools to model structured activities. Methods to learn the topology or structure of BNs from data [102] have also been investigated in the machine learning community. However, to learn the local CPDs for large networks requires very large amounts of training data or extensive hand-tuning by experts both of which limit the applicability of DBNs in large scale settings.

2) Petri Nets: Petri nets were defined by Petri [103] as a mathematical tool for describing relations between conditions and events. Petri nets are particularly useful to model and visualize behaviors such as sequencing, concurrency, synchronization, and resource sharing [104], [105]. Petri nets are bipartite graphs consisting of two types of nodes—places and transitions. Places refer to the state of an entity and transitions refer to changes in the state of the entity. Consider an example of a car pickup activity represented by a probabilistic Petri net as shown in Fig. 9. In this figure, the places are labeled $p_1, \ldots, p_5$ and transitions $t_1, \ldots, t_6$. In this PN, p_1 and p_3 are the start nodes and p_5 is the terminal node. When a car enters the scene, a "token" is placed in place p_1. The transition t_1 is enabled in this state, but it cannot fire until the condition associated with it is satisfied, i.e., when the car stops near a parking slot. When this occurs, the token is removed from p_1 and placed in p_2. Similarly, when a person enters the parking lot, a token is placed in p_3 and transition t_5 fires after the person disappears near the parked car. The token is then removed from p_3 and placed in p_4.

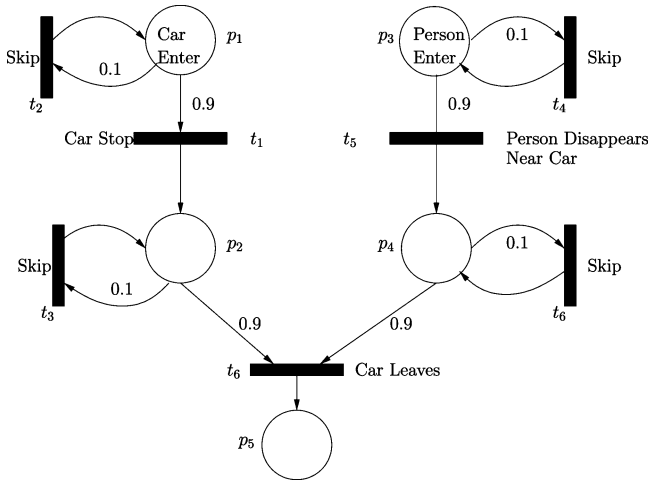


Fig. 9. Probabilistic Petri net representing a pickup-by-car activity. Figure taken from [108].

Now with a token in each of the enabling places of transition t_6, it is ready to fire when the associated condition, i.e., car leaving the parking lot is satisfied. Once the car leaves, t_6 fires and both tokens are removed and a token placed in the final place p_5. This example illustrates sequencing, concurrency, and synchronization.

Petri nets were used by Castel *et al.* [106] to develop a system for high-level interpretation of image sequences. In their approach, the structure of the Petri net was specified *a priori*. This can be tedious for large networks representing complex activities. Ghanem *et al.* [107] proposed a method to semiautomate this task by automatically mapping a small set of logical, spatial, and temporal operators to the graph structure. Using this method, they developed an interactive tool for querying surveillance videos by mapping user queries to Petri nets. However, these approaches were based on deterministic Petri nets. Hence, they cannot deal with uncertainty in the low-level modules as is usually the case with trackers, object detectors, etc. Further, real-life human activities do not conform to hard-coded models—the models need to allow deviations from the expected sequence of steps while penalizing significant deviations. To address this issue, Albanese *et al.* [108] proposed the concept of a probabilistic Petri net (PPN) (see Fig. 9). In a PPN, the transitions are associated with a weight that encodes the probability with which that transition fires. By using skip transitions and penalizing them with low probability, robustness is achieved to missing observations in the input stream. Further, the uncertainty in the identity of an object or the uncertainty in the unfolding of an activity can be efficiently incorporated into the tokens of the Petri net.

Though Petri nets are an intuitive tool for expressing complex activities, they suffer from the disadvantage of having to manually describe the model structure. The problem of learning the structure from training data has not yet been formally addressed.

3) Other Graphical Models: Other graphical models have been proposed to deal with the drawbacks in DBNs—most significantly, the limitation to sequential activities. Graphical models that specifically model more complex temporal relations such as sequentiality, duration, parallelism, synchrony, etc., have been proposed in the DBN framework. Examples include the work of Pinhanez and Bobick [109] who use a simplified version of Allen's interval algebra to model sophisticated temporal ordering constraints such as past, now, and future. This structure is termed the past–now–future (PNF) network. Similarly, Shi *et al.* [110], [111] have proposed using propagation nets to represent activities using partially ordered temporal intervals. In their approach, an activity is constrained by temporal and logical ordering and duration of the activity intervals. More recently, Hamid *et al.* [112] considered a temporally extended activity as a sequence of event labels. Due to contextual and activity specific constraints, the sequence labels are observed to have some inherent partial ordering. For example, in a kitchen setting, the refrigerator would have to be opened before the eggs can be accessed. Using these constraints, they consider an activity model as a set of subsequences, which encode the partial ordering constraints of varying lengths. These subsequences are efficiently represented using Suffix trees. The advantage of the Suffix-tree representation is that the structure of the activity can be learned from training data using standard graph-theoretic methods.

B. Syntactic Approaches

1) Grammars: Grammars express the structure of a process using a set of production rules. To draw a parallel to grammars in language modeling, production rules specify how sentences (activities) can be constructed from words (activity primitives), and how to recognize if a sentence (video) conforms to the rules of a given grammar (activity model). One of the earliest use of grammars for visual activity recognition was proposed by Brand [113], who used a grammar to recognize hand manipulations in sequences containing disassembly tasks. He made use of simple grammars with no probabilistic modeling. Ryoo and Aggarwal [114] used the context-free grammar (CFG) formalism to model and recognize composite human activities and multiperson interactions. They followed a hierarchical approach where the lower levels are composed of HMMs and BNs. The higher level interactions are modeled by CFGs. Context-free grammar approaches present a sound theoretical basis for modeling structured processes. In syntactic approaches, one only needs to enumerate the list of primitive events that need to be detected and the set of production rules that define higher level activities of interest. Once the rules of a CFG have been formulated, efficient algorithms to parse them exist [115], [116], which have made them popular in real-time applications.

Because deterministic grammars expect perfect accuracy in the lower levels, they are not suited to deal with errors in low-level tasks such as tracking errors and missing observations. In complex scenarios involving several agents requiring temporal relations that are more complex than just sequencing, such as parallelism, overlap, synchrony, it is difficult to formulate the grammatical rules manually. Learning the rules of the grammar from training data is a promising alternative, but it has proved to be extremely difficult in the general case [117].

2) Stochastic Grammars: Algorithms for detection of low-level primitives are frequently probabilistic in nature. Thus, stochastic context-free grammars (SCFGs), which are a probabilistic extension of CFGs, were found to be suitable

$$S \rightarrow BOARDING_N$$
$$BOARDING \rightarrow appear_0 CHECK_1 disappear_1$$
$$(isPerson(appear.class) \wedge isInside(appear.loc, Gate) \wedge isInside(disappear.loc, Plane))$$
$$CHECK \rightarrow moveclose_0 CHECK_1$$
$$CHECK \rightarrow moveaway_0 CHECK_1$$
$$CHECK \rightarrow moveclose_0 moveaway_1 CHECK_1$$
$$(isPerson(moveclose.class) \wedge moveclose.idr = moveaway.idr)$$

Fig. 10. Example of an attribute grammar for a passenger boarding an airplane taken from [120].

for integration with real-life vision modules. SCFGs were used by Ivanov and Bobick [118] to model the semantics of activities whose structure was assumed to be known. They used HMMs for low-level primitive detection. The grammar production rules were augmented with probabilities and a "skip" transition was introduced. This resulted in increased robustness to insertion errors in the input stream and also to errors in low-level modules. Moore *et al.* [119] used SCFGs to model multitasked activities—activities that have several independent threads of execution with intermittent dependent interactions with each other as demonstrated in a blackjack game with several participants.

In many cases, it is desirable to associate additional attributes or features to the primitive events. For example, the exact location in which the primitive event occurs may be significant for describing an event, but this may not be effectively encoded in the (finite) primitive event set. Thus, attribute grammars achieve greater expressive power than traditional grammars. Probabilistic attribute grammars have been used by Joo and Chellappa [120] for multiagent activities in surveillance settings. In the example shown in Fig. 10, one can see the production rules and the primitive events such as "appear," "disappear," "moveclose," "moveaway," etc., in the description of the activity. The primitive events are further associated with attributes such as location (loc) where the appearance and disappearance events occur, classification (class) into a set of objects, identity (idr) of the entity involved, etc.

While SCFGs are more robust than CFGs to errors and missed detections in the input stream, they share many of the temporal relation modeling limitations of CFGs as discussed above.

C. Knowledge and Logic-Based Approaches

1) Logic-Based Approaches: Logic-based methods rely on formal logical rules to describe common sense domain knowledge to describe activities. Logical rules are useful to express domain knowledge as input by a user or to present the results of high-level reasoning in an intuitive and human-readable format. Declarative models [121] describe all expected activities in terms of scene structure, events, etc. The model for an activity consists of the interactions between the objects of the scene. Medioni *et al.* [122] propose a hierarchical representation to recognize a series of actions performed by a single agent. Symbolic descriptors of actions are extracted from low-level features through several midlevel layers. Next, a rule-based method is used to approximate the probability of occurrence of a specific activity by matching the properties of the agent with the expected distributions (represented by a mean and a variance) for a particular action. In a later work, Hongeng *et al.* [123] extended this representation by considering an activity to be composed of several action threads. Each action thread is modeled as a stochastic finite state automaton. Constraints between the various threads are propagated in a temporal logic network. Shet *et al.* [124] propose a system that relies on logic programming to represent and recognize high-level activities. Low-level modules are used to detect primitive events. The high-level reasoning engine is based on Prolog and recognizes activities, which are represented by logical rules between primitives. These approaches do not explicitly address the problem of uncertainty in the observation input stream. To address this issue, a combination of logical and probabilistic models was presented in [125], where each logical rule is represented as first-order logic formula. Each rule is further provided with a weight, where the weight indicates a belief in the accuracy of the rule. Inference is performed using a Markov-logic network.

While logic-based methods are a natural way of incorporating domain knowledge, they often involve expensive constraint satisfaction checks. Further, it is not clear how much domain knowledge should be incorporated in a given setting—incorporating more knowledge can potentially make the model rigid and nongeneralizable to other settings. Further, the logic rules require extensive enumeration by a domain expert for every deployment.

2) Ontologies: In most practical deployments that use any of the aforementioned approaches, symbolic activity definitions are constructed in an empirical manner, for example, the rules of a grammar or a set of logical rules are specified manually. Though empirical constructs are fast to design and even work very well in most cases, they are limited in their utility to specific deployments for which they have been designed. Hence, there is a need for a centralized representation of activity definitions or ontologies for activities that are independent of algorithmic choices. Ontologies standardize activity definitions, allow for easy portability to specific deployments, enable interoperability of different systems, and allow easy replication and comparison of system performance. Several researchers have proposed ontologies for specific domains of visual surveillance. For example, Chen *et al.* [126] proposed an ontology for analyzing social interaction in nursing homes, Hakeem *et al.* for classification of meeting videos [127], and Georis *et al.* [8] for activities in a bank monitoring setting. To consolidate these efforts and to build a common knowledge base of domain ontologies, the Video Event Challenge Workshop was held in 2003. As a result of this workshop, ontologies have been defined for six domains of video surveillance [128]: 1) perimeter and internal security; 2) railroad crossing surveillance; 3) visual bank monitoring; 4) visual metro monitoring; 5) store security; and 6) airport-tarmac security. An example from the ontology output is shown in Fig. 11, which describes

```
PROCESS(cruise-parking-lot(vehicle v, parking-lot lot),
Sequence(enter(v, lot),
  set-to-zero(i),
  Repeat-Until(
    AND(move-in-circuit(v), inside(v, lot), increment(i)),
    equal(i, n)),
  exit(v, lot)))
```

Fig. 11. Ontology for car cruising in parking lot activity. Example taken from [128].

car cruising activity. This ontology keeps track of the number of times the car moves around in a circuit inside the parking lot without stopping. When this exceeds a set threshold, a cruising activity is detected. The workshop also led to the development of two formal languages—the video event representation language (VERL) [129], [130], which provides an ontological representation of complex events in terms of simpler subevents, and the video event markup language (VEML), which is used to annotate VERL events in videos.

Though ontologies provide concise high-level definitions of activities, they do not necessarily suggest the right "hardware" to "parse" the ontologies for recognition tasks.

VI. Directions for Future Work and Conclusion

A lot of enthusiasm has been generated in the vision community by recent advances in machine recognition of activities. However, several important issues remain to be addressed. In this section, we briefly discuss some of these issues.

A. Real-World Conditions

Most action and activity recognition systems are currently designed and tested on video sequences acquired in constrained conditions. Factors that can severely limit the applicability in real-world conditions include noise, occlusions, shadows, etc. Errors in feature extraction can easily propagate to higher levels. For real-world deployment, action recognition systems need to be tested against such real-world conditions. Methods that are robust to these factors also need to be investigated. Many practically deployed systems do not record videos at high spatio–temporal resolution in part due to the difficulty in storing the large data that is produced. Hence, dealing with low-resolution video is an important issue. In the approaches discussed so far, it is assumed that reliable features can be extracted in a given setting such as optical flow or background subtracted blobs. In analyzing actions in far-field settings, this assumption does not usually hold. While researchers have addressed these issues in specific settings (cf., [131] and [132]), a systematic and general approach is still lacking. Hence, more research needs to be done to address these practical issues.

B. Invariances in Human Action Analysis

One of the most significant challenges in action recognition is to find methods that can explain and be robust to the wide variability in features that are observed within the same action class. Sheikh *et al.* [133] have identified three important sources that give rise to variability in observed features. They are as follows:

1) viewpoint;
2) execution rate;
3) anthropometry

Any real-world action recognition system needs to be invariant to these factors. In this section, we will review some efforts in this direction that have been pursued in the research community.

1) View Invariance: While it may be easy to build statistical models of simple actions from a single view, it is extremely challenging to generalize them to other views. This is due to the wide variations in motion and structure features induced by camera perspective effects and occlusions. One way to deal with the problem is to store templates from several canonical views as done in [30] and interpolate across the stored views as proposed by [134]. This approach, however, is not scalable because one does not know how many views to consider as canonical. Another approach is to assume that point correspondences across views are available as in [32] and compute a transformation that maps a stored model to an example from an arbitrary view. Similarly, Seitz and Dyer [135] present an approach to recognize cyclic motion that is affine invariant by assuming that feature correspondence between successive time instants is known. It was shown by Rao and Shah [136] that extrema in ST curvature of trajectories are preserved across views, which were exploited to perform view-invariant action recognition. Another example is the work of Parameswaran *et al.* [137] who define a view-invariant representation of actions based on the theory of 2-D and 3-D invariants. In their approach, they consider an action to be a sequence of *poses*. They assume that there exists at least one *key pose* in the sequence in which five points are aligned on a plane in the 3-D world coordinates. Using this assumption, they derive a set of view-invariant descriptors. More recently, the notion of motion history [30] was extended to 3-D by Weinland *et al.* [138] where the authors combine views from multiple cameras to arrive at a 3-D binary occupancy volume. Motion history is computed over these 3-D volumes and view-invariant features are extracted by computing circular fast Fourier transform (FFT) of the volume. All these approaches are strongly tied to the specific choice of feature. There is no general approach of achieving view invariance that can be extended to several features, thus making it an open research issue.

2) Execution Rate Invariance: The second major source of observed variability in features arises from the differences in execution rates while performing the same action. Variations in execution style exist both in interperson and intraperson settings. State–space approaches are robust to minor changes in execution rates, but are not truly rate invariant, because they do not explicitly model transformations of the temporal axis. Mathematically, the variation in execution rate is modeled as a warping function of the temporal scale. The simplest case of linear time warps can be usually dealt with fairly easily. To model highly nonlinear warping functions, the most common method is dynamic time warping (DTW) of the feature sequence such as in [134], [139], and [140]. Recently, Veeraraghavan *et al.* [141] proposed using DTW with constraints to account for the fact that the space of all time-warp functions does not produce physically meaningful actions. DTW is a promising method because it is independent of the choice of feature. The only requirement is

that a distance metric be defined on the feature space. However, DTW requires accurate temporal alignment of test and gallery sequences, i.e., the start and end time instants have to be aligned. Further, the distance computations involved can be prohibitive for long sequences involving many templates. Thus, more efficient methods are required to achieve real-time performance.

3) Anthropometric Invariance: Anthropometric variations such as those induced by the size, shape, gender, etc., of humans is another important class of variabilities that requires careful attention. Unlike viewpoint and execution-rate variabilities that have received significant attention, a systematic study of anthropometric variations has been receiving interest only in recent years. Ad hoc methods that normalize the extracted features to compensate for changes in size, scale, etc., are usually employed when no further information is available. Drawing on studies on human anthropometry Gritai *et al.* [142] suggested that the anthropometric transformation between two different individuals can be modeled as a projective transformation of the image coordinates of body joints. Based on this, they define a similarity metric between actions by using epipolar geometry to provide constraints on actions performed by different individuals. Further research is needed to understand the effects of anthropometric variations and building algorithms to achieve invariance to this factor.

C. Evaluation of Complex Systems

Establishing standardized test beds is a fundamental requirement to compare algorithms and assess progress. It is encouraging to see that several data sets have been made available by research groups and new research is expected to report results on these data sets. Examples include the University of Central Florida (UCF) activity data set [143], Transportation Security Administration (TSA) airport tarmac data set [9], Free Viewpoint National Institute for Research in Computer Science and Control (INRIA) data set [138], and the Royal Institute of Technology (KTH) actions data set [47]. However, most of these data sets consist of simple actions such as opening a closet door, lifting an object, etc. Very few common data sets exist for evaluating higher level complex activities and reasoning algorithms. Complex activity recognition systems consist of a slew of lower level detection and tracking modules. Hence, a straightforward comparison of systems is not easy. One approach to evaluate complex systems is to create ground truth corresponding to outputs from a predefined set of low-level modules. Evaluation would then focus solely on the high-level reasoning engines. While this is one criterion of evaluation, the other criterion is the ability to deal with errors in low-level modules. Participation from the research community is required to address this important issue.

D. Integration With Other Modalities

A vision-based system to recognize human activities can be seen as a crucial stepping stone toward the larger goal of designing machine intelligence systems. To draw a parallel with natural intelligence, humans rely on several modalities including the five classical senses—vision, audition, tactition, olfaction, and gustation—and other senses such as thermoception (temperature) and equilibrioception (balance and acceleration) for everyday tasks. It has also been realized that alternate modalities can improve the performance of vision-based systems, e.g., inertial sensors in structure from motion (SfM), joint audio–video-based tracking [144], etc. Thus, for the longer term pursuit to create machine intelligence, or for the shorter term pursuit of increasing the robustness of action/activity detection modules, integration with other modalities such as audio, temperature, motion, and inertial sensors needs to be investigated in a more systematic manner.

E. Intention Reasoning

Most of the approaches for recognizing and detecting action and activities are based on the premise that the action/activity has already occurred. Reasoning about the intentions of humans and inferring what is going to happen presents a significant intellectual challenge. Security applications are among the first that stand to benefit from such a system, where detection of threat is of utmost importance.

VII. Conclusion

Providing a machine the ability to see and understand as humans do has long fascinated scientists, engineers, and even the common man. Synergistic research efforts in various scientific disciplines, computer vision, artificial intelligence, neuroscience, linguistics, etc., have brought us closer to this goal than at any other point in history. However, several more technical and intellectual challenges need to be tackled before we get there. The advances made so far need to be consolidated, in terms of their robustness to real-world conditions and real-time performance. This would then provide a firmer ground for further research.

References

[1] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception Psychophys.*, vol. 14, no. 2, pp. 201–211, 1973.
[2] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Comput. Vis. Image Understand.*, vol. 73, no. 3, pp. 428–440, 1999.
[3] C. Cedras and M. Shah, "Motion-based recognition: A survey," *Image Vis. Comput.*, vol. 13, no. 2, pp. 129–155, 1995.
[4] D. M. Gavrila, "The visual analysis of human movement: A survey," *Comput. Vis. Image Understand.*, vol. 73, no. 1, pp. 82–98, 1999.
[5] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understand.*, vol. 104, no. 2, pp. 90–126, 2006.
[6] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, pp. 1–45, 2006.
[7] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa, "Matching shape sequences in video with an application to human movement analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1896–1909, Dec. 2005.
[8] B. Georis, M. Maziere, F. Bremond, and M. Thonnat, "A video interpretation platform applied to bank agency monitoring," in *Proc. 2nd Workshop Intell. Distributed Surveillance Syst.*, 2004, pp. 46–50.
[9] N. Vaswani, A. K. Roy-Chowdhury, and R. Chellappa, "Shape activity: A continuous-state HMM for moving/deforming shapes with application to abnormal activity detection," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1603–1616, Oct. 2005.
[10] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The Human ID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.
[11] Y. Rui, T. S. Huang, and S. F. Chang, "Image retrieval: Current techniques, promising directions and open issues," *J. Vis. Commun. Image Represent.*, vol. 10, no. 4, pp. 39–62, 1999.

[12] S. F. Chang, "The Holy Grail of content-based media analysis," *IEEE Multimedia Mag.*, vol. 9, no. 2, pp. 6–10, Apr. 2002.
[13] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 819–826.
[14] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.
[15] W. Hu, D. Xie, T. Tan, and S. Maybank, "Learning activity patterns using fuzzy self-organizing neural network," *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 34, no. 3, pp. 1618–1626, Jun. 2004.
[16] A. Pentland, "Smart rooms, smart clothes," in *Proc. Int. Conf. Pattern Recognit.*, 1998, vol. 2, pp. 949–953.
[17] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan, "Computational studies of human motion: Part 1, tracking and motion synthesis," *Found. Trends Comput. Graphics Vis.*, vol. 1, no. 2-3, pp. 77–254, 2005.
[18] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Comput. Surv.*, vol. 27, no. 3, pp. 433–466, 1995.
[19] T. Huang, D. Koller, J. Malik, G. H. Ogasawara, B. Rao, S. J. Russell, and J. Weber, "Automatic symbolic traffic scene analysis using belief networks," in *Proc. Nat. Conf. Artif. Intell.*, 1994, pp. 966–972.
[20] A. K. Roy-Chowdhury and R. Chellappa, "A factorization approach to activity recognition," in *Proc. CVPR Workshop Event Mining*, 2003, p. 41.
[21] A. M. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2000, pp. 751–767.
[22] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, vol. 8, no. 2, pp. 179–187, Feb. 1962.
[23] H. Freeman, "On the encoding of arbitrary geometric configurations," *IRE Trans. Electron. Comput.*, vol. 10, no. 2, pp. 260–268, Apr. 1961.
[24] D. G. Kendall, "Shape manifolds, procrustean metrics and complex projective spaces," *Bull. Lond. Math. Soc.*, vol. 16, pp. 81–121, 1984.
[25] H. Blum and R. N. Nagel, "Shape description using weighted symmetric axis features," *Pattern Recognit.*, vol. 10, no. 3, pp. 167–180, 1978.
[26] A. Bissacco, P. Saisan, and S. Soatto, "Gait recognition using dynamic affine invariants," presented at the Int. Symp. Math. Theory Netw. Syst., Leuven, Belgium, Jul. 5–9, 2004.
[27] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, no. 2-3, pp. 107–123, 2005.
[28] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Vis. Surveillance Performance Eval. Tracking Surveillance*, 2005, pp. 65–72.
[29] R. Polana and R. C. Nelson, "Detection and recognition of periodic, nonrigid motion," *Int. J. Comput. Vis.*, vol. 23, no. 3, pp. 261–282, 1997.
[30] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.
[31] A. F. Bobick, "Movement, activity, and action: The role of knowledge in the perception of motion," *Philosoph. Trans. Roy. Soc. Lond. B*, vol. 352, pp. 1257–1265, 1997.
[32] T. F. Syeda-Mahmood, M. Vasilescu, and S. Sethi, "Recognizing action events from multiple viewpoints," in *Proc. IEEE Workshop Detection Recognit. Events Video*, 2001, pp. 64–72.
[33] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 984–989.
[34] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.
[35] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[36] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2001, pp. 585–591.
[37] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
[38] Y. Yacoob and M. J. Black, "Parameterized modeling and recognition of activities," *Comput. Vis. Image Understand.*, vol. 73, no. 2, pp. 232–247, 1999.
[39] A. M. Elgammal and C. S. Lee, "Inferring 3D body pose from silhouettes using activity manifold learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 681–688.
[40] R. Pless, "Image spaces and video trajectories: Using Isomap to explore video sequences," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1433–1440.
[41] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanism," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 7, no. 5, pp. 923–932, May 1990.
[42] R. A. Young, R. M. Lesperance, and W. W. Meyer, "The Gaussian derivative model for spatial-temporal vision: I. Cortical model," *Spatial Vis.*, vol. 14, no. 3–4, pp. 261–319, 2001.
[43] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[44] O. Chomat and J. L. Crowley, "Probabilistic recognition of activity using local appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1999, vol. 02, pp. 104–109.
[45] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 2, pp. 123–130.
[46] J. C. Niebles, H. Wang, and L. F. Fei, "Unsupervised learning of human action categories using spatial-temporal words," in *Proc. British Mach. Vis. Conf.*, 2006, pp. 1249–1258.
[47] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. Int. Conf. Pattern Recognit.*, 2004, pp. 32–36.
[48] S. Savarese, A. Del Pozo, J. C. Niebles, and L. Fei-Fei, "Spatial-temporal correlations for unsupervised action classification," presented at the IEEE Workshop Motion Video Comput., Copper Mountain, CO, Jan. 8–9, 2008.
[49] S. Nowozin, G. Bakir, and K. Tsuda, "Discriminative subsequence mining for action classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[50] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 17–31, 2007.
[51] S. F. Wong, T. K. Kim, and R. Cipolla, "Learning motion categories using both semantic and structural information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–6.
[52] Y. Song, L. Goncalves, and P. Perona, "Unsupervised learning of human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 814–827, Jul. 2003.
[53] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[54] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 405–412.
[55] P. A. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[56] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 166–173.
[57] Y. Ke, R. Sukthankar, and M. Hebert, "Spatio-temporal shape and flow correlation for action recognition," in *Proc. Visual Surveillance Workshop*, 2007, pp. 1–8.
[58] M. A. O. Vasilescu, "Human motion signatures: Analysis, synthesis, recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2002, pp. 456–460.
[59] T. K. Kim, S. F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[60] L. Wolf, H. Jhuang, and T. Hazan, "Modeling appearances with low-rank SVM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–6.
[61] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[62] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1992, pp. 379–385.

[63] J. Schlenzig, E. Hunter, and R. Jain, "Recursive identification of gesture inputs using hidden Markov models," in *Proc. 2nd IEEE Workshop Appl. Comput. Vis.*, 1994, pp. 187–194.

[64] A. D. Wilson and A. F. Bobick, "Learning visual behavior for gesture analysis," in *Proc. Int. Symp. Comput. Vis.*, 1995, pp. 229–234.

[65] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1371–1375, Dec. 1998.

[66] A. Kale, A. Sundaresan, A. N. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa, "Identification of humans using gait," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1163–1173, Sep. 2004.

[67] Z. Liu and S. Sarkar, "Improved gait recognition by gait dynamics normalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 863–876, Jun. 2006.

[68] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1997, pp. 994–999.

[69] D. J. Moore, I. A. Essa, and M. H. Hayes, "Exploiting human actions and object context for recognition tasks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1999, pp. 80–86.

[70] S. Hongeng and R. Nevatia, "Large-scale event detection using semi-hidden Markov models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1455–1462.

[71] N. P. Cuntoor and R. Chellappa, "Mixed-state models for nonstationary multiobject activities," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 106–119, 2007.

[72] A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto, "Recognition of human gaits," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 2, pp. 52–57.

[73] M. C. Mazzaro, M. Sznaier, and O. Camps, "A model (in) validation approach to gait classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1820–1825, Nov. 2005.

[74] N. P. Cuntoor and R. Chellappa, "Epitomic representation of human activities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[75] P. Saisan, G. Doretto, Y. N. Wu, and S. Soatto, "Dynamic texture recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. 58–63.

[76] A. B. Chan and N. Vasconcelos, "Classifying video with kernel dynamic textures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–6.

[77] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the em algorithm," *J. Time Series Anal.*, vol. 3, no. 4, pp. 253–264, 1982.

[78] P. V. Overschee and B. D. Moor, "Subspace algorithms for the stochastic identification problem," *Automatica*, vol. 29, no. 3, pp. 649–660, 1993.

[79] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, CRG-TR-96-2, 1996.

[80] K. D. Cock and B. D. Moor, "Subspace angles between ARMA models," *Syst. Control Lett.*, vol. 46, pp. 265–270, 2002.

[81] R. J. Martin, "A metric for ARMA processes," *IEEE Trans. Signal Process.*, vol. 48, no. 4, pp. 1164–1170, Apr. 2000.

[82] R. Vidal and P. Favaro, "Dynamicboost: Boosting time series generated by dynamical systems," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–6.

[83] S. V. N. Vishwanathan, A. J. Smola, and R. Vidal, "Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes," *Int. J. Comput. Vis.*, vol. 73, no. 1, pp. 95–119, 2007.

[84] A. Bissacco and S. Soatto, "On the blind classification of time series," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–7.

[85] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[86] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *Int. J. Comput. Vis.*, vol. 51, no. 2, pp. 91–109, 2003.

[87] C. Bregler, "Learning and recognizing human dynamics in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1997, p. 568.

[88] B. North, A. Blake, M. Isard, and J. Rittscher, "Learning and classification of complex dynamics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 1016–1034, Sep. 2000.

[89] V. Pavlovic, J. M. Rehg, and J. MacCormick, "Learning switching linear models of human motion," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2000, pp. 981–987.

[90] V. Pavlovic and J. M. Rehg, "Impact of dynamic model learning on classification of human motion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2000, pp. 1788–1795.

[91] S. M. Oh, J. M. Rehg, T. R. Balch, and F. Dellaert, "Data-driven MCMC for learning and inference in switching linear dynamic systems," in *Proc. Nat. Conf. Artif. Intell.*, 2005, pp. 944–949.

[92] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufmann, 1988.

[93] H. Buxton and S. Gong, "Visual surveillance in a dynamic and uncertain world," *Artif. Intell.*, vol. 78, no. 1-2, pp. 431–459, 1995.

[94] P. Remagnino, T. Tan, and K. Baker, "Agent orientated annotation in model based visual surveillance," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1998, pp. 857–862.

[95] S. Park and J. K. Aggarwal, "Recognition of two-person interactions using a hierarchical Bayesian network," *ACM J. Multimedia Syst.*, vol. 10, Special Issue on Video Surveillance, no. 2, pp. 164–179, 2004.

[96] S. S. Intille and A. F. Bobick, "A framework for recognizing multi-agent action from visual evidence," in *Proc. Nat. Conf. Artif. Intell.*, 1999, pp. 518–525.

[97] S. Gong and T. Xiang, "Recognition of group activities using dynamic probabilistic networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 742–749.

[98] R. L. Kashyap, "Bayesian comparison of different classes of dynamic models using empirical data," *IEEE Trans. Autom. Control*, vol. AC-22, no. 5, pp. 715–727, Oct. 1977.

[99] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[100] A. Gupta and L. S. Davis, "Objects in action: An approach for combining action understanding and object perception," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[101] M. I. Jordan, *Learning in Graphical Models*. Cambridge, MA: The MIT Press, 1998.

[102] N. Friedman and D. Koller, "Being Bayesian about Bayesian network structure: A Bayesian approach to structure discovery in Bayesian networks," *Mach. Learn.*, vol. 50, no. 1–2, pp. 95–125, 2003.

[103] C. A. Petri, "Communication with automata," Defense Tech. Inf. Cntr., Fort Belvoir, VA, DTIC Res. Rep. AD0630125, 1966.

[104] R. David and H. Alla, "Petri nets for modeling of dynamic systems a survey," *Automatica*, vol. 30, no. 2, pp. 175–202, 1994.

[105] T. Murata, "Petri nets: Properties, analysis and applications," *Proc. IEEE*, vol. 77, no. 4, pp. 541–580, Apr. 1989.

[106] C. Castel, L. Chaudron, and C. Tessier, "What is going on? a high-level interpretation of a sequence of images," in *Proc. ECCV Workshop Conceptual Descriptions Images*, 1996, pp. 13–27.

[107] N. Ghanem, D. DeMenthon, D. Doermann, and L. Davis, "Representation and recognition of events in surveillance video using Petri nets," in *Proc. 2nd IEEE Workshop Event Mining*, 2004, p. 112.

[108] M. Albanese, R. Chellappa, V. Moscato, A. Picariello, V. S. Subrahmanian, P. Turaga, and O. Udrea, "A constrained probabilistic petri net framework for human activity detection in video," *IEEE Trans. Multimedia*, to be published.

[109] C. S. Pinhanez and A. F. Bobick, "Human action detection using pnf propagation of temp oral constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1998, p. 898.

[110] Y. Shi, Y. Huang, D. Minen, A. Bobick, and I. Essa, "Propagation networks for recognizing partially ordered sequential action," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, vol. 2, pp. 862–869.

[111] Y. Shi, A. F. Bobick, and I. A. Essa, "Learning temporal sequence model from partially labeled data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1631–1638.

[112] R. Hamid, A. Maddi, A. Bobick, and I. Essa, "Structure from statistics—unsupervised activity analysis using suffix trees," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

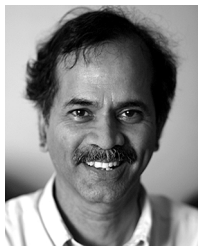
[113] M. Brand, "Understanding manipulation in video," in *Proc. 2nd Int. Conf. Autom. Face Gesture Recognit.*, 1996, p. 94.

[114] M. S. Ryoo and J. K. Aggarwal, "Recognition of composite human activities through context-free grammar based representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1709–1718.
[115] J. Earley, "An efficient context-free parsing algorithm," *Commun. ACM*, vol. 13, no. 2, pp. 94–102, 1970.
[116] A. V. Aho and J. D. Ullman, *The Theory of Parsing, Translation, and Compiling, Volume 1: Parsing*. Englewood Cliffs, NJ: Prentice-Hall, 1972.
[117] C. D. L. Higuera, "Current trends in grammatical inference," in *Proc. Joint IAPR Int. Workshops Adv. Pattern Recognit.*, 2000, pp. 28–31.
[118] Y. A. Ivanov and A. F. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 852–872, Aug. 2000.
[119] D. Moore and I. Essa, "Recognizing multitasked activities from video using stochastic context-free grammar," in *Proc. 18th Nat. Conf. Artif. Intell.*, 2002, pp. 770–776.
[120] S. W. Joo and R. Chellappa, "Recognition of multi-object events using attribute grammars," in *Proc. Int. Conf. Image Process.*, 2006, pp. 2897–2900.
[121] N. Rota and M. Thonnat, "Activity recognition from video sequences using declarative models," in *Proc. 14th Eur. Conf. Artif. Intell.*, 2000, pp. 673–680.
[122] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 8, pp. 873–889, Aug. 2001.
[123] S. Hongeng, R. Nevatia, and F. Bremond, "Video-based event recognition: Activity representation and probabilistic recognition methods," *Comput. Vis. Image Understand.*, vol. 96, no. 2, pp. 129–162, 2004.
[124] V. D. Shet, D. Harwood, and L. S. Davis, "Vidmap: Video monitoring of activity with prologue," in *Proc. IEEE Conf. Adv. Video Signal Based Surveillance*, 2005, pp. 224–229.
[125] S. Tran and L. S. Davis, "Visual event modeling and recognition using Markov logic networks," presented at the IEEE Eur. Conf. Comput. Vis., Marseille, France, Oct. 2008.
[126] D. Chen, J. Yang, and H. D. Wactlar, "Towards automatic analysis of social interaction patterns in a nursing home environment from video," in *Proc. 6th ACM SIGMM Int. Workshop Multimedia Inf. Retrieval*, 2004, pp. 283–290.
[127] A. Hakeem and M. Shah, "Ontology and taxonomy collaborated framework for meeting classification," in *Proc. Int. Conf. Pattern Recognit.*, 2004, pp. 219–222.
[128] S. Guler, J. B. Burns, A. Hakeem, Y. Sheikh, M. Shah, M. Thonnat, F. Bremond, N. Maillot, T. V. Vu, I. Haritaoglu, R. Chellappa, U. Akdemir, and L. Davis, "An ontology of video events in the physical security and surveillance domain," [Online]. Available: http://www.ai.sri.com/~burns/EventOntology, work done as part of the ARDA video event Challenge Workshop, 2003
[129] J. Hobbs, R. Nevatia, and B. Bolles, "An ontology for video event representation," in *Proc. IEEE Workshop Event Detection Recognit.*, 2004, p. 119.
[130] A. R. J. Francois, R. Nevatia, J. Hobbs, and R. C. Bolles, "Verl: An ontology framework for representing and annotating video events," *IEEE MultiMedia Mag.*, vol. 12, no. 4, pp. 76–86, Oct.-Dec. 2005.
[131] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 781–796, Aug. 2000.
[132] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 726–733.
[133] Y. Sheikh, M. Sheikh, and M. Shah, "Exploring the space of a human action," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 144–149.
[134] T. J. Darrell, I. A. Essa, and A. P. Pentland, "Task-specific gesture analysis in real-time using interpolated views," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 12, pp. 1236–1242, Dec. 1996.
[135] S. M. Seitz and C. R. Dyer, "View-invariant analysis of cyclic motion," *Int. J. Comput. Vis.*, vol. 25, no. 3, pp. 231–251, 1997.
[136] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 203–226, 2002.
[137] V. Parameswaran and R. Chellappa, "View invariance for human action recognition," *Int. J. Comput. Vis.*, vol. 66, no. 1, 2006.
[138] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Comput. Vis. Image Understand.*, vol. 104, no. 2, pp. 249–257, 2006.
[139] K. Takahashi, S. Seki, E. Kojima, and R. Oka, "Recognition of dexterous manipulations from time-varying images," in *Proc. IEEE Workshop Motion Non-Rigid Articulated Objects*, 1994, pp. 23–28.
[140] M. A. Giese and T. Poggio, "Morphable models for the analysis and synthesis of complex motion patterns," *Int. J. Comput. Vis.*, vol. 38, no. 1, pp. 59–73, 2000.
[141] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury, "The function space of an activity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 959–968.
[142] A. Gritai, Y. Sheikh, and M. Shah, "On the use of anthropometry in the invariant analysis of human actions," in *Int. Conf. Pattern Recognit.*, 2004, pp. 923–926.
[143] C. Rao and M. Shah, "View-invariance in action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. 316–322.
[144] V. Cevher, A. Sankaranarayanan, J. H. McClellan, and R. Chellappa, "Target tracking using a joint acoustic video system," *IEEE Trans. Multimedia*, vol. 9, no. 4, pp. 715–727, Jun. 2007.

Pavan Turaga (S'05) received the B.Tech. degree in electronics and communication engineering from the Indian Institute of Technology, Guwahati, India, in 2004. He is working towards the Ph.D. degree in electrical engineering at the Department of Electrical and Computer Engineering, University of Maryland, College Park.

He is a recipient of the University of Maryland graduate school fellowship for 2004–2006. His research interests are in statistics and machine learning with applications to computer vision and pattern analysis.

Mr. Turaga is a student member of Association for the Advancement of Artificial Intelligence (AAAI). He was selected to participate in the Emerging Leaders in Multimedia Workshop by IBM, New York, in 2008.



Rama Chellappa (F'92) received the B.E. degree (with honors) in electronics and communication engineering from the University of Madras, Madras, India, in 1975, the M.E. degree (with distinction) in electrical and communication engineering from the Indian Institute of Science, Bangalore, India, in 1977, and the M.S.E.E. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, in 1978 and 1981, respectively.

Since 1991, he has been a Professor of Electrical and Computer Engineering and an affiliate Professor of Computer Science at the University of Maryland, College Park. He is also affiliated with the Center for Automation Research (Director), the Institute for Advanced Computer Studies (Permanent Member), the Applied Mathematics program, and the Chemical Physics program. In 2005, he was named a Minta Martin Professor of Engineering. Prior to joining the University of Maryland, he was an Assistant (1981–1986) and Associate Professor (1986–1991) and Director of the Signal and Image Processing Institute (1988–1990) at the University of Southern California (USC), Los Angeles. Over the last 27 years, he has published numerous book chapters and peer-reviewed journal and conference papers. He has also coedited and coauthored many research monographs on Markov random fields, biometrics, and surveillance. His current research interests are in face and gait analysis, 3-D modeling from video, automatic target recognition from stationary and moving platforms, surveillance and monitoring, hyperspectral processing, image understanding, and commercial applications of image processing and understanding.

Dr. Chellappa has served as an Associate Editor of four IEEE TRANSACTIONS. He was a Co-Editor-in-Chief of *Graphical Models and Image Processing*. He also served as the Editor-in-Chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He served as a member of the IEEE Signal Processing Society Board of Governors and as its Vice President of Awards and Membership. Recently, he has been elected to serve a two-year term as the President of the newly constituted IEEE Biometrics Council. He has received several awards, including the National Science Foundation Presidential Young Investigator Award in 1985, four IBM Faculty Development Awards, the 1990 Excellence in Teaching Award from the School of Engineering at USC, the 1992 Best Industry Related Paper Award from the International Association of Pattern Recognition (with Q. Zheng), and the 2000 Technical Achievement Award from IEEE Signal Processing Society. He was elected as

a Distinguished Faculty Research Fellow (1996–1998), and as a Distinguished Scholar Teacher (2003) at the University of Maryland. He coauthored a paper that received the Best Student Paper in the Computer Vision Track at the International Association of Pattern Recognition in 2006. He is a corecipient (with A. Sundaresan) of the 2007 Outstanding Innovator of the Year Award from the Office of Technology Commercialization and received the 2007 A. J. Clark School of Engineering Faculty Outstanding Research Award. He is serving as a Distinguished Lecturer of the IEEE Signal Processing Society for the period 2008–2009 and received the Society's Meritorious Service Award in 2008. He is a Golden Core Member of the IEEE Computer Society, received its Meritorious Service Award in 2004, and has been selected to receive its Technical Achievement Award in 2008. He has served as a General and Technical Program Chair for several IEEE international and national conferences and workshops. He is a Fellow of the International Association for Pattern Recognition.

V. S. Subrahmanian is currently Professor of Computer Science at the University of Maryland, College Park and Director of the University of Maryland's Institute for Advanced Computer Studies (UMIACS). He has worked on nonmonotonic and probabilistic logics, inconsistency management in databases, database models views and inference, rule bases, heterogeneous databases, multimedia databases, probabilistic databases, and agent systems. He has edited two books, one on nonmonotonic reasoning (MIT Press) and one on multimedia databases (Springer-Verlag). He has coauthored an advanced database textbook (Morgan Kaufman, 1997) and a book on heterogeneous software agents. He is the sole author of a textbook on multimedia databases (Morgan Kaufmann).

Prof. Subrahmanian received the NSF Young Investigator Award in 1993 and the Distinguished Young Scientist Award from the Maryland Academy of Science/Maryland Science Center in 1997. He has given invited talks at numerous national and international conferences—in addition, he has served on numerous conference and funding panels, as well as on the program committees of numerous conferences. He has also chaired several conferences. He is or has previously been on the editorial boards of several journals. He has served on DARPA's (Defense Advanced Research Projects Agency) Executive Advisory Council on Advanced Logistics and as an ad hoc member of the U.S. Air Force Science Advisory Board (2001). He also serves on the Board of Directors of the Development Gateway Foundation—an organization that focuses on using information technology in supporting poverty reduction in developing nations.

Octavian Udrea received the B.S. and M.S. degrees from the Polytechnic University of Bucharest, Bucharest, Romania, in 2003 and 2004, respectively, and the Ph.D. degree from the University of Maryland, College Park, in August 2008, all in computer science.

He will join the IBM T. J. Watson Research Center in fall 2008. His primary research interests include knowledge representation, heterogeneous databases, automated code verification, and activity detection in video databases.

Dr. Udrea is a student member of the Association for Computing Machinery (ACM) and the Association for the Advancement of Artificial Intelligence (AAAI).

Social Signal Processing: Survey of an Emerging Domain

Alessandro Vinciarelli [a,b] Maja Pantic [c,d] Hervé Bourlard [a,b]

[a] *IDIAP Research Institute - CP592 - 1920 Martigny (Switzerland)*

[b] *Ecole Polytechnique Fédérale de Lausanne (EPFL) - 1015 Lausanne (CH)*

[c] *Imperial College - 180 Queens Gate - London SW7 2AZ (UK)*

[d] *University of Twente - Drienerlolaan 5 -7522 NB Enschede (The Netherlands)*

Abstract

The ability to understand and manage social signals of a person we are communicating with is the core of social intelligence. Social intelligence is a facet of human intelligence that has been argued to be indispensable and perhaps the most important for success in life. This paper argues that next-generation computing needs to include the essence of social intelligence – the ability to recognize human social signals and social behaviours like turn taking, politeness, and disagreement – in order to become more effective and more efficient. Although each one of us understands the importance of social signals in everyday life situations, and in spite of recent advances in machine analysis of relevant behavioural cues like blinks, smiles, crossed arms, laughter, and similar, design and development of automated systems for Social Signal Processing (SSP) are rather difficult. This paper surveys the past efforts in solving these problems by a computer, it summarizes the relevant findings in social psychology, and it proposes a set of recommendations for enabling the development of the next generation of socially-aware computing.

Key words: Social signals, computer vision, speech processing, human behaviour analysis, social interactions.

1 Introduction

The exploration of how human beings react to the world and interact with it and each other remains one of the greatest scientific challenges. Perceiv-

Email addresses: vincia@idiap.ch (Alessandro Vinciarelli), m.pantic@imperial.ac.uk (Maja Pantic), bourlard@idiap.ch (Hervé Bourlard).

ing, learning, and adapting to the world are commonly labelled as intelligent behaviour. But what does it mean being intelligent? Is IQ a good measure of human intelligence and the best predictor of somebody's success in life? There is now a growing research in cognitive sciences, which argues that our common view of intelligence is too narrow, ignoring a crucial range of abilities that matter immensely for how people do in life. This range of abilities is called *social intelligence* [6][8][19][182] and includes the ability to express and recognise social signals and social behaviours like turn taking, agreement, politeness, and empathy, coupled with the ability to manage them in order to get along well with others while winning their cooperation. Social signals and social behaviours are the expression of ones attitude towards social situation and interplay, and they are manifested through a multiplicity of non-verbal behavioural cues including facial expressions, body postures and gestures, and vocal outbursts like laughter (see Figure 1). Social signals typically last for a short time (milliseconds, like turn taking, to minutes, like mirroring), compared to social behaviours that last longer (seconds, like agreement, to minutes, like politeness, to hours or days, like empathy) and are expressed as temporal patterns of non-verbal behavioural cues. The skills of social intelligence have been argued to be indispensable and perhaps the most important for success in life [66].

When it comes to computers, however, they are socially ignorant [143]. Current computing devices do not account for the fact that human-human communication is always socially situated and that discussions are not just facts but part of a larger social interplay. However, not all computers will need social intelligence and none will need all of the related skills humans have. The current-state-of-the-art categorical computing works well and will always work well for context-independent tasks like making plane reservations and buying and selling stocks. However, this kind of computing is utterly inappropriate for virtual reality applications as well as for interacting with each of the (possibly hundreds) computer systems diffused throughout future smart environments (predicted as the future of computing by several visionaries such as Mark Weiser) and aimed at improving the quality of life by anticipating the users needs. Computer systems and devices capable of sensing agreement, inattention, or dispute, and capable of adapting and responding to these social signals in a polite, unintrusive, or persuasive manner, are likely to be perceived as more natural, efficacious, and trustworthy. For example, in education, pupils' social signals inform the teacher of the need to adjust the instructional message. Successful human teachers acknowledge this and work with it; digital conversational embodied agents must begin to do the same by employing tools that can accurately sense and interpret social signals and social context of the pupil, learn successful context-dependent social behaviour, and use a proper socially-adept presentation language (see e.g. [141]) to drive the animation of the agent. The research area of machine analysis and employment of human social signals to build more natural, flexible computing

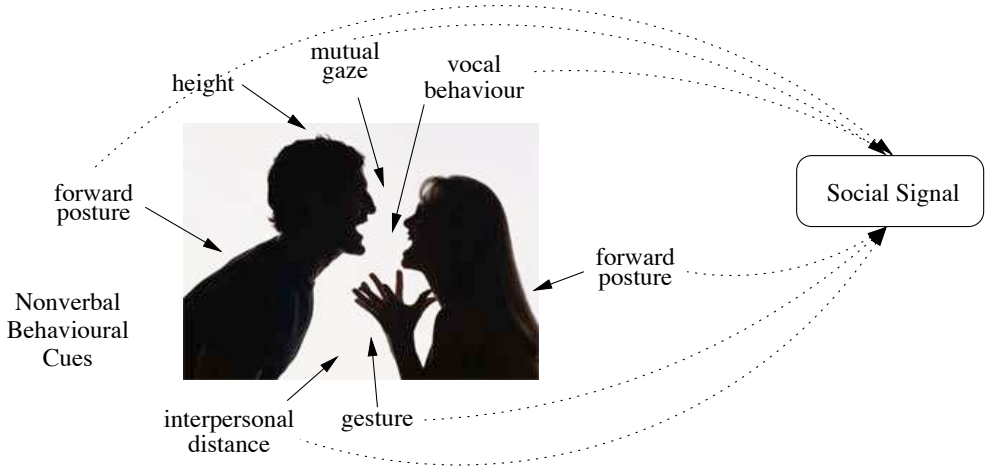


Fig. 1. Behavioural cues and social signals. Multiple behavioural cues (vocal behaviour, posture, mutual gaze, interpersonal distance, etc.) combine to produce a social signal (in this case aggressivity or disagreement) that is evident even if the picture shows only the silhouettes of the individuals involved in the interaction.

technology goes by the general name of Socially-Aware Computing as introduced by Pentland [142][143].

Although the importance of social signals in everyday life situations is evident, and in spite of recent advances in machine analysis and synthesis of relevant behavioural cues like gaze exchange, blinks, smiles, head nods, crossed arms, laughter, and similar [137][138], the research efforts in machine analysis and synthesis of human social signals like attention, empathy, politeness, flirting, (dis)agreement, etc., are still tentative and pioneering efforts. The importance of studying social interactions and developing automated assessing of human social behaviour from audiovisual recordings is undisputable. It will result in valuable multimodal tools that could revolutionise basic research in cognitive and social sciences by raising the quality and shortening the time to conduct research that is now lengthy, laborious, and often imprecise. At the same time, and as outlined above, such tools form a large step ahead in realising naturalistic, socially-aware computing and interfaces, built for humans, based on models of human behaviour.

Social Signal Processing (SSP) [143][145][202][203] is the new research and technological domain that aims at providing computers with the ability to sense and understand human social signals. Despite being in its initial phase, SSP has already attracted the attention of the technological community: the MIT Technology Review magazine identifies reality mining (one of the main applications of SSP so far, see Section 4 for more details), as one of the ten technologies likely to change the world [69], while management experts expect SSP to change organization studies like the microscope has changed medicine few centuries ago [19].

To the best of our knowledge, this is the first attempt to survey the past work done on SSP. The innovative and multidisciplinary character of the research on SSP is the main reason for this state of affairs. For example, in contrast to the research on human affective behaviour analysis that witnessed tremendous progress in the past decade (for exhaustive surveys in the field see, e.g.,[76][140][221]), the research on machine analysis of human social behaviour just started to attract the interest of the research community in computer science. This and the fragmentation of the research over several scientific communities including those in psychology, computer vision, speech and signal processing, make the exercise of surveying the current efforts in machine analysis of human social behaviour difficult.

The paper begins by examining the context in which the research on SSP has arisen and by providing a taxonomy of the target problem domain (Section 2). The paper surveys then the past work done in tackling the problems of machine detection and interpretation of social signals and social behaviours in real-world scenarios (Section 3). Existing research efforts to apply social signal processing to automatic recognition of socially relevant information such as someone's role, dominance, influence, etc., are surveyed next (Section 4). Finally, the paper discusses a number of challenges facing researchers in the field (Section 5). In the authors' opinion, these need to be addressed before the research in the field can enter its next phase – deployment of research findings in real-world applications.

2 Behavioural Cues and Social Signals: A Taxonomy

There is more than words in social interactions [9], whether these take place between humans or between humans and computers [30]. This is well known to social psychologists that have studied nonverbal communication for several decades [96][158]. It is what people experience when they watch a television program in a language they do not understand and still capture a number of important social cues such as differences in status between individuals, overall atmosphere of interactions (e.g., tense vs. relaxed), rapport beteween people (mutual trust vs. mutual distrust), etc.

Nonverbal behaviour is a continuous source of signals which convey information about feelings, mental state, personality, and other traits of people [158]. During social interactions, nonverbal behaviour conveys this information not only for each of the involved individuals, but it also determines the nature and quality of the social relationships they have with others. This happens through a wide spectrum of nonverbal behavioural cues [7][8] that are perceived and displayed mostly unconsciously while producing *social awareness*, i.e. a spontaneous understanding of social situations that does not require attention or

reasoning [98].

The term behavioural cue is typically used to describe a set of temporal changes in neuromuscular and physiological activity that last for short intervals of time (milliseconds to minutes) in contrast to *behaviours* (e.g. social behaviours like politeness or empathy) that last on average longer (minutes to hours). As summarised in [47] among the types of messages (communicative intentions) conveyed by behavioural cues are the following:

- *affective/attitudinal/cognitive states* (e.g. fear, joy, stress, disagreement, ambivalence and inattention),
- *emblems* (culture-specific interactive signals like wink or thumbs up),
- *manipulators* (actions used to act on objects in the environment or self-manipulative actions such as lip biting and scratching),
- *illustrators* (actions accompanying speech such as finger pointing and raised eyebrows), and
- *regulators* (conversational mediators such as the exchange of a look, palm pointing, head nods and smiles).

In most cases, behavioural cues accompany verbal communication and, even if they are invisible, i.e., they are sensed and interpreted outside conscious awareness, they have a major impact on the perception of verbal messages and social situations [96]. Early investigations of verbal and nonverbal components in interaction (in particular [113] as cited in [96]) have suggested that the verbal messages account for just 7% of the overall social perception. This conclusion has been later argued because the actual weight of the different messages (i.e. verbal vs non-verbal) depends on the context and on the specific kind of interaction [45]. However, more recent studies still confirm that the nonverbal behaviour plays a major role in shaping the perception of social situations: e.g., judges assessing the rapport between two people are more accurate when they use only the facial expressions than when they use only the verbal messages exchanged [8]. Overall, the nonverbal social signals seem to be the predominant source of information used in understanding social interactions [9].

The rest of this section provides a taxonomy of the SSP problem domain by listing and explaining the most important behavioural cues and their functions in social behaviour. Behavioural cues that we included in this list are those that the research in psychology has recognized as being the most important in human judgments of social behaviour. Table 1 provides a synopsis of those behavioural cues, the social signals they are related to, and the technologies that can be used to sense and analyse them. For more exhaustive explanations of nonverbal behaviours and the related behavioural cues, readers are referred to [7][47][96][158].

	Example Social Behaviours							Tech.		
Social Cues	emotion	personality	status	dominance	persuasion	regulation	rapport	speech anlysis	computer vision	biometry
Physical appearance										
height			√	√					√	√
attractiveness		√	√	√	√		√		√	√
body shape		√		√					√	√
Gesture and posture										
hand gestures	√	√			√	√	√		√	√
posture	√	√	√	√	√	√	√		√	√
walking		√	√	√					√	√
Face and eyes behaviour										
facial expressions	√	√	√	√	√	√	√		√	√
gaze behaviour	√	√	√	√	√	√	√		√	
focus of attention	√	√	√	√	√	√	√		√	
Vocal behaviour										
prosody	√	√		√	√		√	√		
turn taking	√	√	√	√		√	√	√		
vocal outbursts	√	√		√	√	√	√	√		
silence	√		√				√	√		
Space and Environment										
distance	√	√	√		√		√		√	
seating arrangement				√	√		√		√	

Table 1
The table shows the behavioural cues associated to some of the most important social behaviours as well as the technologies involved in their automatic detection.

2.1 Physical Appearance

The physical appearance includes natural characteristics such as height, body shape, physiognomy, skin and hair color, as well as artificial characteristics such as clothes, ornaments, make up, and other manufacts used to modify/

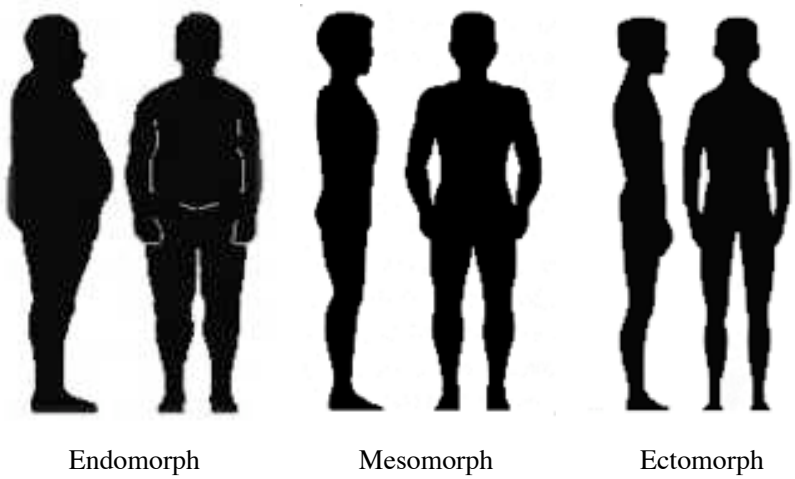


Fig. 2. Somatotypes. The figure shows the three body shapes that tend to elicit the perception of specific personality traits.

accentuate the facial/ body aspects.

The main social signal associated to physical appearance is the *attractiveness*. Attractiveness produces a positive *halo effect* (a phenomenon also known as "*what is beautful is good*" [41]). Attractive people are often judged as having high status and good personality even if no objective basis for such judgments exists [70][208]. Attractive people also have higher probability of starting new social relationships with people they do not know [158]. Other physical characteristics are not necessarily related to the attractiveness, but still have a major influence on social perceptions. The most important are height and *somatotype* (see below). Tall individuals tend to be attributed higher social status and, in some cases, they actually hold a higher status. For example, a survey has shown that the average height of the American CEOs of the Fortune 500 companies is around 7.5 cm higher than the average height of the American population. Moreover, 30% of the same CEOs are taller than 190 cm, while only 4% of the rest of the American population lies in the same range of height [63].

Different *somatotypes* (see Figure 2), tend to elicit the attribution of certain personality traits [25]. For example, *endomorphic* individuals (round, fat and soft) tend to be perceived as more talkative and sympathetic, but also more dependent on others. *Mesomorphic* individuals (bony, muscular and athletic) tend to be perceived as more self-reliant, more mature in behaviour and stronger, while *ectomorphic* individuals (tall, thin and fragile) tend to be perceived as more tense, more nervous, more pessimistic and inclined to be difficult. These judgments are typically influenced by stereotypes that do not necessarily correspond to the reality, but still influence significantly the social perceptions [96].

Following the work of Darwin [37], which was the first to describe body expressions associated with emotions in animals and humans, there have been a number of studies on human body postures and gestures communicating emotions. For example the works in [27][198] investigated perception and display of body postures relevant to basic emotions including happiness, sadness, surprise, fear, disgust, and anger, while the studies in [72][152] investigated bodily expressions of felt and recognized basic emotions as visible in specific changes in arm movement, gait parameters, and kinematics. Overall, these studies have shown that both posture and body/ limb motions change with emotion expressed. Basic research also provides evidence that gestures like head inclination, face touching, and shifting posture often accompany social affective states like shame and embarrassment [26][50]. However, as indicated by researchers in the field (e.g. in [112]), as much as 90% of body gestures are associated with speech, representing typical social signals such as illustrators, emblems, and regulators.

In other words, gestures are used in most cases to regulate interactions (e.g., to yield the turn in a conversation), to communicate a specific meaning (e.g., the *thumbs up* gesture to show appreciation), to punctuate a discourse (e.g., to underline an utterance by rising the index finger), to greet (e.g., by waving hands to say goodbye), etc. [123]. However, in some cases gestures are performed unconsciously and they are interesting from an SSP point of view because they account for *honest* information [146], i.e., they leak cues related to the actual attitude of a person with respect to a social context. In particular, *adaptors* express boredom, stress and negative feelings towards others. Adaptors are usually displayed unconsciously and include self-manipulations (e.g., scratching, nose and ear touching, hair twisting), manipulation of small objects (e.g., playing with pens and papers), and self-protection gestures (e.g., folding arms or rythmicly moving legs) [96].

Postures are also typically assumed unconsciously and, arguably, they are the most reliable cues about the actual attitude of people towards social situations [158]. One of the main classifications of postural behaviours proposes three main criteria to assess the social meaning of postures [166]. The first criterion distinguishes between *inclusive* and *non-inclusive* postures and accounts for how much a given posture takes into account the presence of others. For example, facing in the opposite direction with respect to others is a clear sign of non-inclusion. The second criterion is *face-to-face vs. parallel body orientation* and concerns mainly people involved in conversations. Face-to-face interactions are in general more active and engaging (the frontal position addresses the need of continuous mutual monitoring), while people sitting parallel to each other tend to be either buddies or less mutually interested. The

Congruent postures

Non–congruent postures

Fig. 3. Postural congruence. The figure on the left shows how people deeply involved in an interaction tend to assume the same posture. In the other picture, the forward inclination of the person on the right is not reciprocated by the person on the left.

third criterion is *congruence vs. incongruence*: symmetric postures tend to account for a deep psychological involvement (see left picture in Figure 3), while non-symmetric ones correspond to the opposite situation. The postural congruence is an example of a general phenomenon called *chameleon effect* or *mirroring* [22], that consists of the mutual imitation of people as a mean to display affiliation and liking. Postural behaviour includes also walking and movements that convey social information such as status, dominance and affective state [109].

2.3 Face and Eye Behaviour

The human face is involved in an impressive variety of different activities. It houses the majority of our sensory apparatus: eyes, ears, mouth and nose, allowing the bearer to see, hear, taste and smell. Apart from these biological functions, the human face provides a number of signals essential for interpersonal communication in our social life. The face houses the speech production apparatus and is used to identify other members of the species, to regulate the conversation by gazing or nodding, and to interpret what has been said by lip reading. It is our direct and naturally preeminent means of communicating and understanding somebody's affective state and intentions on the basis of the shown facial expression [89]. Personality, attractiveness, age and gender can be also seen from someone's face [8]. Thus the face is a multi-signal sender/receiver capable of tremendous flexibility and specificity. It is therefore not surprising that the experiments (see beginning of Section 2) about the relative weight of the different nonverbal components in shaping social perceptions always show that facial behaviour plays a major role [8][68][113].

Two major approaches to facial behaviour measurement in psychological re-

Fig. 4. Basic emotions. Prototypic facial expressions of six basic emotions (disgust, happiness, sadness, anger, fear, and surprise).

search are message and sign judgment [23]. The aim of message judgment is to infer what underlies a displayed facial expression, such as affect or personality, while the aim of sign judgment is to describe the *surface* of the shown behavior, such as facial movement or facial component shape. Thus, a brow furrow can be judged as *anger* in a message-judgment and as a facial movement that lowers and pulls the eyebrows closer together in a sign-judgment approach. While message judgment is all about interpretation, sign judgment attempts to be objective, leaving inference about the conveyed message to higher order decision making.

As indicated in [23], most commonly used facial expression descriptors in message judgment approaches are the six basic emotions (fear, sadness, happiness, anger, disgust, surprise; see Fig. 4), proposed by Ekman and discrete emotion theorists, who suggest that these emotions are universally displayed and recognized from facial expressions [89]. In sign judgment approaches [24], a widely used method for manual labeling of facial actions is the Facial Action Coding System (FACS) [48].

FACS associates facial expression changes with actions of the muscles that produce them. It defines 9 different Action Units (AUs) in the upper face, 18 in the lower face, 11 for head position, 9 for eye position, and 14 additional descriptors for miscellaneous actions. AUs are considered to be the smallest visually discernable facial movements. Using FACS, human coders can manually code nearly any anatomically possible facial expression, decomposing it into the specific AUs that produced the expression. As AUs are independent of interpretation, they can be used for any higher order decision making process including recognition of basic emotions (EMFACS; see [48]), cognitive states like interest and puzzlement [32], psychological states like suicidal depression [50] or pain [212], social behaviours like accord and rapport [8][32], personality traits like extraversion and temperament [50], and social signals like status, trustworthiness, emblems (i.e., culture-specific interactive signals like wink), regulators (i.e., conversational mediators like nod and gaze exchange), and illustrators (i.e., cues accompanying speech like raised eyebrows) [8][46][47]. FACS provides an objective and comprehensive language for describing facial expressions and relating them back to what is known about their meaning

from the behavioral science literature. Because it is comprehensive, FACS also allows for the discovery of new patterns related to emotional or situational states. For example, what are the facial behaviors associated with social signals such as empathy, persuasion, and politeness? An example where subjective judgments of expression failed to find relationships which were later found with FACS is the failure of naive subjects to differentiate deception and intoxication from facial display, whereas reliable differences were shown with FACS [165]. Research based upon FACS has also shown that facial actions can show differences between those telling the truth and lying at a much higher accuracy level than naive subjects making subjective judgments of the same faces [56]. Exhaustive overview of studies on facial and gaze behaviour using FACS can be found in [50].

2.4 Vocal Behaviour

The vocal nonverbal behaviour includes all spoken cues that surroud the verbal message and influence its actual meaning. The effect of vocal nonverbal behaviour is particularly evident when the tone of a message is ironic. In this case the face value of the words is changed into its opposite by just using the appropriate vocal intonation. The vocal nonverbal behaviour includes five major components: *voice quality*, *linguistic* and *non-linguistic vocalizations*, *silences*, and *turn-taking patterns*. Each one of them relates to social signals that contribute to different aspects of the social perception of a message.

The *voice quality* corresponds to the prosodic features, i.e., pitch, tempo, and energy (see Section 3.3.4 for more details) and, in perceptual terms, accounts for *how* something is said [31]. The prosody conveys a wide spectrum of socially relevant cues: emotions like anger or fear are often accompanied by energy bursts in voice (shouts) [168], the pitch influences the perception of dominance and extroversion (in general it is a personality marker [167]), the speaking fluency (typically corresponding to high rythm and lack of hesitations) increases the perception of competence and results into higher persuasiveness [167]. The vocalizations include also effects that aim at giving particular value to certain utterances or parts of the discourse, e.g., the pitch accents (sudden increases of the pitch to underline a word) [79], or changes in rhythm and energy aiming at structuring the discourse [80].

The *linguistic vocalizations* (also known as *segregates*) include all the non-words that are typically used as if they were actual words, e.g., "*ehm*","*ah-ah*", "*uhm*", etc. Segregates have two main functions, the first is to replace words that for some reason cannot be found, e.g., when people do not know how to answer a question and simply utter a prolonged "*ehm*". They are often referred to as *disfluencies* and often account for a situation of embarassment or

difficulty with respect to a social interaction [64]. The second important function is the so-called *back-channeling*, i.e., the use of segregates to accompany someone else speaking. In this sense they can express attention, agreement, wonder, as well as the attempt of grabbing the floor or contradicting [176].

The *non-linguistic vocalizations*, also known as vocal outbursts, include nonverbal sounds like laughing, sobbing, crying, whispering, groaning, and similar, that may or may not accompany words, and provide some information about the attitude towards social situations. For instance, laughter tends to reward desirable social behaviour [90] and shows affiliation efforts, while crying is often involved in *mirroring* (also known as *chameleon effect* [22]), that is in the mutual imitation of people connected by strong social bonds [91]. Also, research in psychology has shown that listeners tend to be accurate in decoding some basic emotions as well as some non-basic affective and social signals such as distress, anxiety, boredom, and sexual interest from vocal outbursts like laughs, yawns, coughs, and sighs [163].

The silence is often interpreted as simple non-speech, but actually plays a major role in the vocal behaviour [219]. There are three kinds of silence in speech: *hesitation silence*, *psycholinguistic silence*, and *interactive silence* [158]. The first takes place when a speaker has difficults in talking, e.g., because she is expressing a difficult concept or must face a hostile attitude in listeners. Sometimes, hesitation silences give rise to segregates that are used to *fill* the silence space (hence segregates are called sometimes fillers). The psycholinguistic silences take place when the speaker needs time to encode or decode the speech. This kind of silences happen often at the beginning of an intervention because the speaker needs to think about the next words. In this sense, this is often a sign of difficulty and problems in dealing with a conversation. The interactive silences aim at conveying messages about the interactions taking place: silence can be a sign of respect for people we want to listen to, a way of ignoring persons we do not want to answer to, as well as a way to attract the attention to other forms of communication like mutual gaze or facial expressions.

Another important aspect of vocal nonverbal behaviour is turn-taking [154]. This includes two main components: the regulation of the conversations, and the coordination (or the lack of it) during the speaker transitions. The regulation in conversations includes behaviours aimed at maintaining, yielding, denying, or requesting the turn. Both gaze and voice quality (e.g. coughing) are used to signal *transition relevant points* [217]. When it comes to vocal nonverbal cues as conversation regulators, specific pitch and energy patterns show the intention of yielding the turn rather than maintaining the floor. Also, linguistic vocalizations (see above) are often used as a form of back-channeling to request the turn. The second important aspect in turn-taking is the coordination at the speaker transitions [20]. Conversations where the latency times between turns are too long sound typically awkward. The reason is that in flu-

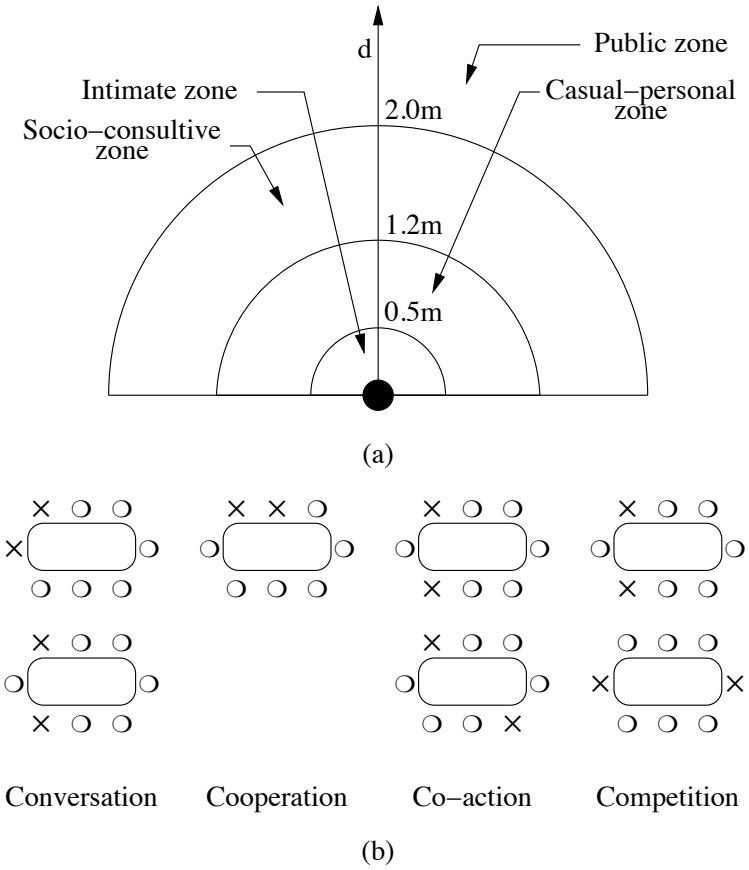


Fig. 5. Space and seating. The upper part of the figure shows the concentric zones around each individual associated to different kinds of rapport (d stands for distance). The lower part of the figure shows the preferred seating arrangements for different kinds of social interactions.

ent conversations, the mutual attention reduces the above phenomenon and results into synchronized speaker changes, where the interactants effectively interpret the signals aimed at maintaining or yielding their turns. Overlapping speech is another important phenomenon that accounts for disputes as well as status and dominance displays [180]. Note, however, that the amount of overlapping speech accounts for up to 10% of the total time even in normal conversations [175].

2.5 Space and Environment

The kind and quality of the relationships between individuals influences their interpersonal distance (the physical space between them). One of the most common classifications of mutual distances between individuals suggests the existence of four concentric zones around a person accounting for different kinds of relationships with the others [77]: the *intimate zone*, the *casual-*

personal zone, the *socio-consultive zone* and the *public zone* (see Figure 5a).

The *intimate zone* is the innermost region and it is open only to the closest family members and friends. Its dimension, like in the case of the other zones, depends on the culture and, in the case of western Europe and United States, the intimate zone corresponds to a distance of 0.4-0.5 meters. In some cases, e.g., crowded buses or elevators, the intimate zone must be necessarily opened to strangers. However, whenever there is enough space, people tend to avoid entering the intimate zone of others. The *casual-personal zone* ranges (at least in USA and Western Europe) between 0.5 and 1.2 meters and it typically includes people we are most familiar with (colleagues, friends, etc.). To open such an area to another person in absence of constraints is a major signal of friendship. The *socio-consultive* distance is roughly between 1 and 2 meters (again in USA and Western Europe) and it is the area of formal relationships. Not surprisingly, professionals (lawyers, doctors, etc.) typically receive their clients sitting behind desks that have a profundity of around 1 meter, so that the distance with respect to their clients is in the range corresponding to the socio-consultive zone. The *public zone* is beyond 2 meters distance and it is, in general, outside the reach of interaction potential. In fact, any exchange taking place at such a distance is typically due to the presence of some obstacle, e.g., a large meeting table that requires people to talk at distance.

Social interactions take place in environments that influence behaviours and perceptions of people with their characteristics. One of the most studied environmental variables is the seating arrangement, i.e., the way people take place around a table for different purposes [96][158]. Figure 5b shows the seating positions that people tend to use to perform different kinds of tasks (the circles are the empty seats) [164]. The seating position depends also on the personality of people: dominant and higher status individuals tend to seat at the shorter side of rectangular tables, or in the middle of the longer sides (both positions ensure high visibility and make easier the control of the conversation flow) [106]. Moreover, extrovert people tend to privilege seating arrangements that minimize interpersonal distances, while introvert ones do the opposite [164].

3 The State of the Art

The problem of machine analysis of human social signals includes four subproblem areas (see Figure 5):

(1) recording the scene,
(2) detecting people in it,
(3) extracting audio and/or visual behavioural cues displayed by people de-

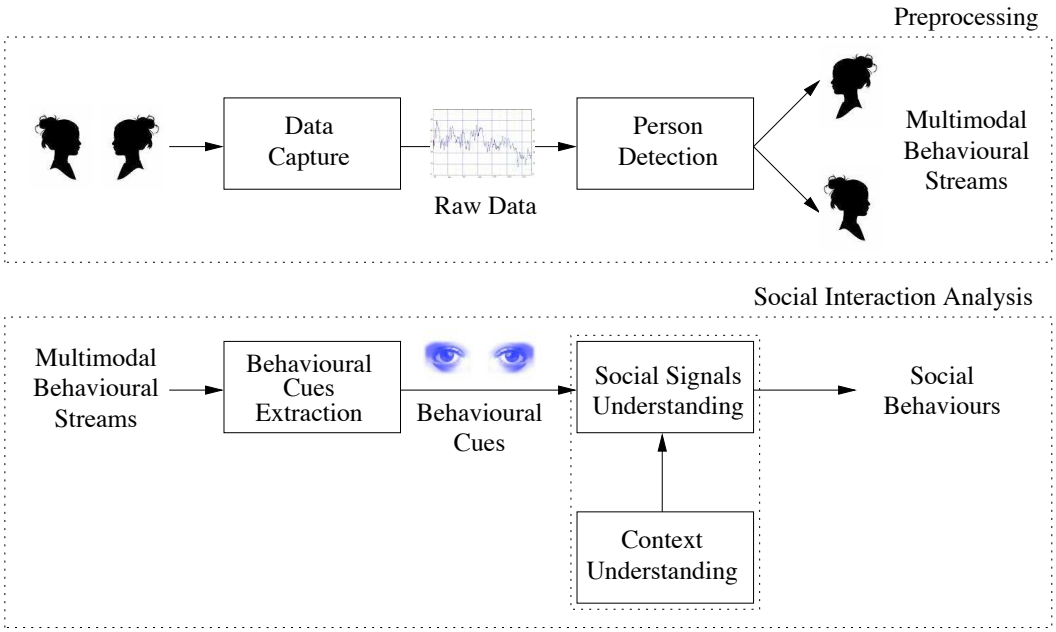


Fig. 6. Machine analysis of social signals and behaviours: a general scheme. The process includes two main stages: The *preprocessing*, takes as input the recordings of social interaction and gives as output the multimodal behavioural streams associated with each person. The *social interaction analysis* maps the multimodal behavioural streams into social signals and social behaviours.

tected in the scene and interpreting this information in terms of social signals conveyed by the observed behavioural cues,

(4) sensing the context in which the scene is recorded and classifying detected social signals into the target social-behaviour-interpretative categories in a context-sensitive manner.

The survey of the past work is divided further into four parts, each of which is dedicated to the efforts in one of the above-listed subproblem areas.

3.1 Data Capture

Data capture refers to using sensors of different kinds to capture and record social interactions taking place in real world scenarios. The choice of the sensors and their arrangement in a specific recording setup determine the rest of the SSP process and limit the spectrum of behavioral cues that can be extracted. For example, no gaze behavior analysis can be performed, if appropriate detectors are not included in the capture system.

The most common sensors are microphones and cameras and they can be arranged in structures of increasing complexity: from a single camera and/

or microphone to capture simple events like oral presentations [201], to fully equipped *smart meeting rooms* where several tens of audio and video channels (including microphone arrays, fisheye cameras, lapel microphones, etc.) are setup and synchronized to capture complex interactions taking place in a group meeting [110][205]. The literature shows also examples of less common sensors such as cellular phones or smart badges equipped with proximity detectors and vocal activity measurement devices [43][144], and systems for the measurement of physiological activity indicators such as blood pressure and skin conductivity [76]. Recent efforts have tried to investigate the neurological basis of social interactions [2] through devices like *functional Magnetic Resonance Imaging* (fMRI) [119], and *Electroencephalography* (EEG) signals [193].

The main challenges in human sensing research domain are *privacy* and *passiveness*. The former involves ethical issues to be addressed when people are recorded during spontaneous social interactions. This subject is outside the scope of this paper, but the *informed consent priciple* [51] should be always respected meaning that human subjects should always be aware of being recorded (e.g., like in broadcast material). Also, the subjects need to authorize explicitly the use and the diffusion of the data and they must have the right of deleting, partially or totally, the recordings where they are portrayed.

The second challenge relates to creating capture systems that are *passive* [125], i.e., unintrusive changing the behaviour of the recorded individuals as little as possible (in principle, the subjects should not even realize that they are recorded). This is a non-trivial problem because passive systems should involve only non-invasive sensors and the output of these is, in general, more difficult to process effectively. On the other hand, data captured by more invasive sensors are easier to process, but at the same time such recording setups tend to change the behaviour of the recorded individuals. Recording human naturalistic behaviour while eliciting specific behaviours and retaining the naturalism/ spontaneity of the behaviour is a very difficult problem tackled recently by several research groups [29][135].

3.2 Person Detection

The sensors used for data capture output signals that can be analyzed automatically to extract the behavioural cues underlying social signals and behaviours. In some cases, the signals corresponding to different individuals are separated at the origin. For example, physiological signals are recorded by invasive devices physically connected (e.g. through electrodes) to each person. Thus, the resulting signals can be attributed withouth ambiguity to a given individual. However, it happens more frequently that the signals contain spurious information (e.g. background noise), or they involve more than

one individual. This is the case of the most commonly used sensors, microphones and cameras, and it requires the application of algorithms for *person detection* capable of isolating the signal segments corresponding to a single individual. The rest of this section discusses how this can be done for multiparty audio and video recordings.

3.2.1 Person Detection in Multiparty Audio Recordings

In the case of audio recordings, person detection is called *speaker segmentation* or *speaker diarization* and consists of splitting the speech recordings into intervals corresponding to a single voice, recognizing automatically *who talks when* (see [189] for an extensive survey). The speaker diarization is the most general case and it includes three main stages: the first is the segmentaion of the data into speech and non-speech segments, the second is the detection of the speaker transitions, and the third is the so-called *clustering*, i.e. the grouping of speaker segments corresponding to a single individual (i.e. to a single voice). In some cases (e.g. broadcast data), no silences are expected between one speaker and the following, thus the first step is not necessary. Systems that do not include a speech/ non-speech segmentation are typically referred to as *speaker segmentation* systems.

Speech and non-speech segmentation is typically performed using machine learning algorithms trained over different audio classes represented in the data (non-speech can include music, background noises, silence, etc.). Typically used techniques include Artificial Neural Networks [5], k Nearest Neighbours [107], Gaussian Mixture Models [61], etc. Most commonly used features include the basic information that can be extracted from any signal (e.g. energy and autocorrelation [156]), as well as the features typically extracted for speech recognition like *Mel Frequency Cepstrum Coefficients* (MFCC), *Linear Predictive Coding* (LPC), etc. (see [84] for an extensive survey).

The detection of the speaker transitions is performed by splitting the speech segments into short intervals (e.g. $2 - 3$ seconds) and by measuring the *difference* (see below) between two consecutive intervals: the highest values of the difference correspond to the speaker changes. The approaches is based on the assumption that the data include at least two speakers. If this is not the case, simple differences in the intonation or the background noise might be detected as speaker changes. The way the difference is estimated allows one to distinguish between the different approaches to the task: in general each interval is modeled using a single Gaussian (preferred to the GMMs because it simplifies the calculations) and the difference is estimated with the symmetric Kullback-Leibler divergence [14]. Alternative approaches [157] use a penalized-likelihood-ratio test to verify whether a single interval is modeled better by a single Gaussian (no speaker change) or by two Gaussians (speaker

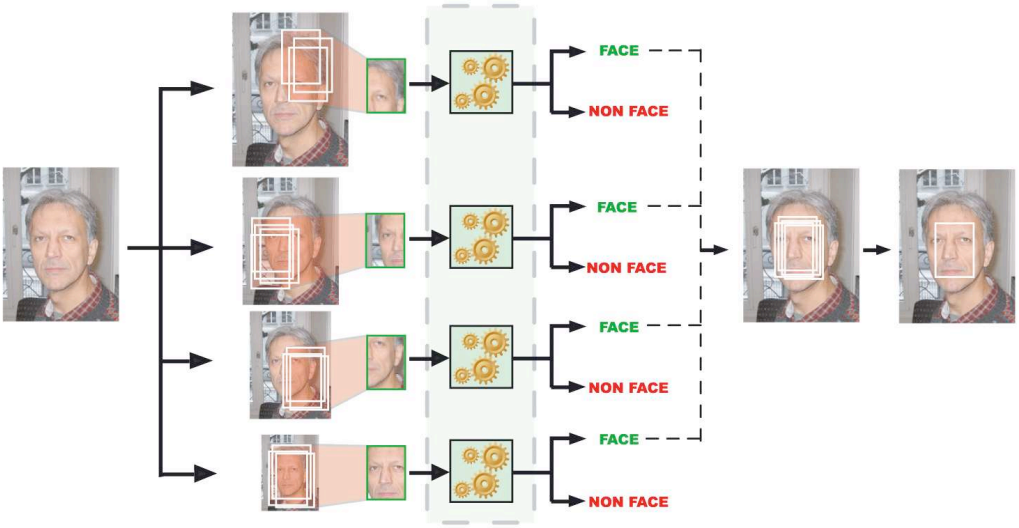


Fig. 7. Face detection. General scheme of an appearance based approach for face detection (picture from "*A tutorial on face detection and recognition*", by S.Marcel, www.idiap.ch/∼marcel).

change).

The last step of both speaker diarization and segmentation is *clustering*, i.e. grouping of the segments corresponding to a single voice into a unique cluster. This is commonly carried out through iterative approaches [14][117][157] where the clusters are initialized using the intervals between the speaker changes detected at the previous step (each interval is converted into a set of feature vectors using common speech processing techniques [84][156]), and then they are iteratively merged based on the similarity of the models used to represent them (single Gaussians or GMMs). The merging process is stopped when a criterion (e.g. the total likelihood of the cluster models starts to decrease) is met.

Most recent approaches tend to integrate three steps above-mentioned into a single framework by using hidden Markov models or dynamic Bayesian networks that align feature vectors extracted at regular time steps (e.g. 30 *ms*) and sequences of states corresponding to speakers in an unsupervised way [4][5].

3.2.2 Person Detection in Multiparty Video Recordings

In the case of video data, the person detection consists in locating faces or full human figures (that must be eventually tracked). Face detection is typically the first step towards facial expression analysis [139] or gaze behavior analy-

sis [199] (see [81][215] for extensive surveys on face detection techniques). The detection of full human figures is more frequent in surveillance systems where the only important information is the movement of people across wide public spaces (e.g. train stations or streets) [62][115]. In the SSP framework, the detection of full human figures can be applied to study social signals related to space and distances (see Section 2.5), but to the best of our knowledge no attempts have been made yet in this direction.

Early approaches to face detection (see e.g. [161][181]) were based on the hypothesis that the presence of a face can be inferred from the pixel values. Thus they apply classifiers like Neural Networks or Support Vector Machines directly over small portions of the video frames (e.g. patches of 20×20 pixels) and map them into a face/ non-face classes. The main limitation of such techniques is that it is difficult to train classifiers for a *non-face* class that can include any kind of visual information (see Figure 7). Other approaches (e.g. [82][58]) try to detect human skin areas in images and then use their spatial distribution to identify faces and facial features (eyes, mouth and nose). The skin areas are detected by clustering the pixels in the color space. Alternative approaches (e.g. [101]) detect separately individual face elements (eyes, nose and mouth) and detect a face where such elements have the appropriate relative positions. These approaches are particularly robust to rotations because they depend on the relative position of face elements, rather than on the orientation with respect to a general reference frame in the image.

Another method that can handle out-of-plane head motions is the statistical method for 3D object detection proposed in [169]. Other such methods, which have been recently proposed, include those in [83][207]. Most of these methods emphasize statistical learning techniques and use appearance features. Arguably the most commonly employed face detector in automatic facial expression analysis is the real-time face detector proposed in [204]. This detector consists of a cascade of classifiers trained by AdaBoost. Each classifier employs integral image filters, also called "box filters," which are reminiscent of Haar Basis functions, and can be computed very fast at any location and scale. This is essential to the speed of the detector. For each stage in the cascade, a subset of features is chosen using a feature selection procedure based on AdaBoost. There are several adapted versions of the face detector described in [204] and the one that is often used is that proposed in [52].

The main challenge in detecting human figures is that people wear clothes of different color and appearance, so the pixel values are not a reliable feature for human body detection (see Figure 8). For this reason, some approaches extract features like the histograms of the edge directions (e.g. [34][223]) from local regions of the images (typically arranged in a regular grid), and then make a decision using classifiers like the Support Vector Machines. The same approach can be improved in the case of the videos, by adding motion information

Fig. 8. People detection. Examples of people detection in public spaces (pictures from [216]).

extracted using the optical flow [35]. Other approaches (e.g. [114][194]) try to detect individual body parts and then use general rules of human body anatomy to reason about the body pose (individual body parts have always the same shape and they have the same relative position). For exhaustive survey, see [153].

3.3 Social Signals Detection

Once people in the observed scene are detected, the next step in the SSP process is to extract behavioural cues displayed by these people. Those cues include one or more synchronized audio and/or video signals that convey the information about the behaviour of the person. They are the actual source from which socially-relevant behavioural cues are extracted. The next sections discuss the main approaches to social signals detection from audio and/or visual signals captured while monitoring a person.

3.3.1 Detection of Social Signals from Physical Appearance

To the best of our knowledge, only few works address the problem of analyzing the physical appearance of people. However, these works do not aim to interpret this information in terms of social signals. Some approaches have tried to measure automatically the beauty of faces [1][44][73][75][211]. The work in [1] detects separately the face elements (eyes, lips, etc.) and then maps the ratios between their dimensions and distances into beauty judgments through classifiers trained on images assessed by humans. The work in [44] models the symmetry and the proportions of a face through the geometry of several landmarks (e.g. the corners of the eyes and the tip of the nose), and then applies machine learning techniques to match human judgments. Other techniques (e.g., [131]) use 3D models of human heads and the distance with respect to

average faces extracted from large data sets to assess personal beauty. Faces closest to the average seem to be judged as more attractive than others.

Also few works were proposed where the body shape, the color of skin, hair, and clothes are extracted automatically (through a clustering of the pixels in the color space) for identification and tracking purposes [16][36][214]. However these works do not address social signal understanding and are therefore out of the scope of this paper.

3.3.2 Detection of Social Signals from Gesture and Posture

Gesture recognition is an active research domain in computer vision and pattern recognition research communities, but no efforts have been made, so far, to interpret the social information carried by gestural behaviours. In fact, the efforts are directed mostly towards the use of gestures as an alternative to keyboard and mouse to operate computers (e.g., [132][172][213]), or to the automatic reading of sign languages (e.g., [40][97]). Also few efforts have been reported towards human affect recognition from body gestures (for an overview see [76][221]). There are two main challenges in recognizing gestures: detecting the body parts involved in the gesture (in general the hands), and modeling the temporal dynamic of the gesture.

The first problem is addressed by selecting appropriate visual features: these include, e.g., histograms of oriented gradients (e.g., [183][184]), optical flow (e.g., [3][188]), spatio-temporal salient points (e.g., [129]) and space-time volumes (e.g., [67]). The second problem is addressed by using techniques such as Dynamic Time Warping (e.g., [129]), Hidden Markov Models (e.g. [3]), and Conditional Random Fields (e.g., [179]).

Like in the case of gestures, machine recognition of walking style (or gait) has been investigated as well, but only for purposes different from SSP, namely recognition and identification in biometric applications [100][102][206]. The common approach is to segment the silhouette of the human body into individual components (legs, arms, trunk, etc.), and then to represent their geometry during walking through vectors of distances [206], symmetry operators [78], geometric features of body and stride (e.g. distance between head and feets or pelvis) [17], etc.

Also automatic posture recognition has been addressed in few works, mostly aiming at surveillance [57] and activity recognition [206] (See [54][116][153] for extensive overviews of the past work in the field). However, there are few works where the posture is recognized as a social signal, namely to estimate the interest level of children learning to use computers [124], to recognize the affective state of people [38][74] (see [76][221] for exhaustive overview of research efforts in the field), and the influence of culture on affective postures [95].

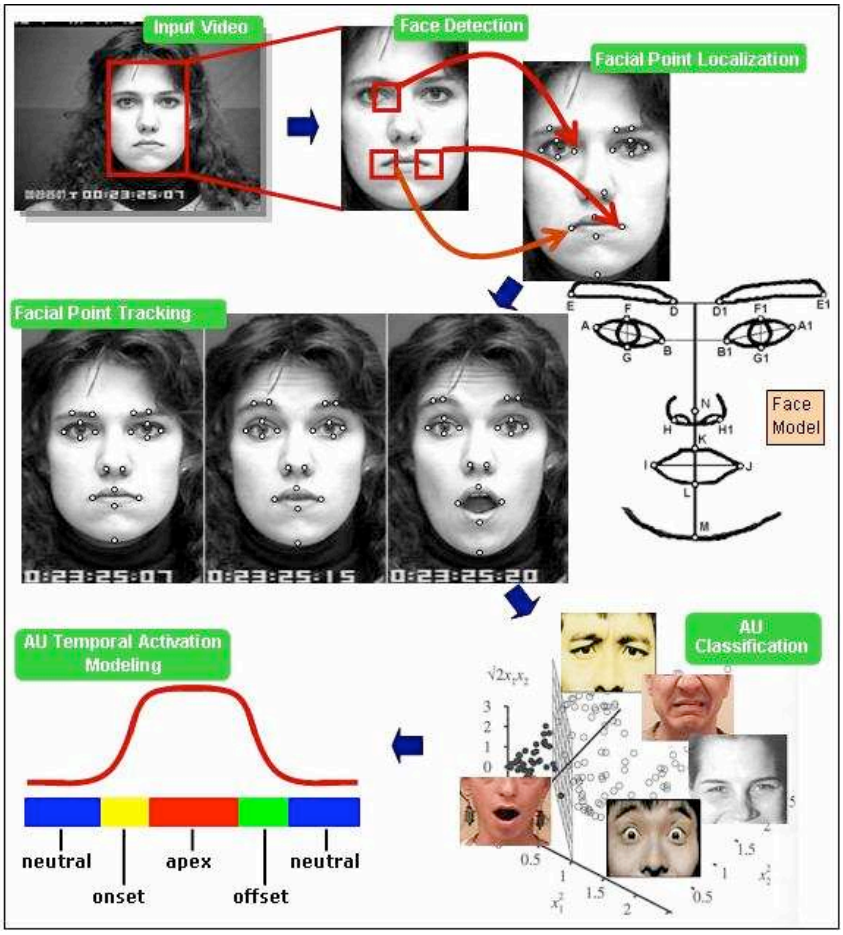


Fig. 9. AU detection. Outline of a geometric-feature-based system for detection of facial AUs and their temporal phases (onset, apex, offset, neutral) proposed in [196].

3.3.3 Detection of Social Signals from Gaze and Face

The problem of machine recognition of human gaze and facial behaviour includes three subproblem areas (see Figure 9): finding faces in the scene, extracting facial features from the detected face region, analyzing the motion of eyes and other facial features and/or the changes in the appearance of facial features, and classifying this information into some facial-behaviour-interpretative categories (e.g., facial muscle actions (AUs), emotions, social behaviours, etc.).

Numerous techniques have been developed for face detection, i.e., identification of all regions in the scene that contain a human face (see Section 3.2). Most of the proposed approaches to facial expression recognition are directed toward static, analytic, 2-D facial feature extraction [135][185]. The usually extracted facial features are either geometric features such as the shapes of the facial components (eyes, mouth, etc.) and the locations of facial fiducial points (corners of the eyes, mouth, etc.) or appearance features represent-

ing the texture of the facial skin in specific facial areas including wrinkles, bulges, and furrows. Appearance-based features include learned image filters from Independent Component Analysis (ICA), Principal Component Analysis (PCA), Local Feature Analysis (LFA), Gabor filters, integral image filters (also known as box-filters and Haar-like filters), features based on edge-oriented histograms, and similar [135]. Several efforts have been also reported which use both geometric and appearance features (e.g. [185]). These approaches to automatic facial expression analysis are referred to as hybrid methods. Although it has been reported that methods based on geometric features are often outperformed by those based on appearance features using, e.g., Gabor wavelets or eigenfaces, recent studies show that in some cases geometric features can outperform appearance-based ones [135][136]. Yet, it seems that using both geometric and appearance features might be the best choice in the case of certain facial expressions [136].

Contractions of facial muscles (i.e., AUs explained in section 2.3), which produce facial expressions, induce movements of the facial skin and changes in the location and/or appearance of facial features. Such changes can be detected by analyzing optical flow, facial-point- or facial-component-contour-tracking results, or by using an ensemble of classifiers trained to make decisions about the presence of certain changes based on the passed appearance features. The optical flow approach to describing face motion has the advantage of not requiring a facial feature extraction stage of processing. Dense flow information is available throughout the entire facial area, regardless of the existence of facial components, even in the areas of smooth texture such as the cheeks and the forehead. Because optical flow is the visible result of movement and is expressed in terms of velocity, it can be used to represent directly facial expressions. Many researchers adopted this approach (for overviews, see [135][139][185]). Until recently, standard optical flow techniques were arguably most commonly used for tracking facial characteristic points and contours as well. In order to address the limitations inherent in optical flow techniques such as the accumulation of error and the sensitivity to noise, occlusion, clutter, and changes in illumination, recent efforts in automatic facial expression recognition use sequential state estimation techniques (such as Kalman filter and Particle filter) to track facial feature points in image sequences [135][136][222].

Eventually, dense flow information, tracked movements of facial characteristic points, tracked changes in contours of facial components, and/or extracted appearance features are translated into a description of the displayed facial behaviour. This description (facial expression interpretation) is usually given either in terms of shown affective states (emotions) or in terms of activated facial muscles (AUs) underlying the displayed facial behaviour. Most facial expressions analyzers developed so far target human facial affect analysis and attempt to recognize a small set of prototypic emotional facial expressions like happiness and anger [140][221]. However, several promising prototype systems

were reported that can recognize deliberately produced AUs in face images (for overviews, see [135][185]) and even few attempts towards recognition of spontaneously displayed AUs (e.g., [103][108]) and towards automatic discrimination between spontaneous and posed facial behaviour such as smiles [195], frowns [197], and pain [104], have been recently reported as well. Although still tentative, few studies have also been recently reported on separating emotional states from non-emotional states and on recognition of non-basic affective states in visual and audiovisual recordings of spontaneous human behaviour(e.g., for overview see [170][220]). However, although messages conveyed by AUs like winks, blinks, frowns, smiles, gaze exchanges, etc., can be interpreted in terms of social signals like turn taking, mirroring, empathy, engagement, etc., no efforts have been reported so far on automatic recognition of social behaviours in recordings of spontaneous facial behaviour. Hence, while the focus of the research in the field started to shift to automatic (non-basic-) emotion and AU recognition in spontaneous facial expressions (produced in a reflex-like manner), efforts towards automatic analysis of human social behaviour from visual and audiovisual recordings of human spontaneous behaviour are still to be made.

While the older methods for facial behaviour analysis employ simple approaches including expert rules and machine learning methods such as neural networks to classify the relevant information from the input data into some facial-expression-interpretative categories (e.g., basic emotion categories), the more recent (and often more advanced) methods employ probabilistic, statistical, and ensemble learning techniques, which seem to be particularly suitable for automatic facial expression recognition from face image sequences (for comprehensive overviews of the efforts in the field, see [135][221]). Note, however, the present systems for facial expression analysis typically depend on accurate head, face and facial feature tracking as input and are still very limited in performance and robustness.

3.3.4 Detection of Social Signals from Vocal Behaviour

The behavioural cues in speech include voice quality, vocalizations (linguistic and non-linguistic), and silences (see Section 2.4 for details). All of them have been the subject of extensive research in speech, but they have rarely been interpreted in terms of social information, even if they account for roughly 50% of the total time in spontaneous conversations [21]. With few exceptions, the detection of vocal behaviour has aimed at the improvement of Automatic Speech Recognition (ASR) systems, where the vocal non-verbal behaviour represents a form of noise rather than an information.

The voice quality corresponds to the prosody and includes three major aspects, often called the *Big Three*: *pitch*, *tempo* and *energy* [31]. The pitch is

the frequency of oscillation of the vocal folds during the emission of voice and it is the characteristic that alone contributes more than anything else to the sound of a voice [120][150]. The measurement of the pitch, often called *fundamental frequency* (or $F0$) because most of the speech energy is concentrated over components corresponding to its integer multiples, can be performed with several standard methods proposed in the literature [84][156]. The pitch is typically extracted as the frequency corresponding to the first peak of the Fourier Transform of short analysis windows (in general 30 ms). Several tools publicly available on the web, e.g. *Wavesurfer* [1] [177] and *Praat* [2] [18], implement algorithms extracting the pitch from speech recordings. The tempo is typically estimated through the speaking rate, i.e. the number of phonetically relevant units, e.g. vowels [149], per second. Other methods are based on measures extracted from the speech signal like the first spectral moment of the energy [121][122] and typically aim at improving speech recognition systems through speaking rate adaptation. The energy, is a property of any digital signal and simply corresponds to the sum of the square values of the samples [156].

No major efforts have been made so far, to the best of our knowledge, to detect the non-linguistic vocalizations (see Section 2.4). The only exceptions are laughter [92][191][192] due to its ubiquitous presence in social interactions, and crying [118][134]. Laugther is detected by applying binary classifiers such as Support Vector Machines to features commonly applied in speech recognition like the *Mel Frequency Cepstral Coefficients* [92], or by modeling *Perceptual Linear Prediction* features with Gaussian Mixture Models and Neural Networks [191][192]. These efforts are based only on audio signals, but few pioneering efforts towards audiovisual recognition of non-linguistic vocal outbursts have been recently reported. A laughter detector which combines the outputs of an audio-based detector that uses MFCC audio features and a visual detector that uses spatial locations of facial feature points is proposed in [86]. They attained 80% average recall rate using 3 sequences of 3 subjects in a person dependent way. In [147] decision level and feature level fusion with audio- and video-only laughter detection are compared. The work uses PLP features and displacements of the tracked facial points as the audio and visual features respectively. Both fusion approaches outperformed single-modal detectors, achieving on average 84% recall in a person-independent test. Extension of this work based on utilisation of temporal features has been reported in [148].

Linguistic vocalizations have been investigated to detect hesitations in spontaneous speech [105][173][174] with the main purpose of improving speech recognition systems. The disfluencies are typically detected by mapping acous-

[1] Publicly available at `http://www.speech.kth.se/wavesurfer/`.
[2] Publicly available at `http://www.praat.org`.

tic observations (e.g. pitch and energy) into classes of interest with classifiers like Neural Networks or Support Vector Machines. The detection of silence is one of the earliest tasks studied in speech analysis and robust algorithms, based on the distribution of the energy, have been developed since the earliest times of digital signal processing [155][156]. Another important aspect of vocal behaviour, i.e. the turn taking, is typically a side-product of the speaker diarization or segmentation step (see Section 3.2).

3.3.5 Detection of Social Signals in Space and Environment

Physical proximity information has been used in *reality mining* applications (see Section 4) as a social cue accounting for the simple presence or absence of interaction between people [43][144]. These works use special cellular phones equipped to sense the presence of similar devices in the vicinity. Automatic detection of seating arrangements has been proposed as a cue for retrieving meeting recordings in [88]. Also, several video-surveillance approaches developed to track people across public spaces can potentially be used for detection of social signals related to the use of the available space (see Section 3.2 for more details).

3.4 Context Sensing and Social Behaviour Understanding

Context plays a crucial role in understanding of human behavioural signals, since they are easily misinterpreted if the information about the situation in which the shown behavioural cues have been displayed is not taken into account. For example, a smile can be a display of politeness (social signal), contentedness (affective cue), joy of seeing a friend (affective cue/ social signal), irony/ irritation (affective cue/ social signal), empathy (emotional response/ social signal), greeting (social signal), to mention just a few possibilities. It is obvious from these examples that in order to determine the communicative intention conveyed by an observed behavioural cue, one must know the context in which the observed signal has been displayed: where the expresser is (outside, inside, in the car, in the kitchen, etc.), what his or her current task is, are other people involved, when the signal has been displayed (i.e., what is the timing of displayed behavioural signals with respect to changes in the environment), and who the expresser is (i.e., it is not probable that each of us will express a particular affective state by modulating the same communicative signals in the same way).

Note, however, that while W4 (*where*, *what*, *when*, *who*) is dealing only with the apparent perceptual aspect of the context in which the observed human behaviour is shown, human behaviour understanding is about W5+ (*where*, *what*,

when, *who*, *why*, *how*), where the *why* and *how* are directly related to recognizing communicative intention including social behaviours, affective and cognitive states of the observed person. Hence, SSP is about W5+. However, since the problem of context-sensing is extremely difficult to solve, especially for a general case (i.e., general-purpose W4 technology does not exist yet [138][137]), answering the *why* and *how* questions in a W4-context-sensitive manner when analysing human behaviour is virtually unexplored area of research. Having said that, it is not a surprise that most of the present approaches to machine analysis of human behaviour are neither context-sensitive nor suitable for handling longer time scales. Hence, the focus of future research efforts in the field should be primarily on tackling the problem of context-constrained analysis of multimodal social signals shown over longer temporal intervals. Here, we would like to stress the importance of two issues: realizing temporal analysis of social signals and achieving temporal multimodal data fusion.

Temporal dynamics of social behavioural cues (i.e., their timing, co-occurrence, speed, etc.) are crucial for the interpretation of the observed social behaviour [8] [50]. However, present methods for human behaviour analysis do not address the *when* context question - dynamics of displayed behavioural signals is usually not taken into account when analyzing the observed behaviour, let alone analysing the timing of displayed behavioural signals with respect to changes in the environment. Exceptions of this rule include few recent studies on modelling semantic and temporal relationships between facial gestures (i.e., AUs, see Section 2.3) forming a facial expression (e.g. [187]), few studies on discrimination between spontaneous and posed facial gestures like brow actions and smiles based on temporal dynamics of target facial gestures, head and shoulder gestures [195][197], and few studies on multimodal analysis of audio and visual dynamic behaviours for emotion recognition [221]. In general, as already mentioned above, present methods cannot handle longer time scales, model grammars of observed persons behaviours, and take temporal and context-dependent evolvement of observations into account for more robust performance. These remain major challenges facing the researchers in the field.

Social signals are spoken and wordless messages like head nods, winks, *uh* and *yeah* utterances, which are sent by means of body gestures and postures, facial expressions and gaze, vocal expressions and speech. Hence, automated analyzers of human social signals and social behaviours should be multimodal, fusing and analyzing verbal and non-verbal interactive signals coming from different modalities (speech, body gestures, facial and vocal expressions). Most of the present audiovisual and multimodal systems in the field perform decision-level data fusion (i.e., classifier fusion) in which the input coming from each modality is modelled independently and these single-modal recognition results are combined at the end. Since humans display audio and visual expressions in a complementary and redundant manner, the assumption of conditional independence between audio and visual data streams in decision-level fusion

is incorrect and results in the loss of information of mutual correlation between the two modalities. To address this problem, a number of model-level fusion methods have been proposed that aim at making use of the correlation between audio and visual data streams, and relax the requirement of synchronization of these streams(e.g., [55][220]). However, how to model multimodal fusion on multiple time scales and how to model temporal correlations within and between different modalities is largely unexplored. A much broader focus on the issues relevant to multimodal temporal fusion is needed including the optimal level of integrating these different streams, the optimal function for the integration, and how estimations of reliability of each stream can be included in the inference process. In addition, how to build context-dependent multimodal fusion is another open and highly relevant issue.

4 Main Applications of Social Signal Processing

The expression *Social Signal Processing* has been used for the first time in [145] to group under a collective definition several pioneering works of Alex Pentland and his group at MIT. Some of their works [142][143] extracted automatically the social signals detected in dyadic interactions to predict with an accuracy of more than 70% the outcome of salary negotiations, hiring interviews, and speed-dating conversations [33]. These works are based on vocal social signals including overall *activity* (the total amount of energy in the speech signals), *influence* (the statistical influence of one person on the speaking patterns of the others), *consistency* (stability of the speaking patterns of each person), and *mimicry* (the imitation between people involved in the interactions). Other works used cellular phones equipped with proximity sensors and vocal activity detectors to perform what came to be called *reality mining*, or *social sensing*, i.e., automatic analysis of everyday social interactions in groups of several tens of individuals [43][144]. Individuals are represented through vectors accounting for their proximity with others and for the places they are (home, work, etc.). The application of the Principal Component Analysis to such vectors leads to the so called *eigenbehaviours* [43].

In approximately the same period, few other groups worked on the analysis of social interactions in multimedia recordings targeting three main areas: analysis of interactions in small groups, recognition of roles, and sensing of users interest in computer characters. Results for problems that have been addressed by more than one group are reported in Table 2.

The research on interactions in small groups has focused on the detection of dominant persons and on the recognition of collective actions. The problem of dominance is addressed in [85][160], where multimodal approaches combine several nonverbal features, mainly speaking energy and body movement,

Ref.	Data	Time	Source	Performance
Role Recognition				
[13]	Meetings (2 recordings, 3 roles)	0h.45m	acted	50.0% of segments (up to 60 seconds long) correctly classified
[15]	NIST TREC SDR Corpus (35 recordings, publicly available 3 roles)	17h.00m	spontaneous	80.0% of the news stories correctly labeled in terms of role
[42]	The Survival Corpus (11 recordings, publicly available, 5 roles)	4h.30m	acted	90% of precision in role assignment
[59]	AMI Meeting Corpus (138 recordings, publicly available, 4 roles)	45h.00m	acted	67.9% of the data time correctly labeled in terms of role
[200]	Radio news bulletins (96 recordings, 6 roles)	25h.00m	spontaneous	80% of the data time correctly labeled in terms of role
[210]	Movies (3 recordings , 4 roles)	5h.46m	spontaneous	95% of roles correctly assigned
[218]	The Survival Corpus (11 recordings, publicly available, 5 roles)	4h.30m	spontaneous	Up to 65% of analysis windows (around 10 seconds long) correctly classified in terms of role
Collective Action Recognition				
[39]	Meetings (30 recordings, publicly available)	2h.30m	acted	Action Error Rate of 12.5%
[111]	Meetings (60 recordings, publicly available)	5h.00m	acted	Action Error Rate of 8.9%
Interest Level Detection				
[60]	Meetings (50 recordings, 3 interest levels)	unknown	acted	75% Precision
[124]	Children playing with video games (10 recordings, 3 interest levels)	3h.20m	spontaneous	82% recognition rate

to identify at each moment who is the dominant individual. The same kind of features has been applied in [39][111] to recognize the actions performed in meetings like discussions, presentations, etc. In both above applications,

Ref.	Data	Time	Source	Performance
Dominance Detection				
[85]	Meetings from AMI Corpus (34 segments)	3h.00m	acted	Most dominant person correctly detected in 85% of segments
[159]	Meetings (8 meetings)	1h.35m	acted	Most dominant person correctly detected in 75% of meetings
[160]	Meetings (40 recordings)	20h.00m	acted	Most dominant person correctly detected in 60% of meetings

Table 2
Results obtained by Social Signal Processing works. For each work, information about the data (kind of interaction, availability, size, the total duration of the recordings), whether it is real-world or acted data, and the reported performance are summarized.

the combination of the information extracted from different modalities is performed with algorithms Dynamic Bayesian Networks [126] and layered Hidden Markov Models [130].

The recognition of roles has been addressed in two main contexts: broadcast material [15][53][200][210] and small scale meetings [13][42][59][218]. The works in [53][200][210] apply Social Network Analysis [209] to detect the role of people in broadcast news and movies, respectively. The social networks are extracted automatically using speaker adjacences in [53][200] (people are linked when they are adjacent in the sequence of the speakers), and face recognition [210] (people are linked when their faces appear together in a scene). The approach in [15] recognizes the roles of speakers in broadcast news using vocal behaviour (turn taking patterns and intervention duration) and lexical features. The recognition is performed using boosting techniques. The roles in meetings are recognized with a classifier tree applied to nonverbal behaviour features (overlapping speech, number of interventions, back-channeling, etc.) in the case of [13], while speech and fidgeting activity are fed to a multi-SVM classifier in [42][218]. A technique based on the combination of Social Network Analysis and lexical modeling (Boostexter) is presented in [59].

The reaction of users to social signals exhibited by computer characters has been investigated in several works showing that people tend to behave with Embodied Conversational Agents (ECA) as they behave with other humans. The effectiveness of computers as *social actors*, i.e., entities involved in the same kind of interactions as humans, has been explored in [127][128], where computers have been shown to be attributed a personality and to elicit the same reactions as those elicited by persons. Similar effects have been shown

in [28][133], where children interacting with computers have modified their voice to match the speaking characteristics of the animated ECA, showing adaptation patterns typical of human-human interactions [20]. Further evidence of the same phenomenon is available in [10][11], where the interaction between humans and ECA is shown to include the *Chameleon effect* [22], i.e. the mutual imitation of individuals due to reciprocal appreciation or to the influence of one individual on the other.

Psychologists have compared the performance of humans and machines in detecting socially relevant information like gender and movements associated to emotional states [65][151][152]. The results show that machines tend to have a constant performance across a wide range of conditions (different behavioral cues at disposition), while humans have dramatic changes in performance (sometimes dropping at chance level) when certain behavioral cues are no longer at disposition. This seems to suggest that humans do not use the behavioral cues actually at their disposition, but rather rely on task specific behavioral cues without which the tasks cannot be performed effectively [65][151][152]. In contrast, automatic approaches (in particular those based on machine learning) are built to rely on any available behavioral cue and their performance simply depends on how much the available cues are actually correlated with the targeted social information.

5 Conclusions and Future Challenges

Social Signal Processing has the ambitious goal of bringing social intelligence [6][66] in computers. The first results in this research domain have been sufficiently impressive to attract the praise of the technology [69] and business [19] communities. What is more important is that they have established a viable interface between human sciences and engineering - social interactions and behaviours, although complex and rooted in the deepest aspects of human psychology, can be analyzed automatically with the help of computers. This interdisciplinarity is, in our opinion, the most important result of research in SSP so far. In fact, the pioneering contributions in SSP [142][143] have shown that the social signals, typically described as so elusive and subtle that only trained psychologists can recognize them [63], are actually evident and detectable enough to be captured through sensors like microphones and cameras, and interpreted through analysis techniques like machine learning and statistics.

However, although fundamental, these are only the first steps and the journey towards *artificial social intelligence* and *socially-aware computing* is still long. In the rest of this section we discuss four challenges facing the researchers in the field, for which we believe are the crucial turnover issues that need to

be addressed before the research in the field can enter its next phase - the deployment phase.

The first issue relates to *tightening of the collaboration between social scientists and engineers*. The analysis of human behaviour in general, and social behaviour in particular, is an inherently multidisciplinary problem [138][221]. More specifically no automatic analysis of social interactions is possible without taking into account the basic mechanisms governing social behaviours that the psychologists have investigated for decades, such as the *chameleon effect* (mutual imitation of people aimed at showing liking or affiliation) [22][99], the interpersonal adaptation (mutual accommodation of behavioural patterns between interacting individuals) [20][71], the interactional synchrony (degree of coordination during interactions) [93], the presence or roles in groups [12][186], the dynamics of conversations [154][217], etc. The collaboration between technology and social sciences demands a mutual effort of the two disciplines. On one hand, engineers need to include the social sciences in their reflection, while on the other hand, social scientists need to formulate their findings in a form useful for engineers and their work on SSP.

The second issue relates to the need of implementing *multi-cue, multi-modal approaches* to SSP. Nonverbal behaviours cannot be read like words in a book [96][158]; they are not unequivocally associated to a specific meaning and their appearance can depend on factors that have nothing to do with social behaviour. Postures correspond in general to social attitudes, but sometimes they are simply comfortable [166], physical distances typically account for social distances, but sometimes they are simply the effect of physical constraints [77]. Moreover, the same signal can correspond to different social behaviour interpretations depending on context and culture [190] (although many advocate that social signals are natural rather than cultural [171]). In other words, social signals are intrinsically ambiguous and the best way to deal with such problem is to use multiple behavioural cues extracted from multiple modalities. Numerous studies have theoretically and empirically demonstrated the advantage of integration of multiple modalities (at least audio and visual) in human behaviour analysis over single modalities (e.g., [162]). This corresponds, from a technological point of view, to the combination of different classifiers that has extensivley been shown to be more effective than single classifiers, as long as they are sufficiently *diverse*, i.e., account for different aspects of the same problem [94]. It is therefore not surprising that some of the most successful works in SSP so far use features extracted from multiple modalities like in [39][85][111]. Note, however, that the relative contributions of different modalities and the related behavioural cues to affect judgment of displayed behaviour depend on the targeted behavioural category and the context in which the behaviour occurs [49][162].

The third issue relates to *the use of real-world data*. Both psychologists and

engineers tend to produce their data in laboratories and artificial settings (see e.g., [33][68][111]), in order to limit parasitic effects and elicit the specific phenomena they want to observe. However, this is likely to simplify excessively the situation and to improve artificially the performance of the automatic approaches. Social interaction is one of the most ubiquitous phenomena in the world - the media (radio and television) show almost exclusively social interactions (debates, movies, talk-shows) [123]. Also other, less common kinds of data are centered on social interactions, e.g., meeting recordings [110], surveillance material [87], and similar. The use of real-world data will allow analysis of interactions that have an actual impact on the life of the participants, thus will show the actual effects of goals and motivations that typically drive human behaviour. This includes also the analysis of *group interactions*, a task difficult from both technological and social point of view because it involves the need of observing multiple people involved in a large number of one-to-one interactions.

The last, but not least, challenging issue relates to the *the identification of applications likely to benefit from SSP*. Applications have the important advantage of linking the effectiveness of detecting social signals to the reality. For example, one of the earliest applications is the prediction of the outcome in transactions recorded at a call center and the results show that the number of successful calls can be increased by around 20% by stopping early the calls that are not promising [19]. This can have not only a positive impact on the marketplace, but also provide *benchmarking procedures* for the SSP research, one of the best means to improve the overall quality of a research domain as extensively shown in fields where international evaluations take place every year (e.g. video analysis in TrecVid [178]).

Acknowledgements. The work of Dr. Vinciarelli is supported by the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2). The work of Dr. Pantic is supported in part by the EC's 7^{th} Framework Programme (FP7/2007-2013) under grant agreement no. 211486 (SEMAINE), and the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The research that has led to this work has been supported in part by the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement no. 231287 (SSPNet).

References

[1] P. Aarabi, D. Hughes, K. Mohajer, and M. Emami. The automatic measurement of facial beauty. In *Proceedings of IEEE International*

Conference on Systems, Man, and Cybernetics, pages 2644–2647, 2001.

[2] R. Adolphs. Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience*, 4(3):165–178, 2003.

[3] M. Ahmad and S.-W. Lee. Human action recognition using shape and CLG-motion flow from multi-view image sequences. *Pattern Recognition*, 41(7):2237–2252, 2008.

[4] J. Ajmera. *Robust Audio Segmentation*. PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL), 2004.

[5] J. Ajmera, I. McCowan, and H. Bourlard. Speech/music segmentation using entropy and dynamism features in a HMM classification framework. *Speech Communication*, 40(3):351–363, 2003.

[6] K. Albrecht. *Social Intelligence: The new science of success*. John Wiley & Sons Ltd, 2005.

[7] N. Ambady, F. Bernieri, and J. Richeson. Towards a histology of social behavior: judgmental accuracy from thin slices of behavior. In M.P. Zanna, editor, *Advances in Experimental Social Psychology*, pages 201–272. 2000.

[8] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992.

[9] M. Argyle. *The Psychology of Interpersonal Behaviour*. Penguin, 1967.

[10] J.N. Bailenson and N. Yee. Virtual interpersonal touch and digital chameleons. *Journal of Nonverbal Behavior*, 31(4):225–242, 2007.

[11] J.N. Bailenson, N. Yee, K. Patel, and A.C. Beall. Detecting digital chameleons. *Computers in Human Behavior*, 24(1):66–87, 2008.

[12] R.F. Bales. *Interaction Process Analysis: A Method for the Study of Small Groups*. Addison-Wesley, 1950.

[13] S. Banerjee and A.I. Rudnicky. Using simple speech based features to detect the state of a meeting and the roles of the meeting participants. In *Proceedings of International Conference on Spoken Language Processing*, pages 2189–2192, 2004.

[14] C. Barras, X. Zhu, S. Meignier, and J.L. Gauvain. Improving speaker diarization. In *Proceedings of the Rich Transcription Workshop*, 2004.

[15] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. The rules behind the roles: identifying speaker roles in radio broadcasts. In *Proceedings of American Association of Artificial Intelligence Symposium*, pages 679–684, 2000.

[16] C. Ben Abdelkader and Y. Yacoob. Statistical estimation of human anthropometry from a single uncalibrated image. In K. Franke, S. Petrovic, and A. Abraham, editors, *Computational Forensics*. Springer Verlag, 2009.

[17] A.F. Bobick and A. Johnson. Gait recognition using static activity-specific parameters. In *Proceedings of Computer Vision and Pattern Recognition*, pages 423–430, 2001.

[18] P. Boersma and D. Weenink. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.

[19] M. Buchanan. The science of subtle signals. *Strategy+Business*, 48:68–77, 2007.

[20] J.K. Burgoon, L.A. Stern, and L. Dillman. *Interpersonal Adaptation: Dyadic Interaction Patterns*. Cambridge University Press, 1995.

[21] N. Campbell. Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Speech and Language Processing*, 14(4):1171–1178, 2006.

[22] T.L. Chartrand and J.A. Bargh. The chameleon effect: the perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910, 1999.

[23] J.F. Cohn. Foundations of human computing: facial expression and emotion. In *Proceedings of the ACM International Conference on Multimodal Interfaces*, pages 233–238, 2006.

[24] J.F. Cohn and P. Ekman. Measuring facial action by manual coding, facial EMG, and automatic facial image analysis. In J.A. Harrigan, R. Rosenthal, and K.R. Scherer, editors, *Handbook of nonverbal behavior research methods in the affective sciences*, pages 9–64. 2005.

[25] J.B. Cortes and F.M. Gatti. Physique and self-description of temperament. *Journal of Consulting Psychology*, 29(5):432–439, 1965.

[26] M. Costa, W. Dinsbach, A.S.R. Manstead, and P.E.R. Bitti. Social presence, embarrassment, and nonverbal behavior. *Journal of Nonverbal Behavior*, 25(4):225–240, 2001.

[27] M. Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, 28(2):117–139, 2004.

[28] R. Coulston, S. Oviatt, and C. Darves. Amplitude convergence in children's conversational speech with animated personas. In *International Conference on Spoken Language Processing*, pages 2689–2692, 2002.

[29] R. Cowie. Building the databases needed to understand rich, spontaneous human behaviour. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[30] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001.

[31] D. Crystal. *Prosodic Systems and Intonation in English*. Cambridge University Press, 1969.

[32] D.W. Cunningham, M. Kleiner, H.H. Bülthoff, and C. Wallraven. The components of conversational facial expressions. *Proceedings of the Symposium on Applied Perception in Graphics and Visualization*, pages 143–150, 2004.

[33] J.R. Curhan and A. Pentland. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, 92(3):802–811, 2007.

[34] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[35] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proceedings of the European Conference on Computer Vision*, pages 428–441, 2006.

[36] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. *International Journal of Computer Vision*, 37(2):175–185, 2000.

[37] C. Darwin. *The Expression of the Emotions in Man and Animals*. J. Murray, 1872.

[38] R. De Silva and N. Bianchi-Berthouze. Modeling human affective postures: an information theoretic characterization of posture features. *Journal of Computational Animation and Virtual World*, 15(3-4):269–276, 2004.

[39] A. Dielmann and S. Renals. Automatic meeting segmentation using dynamic bayesian networks. *IEEE Transactions on Multimedia*, 9(1):25, 2007.

[40] L. Ding and A.M. Martinez. Recovering the linguistic components of the manual signs in american sign language. In *Proceedings of IEEE International Conference on Advanced Video and Signal-based Surveillance*, pages 447–452, 2007.

[41] K. Dion, E. Berscheid, and E. Walster. What is beautiful is good. *Journal of Personality and Social Psychology*, 24(3):285–290, 1972.

[42] W. Dong, B. Lepri, A. Cappelletti, A.S. Pentland, F. Pianesi, and M. Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 271–278, 2007.

[43] N. Eagle and A. Pentland. Reality mining: sensing complex social signals. *Journal of Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[44] Y. Eisenthal, G. Dror, and E. Ruppin. Facial attractiveness: Beauty and the machine. *Neural Computation*, 18(1):119–142, 2005.

[45] P. Ekman, editor. *Emotion in the human face*. Cambridge University Press, 1982.

[46] P. Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1):205–221, 2003.

[47] P. Ekman and W.V. Friesen. The repertoire of nonverbal behavior. *Semiotica*, 1:49–98, 1969.

[48] P. Ekman, W.V. Friesen, and J.C. Hager. *Facial Action Coding System (FACS): Manual*. A Human Face, Salt Lake City (USA), 2002.

[49] P. Ekman, T.S. Huang, T.J. Sejnowski, and J.C. Hager, editors. *Final Report to NSF of the Planning Workshop on Facial Expression Understanding*. Human Interaction Laboratory, University of California, San Francisco, 1993.

[50] P. Ekman and E.L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, 2005.

[51] R.R. Faden, T.L. Beauchamp, and N.M.P. King. *A History and Theory of Informed Consent*. Oxford University Press, 1986.

[52] I.R. Fasel, B. Fortenberry, and J.R. Movellan. A generative framework for real time object detection and classification. *Computer Vision and Image Understanding*, 98(1):181–210, 2005.

[53] S. Favre, H. Salamin, J. Dines, and A. Vinciarelli. Role recognition in multiparty recordings using Social Affiliation Networks and discrete distributions. In *Proceedings of the ACM International Conference on Multimodal Interfaces*, pages 29–36, 2008.

[54] D.A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan. Computational studies of human motion part 1: Tracking and motion synthesis. *Foundations and Trends in Computer Graphics and Vision*, 1(2):77–254, 2006.

[55] N. Fragopanagos and J.G. Taylor. Emotion recognition in human–computer interaction. *Neural Networks*, 18(4):389–405, 2005.

[56] M.G. Frank and P. Ekman. Appearing truthful generalizes across different deception situations. *Journal of Personality and Social Psychology*, 86(3):486–495, 2004.

[57] T. Gandhi and M.M. Trivedi. Pedestrian protection systems: Issues, survey, and challenges. *IEEE Transactions On Intelligent Transportation Systems*, 8(3):413–430, 2007.

[58] C. Garcia and G. Tziritas. Face detection using quantized skin color regions merging and wavelet packet analysis. *IEEE Transactions on Multimedia*, 1(3):264–277, 1999.

[59] N. Garg, S. Favre, H. Salamin, D. Hakkani-Tür, and A. Vinciarelli. Role recognition for meeting participants: an approach based on lexical information and social network analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 693–696, 2008.

[60] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio. Detecting group interest-level in meetings. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 489–492, 2005.

[61] J.L. Gauvain, L.F. Lamel, and G. Adda. Partitioning and transcription of broadcast news data. In *Proceedings of International Conference on Spoken Language Processing*, pages 1335–1338, 1998.

[62] D.M. Gavrila. Visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[63] M. Gladwell. *Blink: The Power of Thinking without Thinking*. Little Brown & Company, 2005.

[64] C.R. Glass, T.V. Merluzzi, J.L. Biever, and K.H. Larsen. Cognitive assessment of social anxiety: Development and validation of a self-statement questionnaire. *Cognitive Therapy and Research*, 6(1):37–55, 1982.

[65] J.M. Gold, D. Tadin, S.C. Cook, and R.B. Blake. The efficiency of biological motion perception. *Perception and Psychophysics*, 70(1):88–95, 2008.

[66] D. Goleman. *Social intelligence*. Hutchinson, 2006.

[67] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, 2007.

[68] J.E. Grahe and F.J. Bernieri. The importance of nonverbal cues in judging rapport. *Journal of Nonverbal Behavior*, 23(4):253–269, 1999.

[69] K. Greene. 10 emerging technologies 2008. *MIT Technology Review*, february 2008.

[70] M.R. Greenwald, A.G.and Banaji. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1):4–27, 1995.

[71] S.W. Gregory, K. Dagan, and S. Webster. Evaluating the relation of vocal accommodation in conversation partners fundamental frequencies to perceptions of communication quality. *Journal of Nonverbal Behavior*, 21(1):23–43, 1997.

[72] M.M. Gross, E.A. Crane, and B.L. Fredrickson. Effect of felt and recognized emotions on body movements during walking. In *Proceedings of the International Conference on The Expression of Emotions in Health and Disease*, 2007.

[73] H. Gunes and M. Piccardi. Assessing facial beauty through proportion analysis by image processing and supervised learning. *International Journal of Human-Computer Studies*, 64(12):1184–1199, 2006.

[74] H. Gunes and M. Piccardi. Bi-modal emotion reognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345, 2007.

[75] H. Gunes, M. Piccardi, and T. Jan. Comparative beauty classification for pre-surgery planning. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pages 2168–2174, 2004.

[76] H. Gunes, M. Piccardi, and M. Pantic. From the lab to the real world: Affect recognition using multiple cues and modalities. In J. Or, editor, *Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition*, pages 185–218. 2008.

[77] E.T. Hall. *The silent language*. Doubleday, 1959.

[78] J.B. Hayfron-Acquah, M.S. Nixon, and J.N. Carter. Automatic gait recognition by symmetry analysis. *Pattern Recognition Letters*, 24(13):2175–2183, 2003.

[79] J. Hirschberg. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63(1-2):305–340, 1993.

[80] J. Hirschberg and B. Grosz. Intonational features of local and global discourse structure. In *Proceedings of the Speech and Natural Language Workshop*, pages 441–446, 1992.

[81] E. Hjelmas and B.K. Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83(3):236–274, 2001.

[82] C.R.L. Hsu, M. Abdel-Mottaleb, and A.K. Jain. Face detection in colour images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):696–706, 2002.

[83] K.S. Huang and M.M. Trivedi. Robust real-time detection, tracking, and pose estimation of faces in video streams. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 965–968, 2004.

[84] X. Huang, A. Acero, and H.W. Hon. *Spoken language processing*. Prentice Hall, 2001.

[85] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. Ba, J.M. Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez. Using audio and video features to classify the most dominant person in a group meeting. In *Proceedings of the ACM International Conference on Multimedia*, pages 835–838, 2007.

[86] A. Ito, X. Wang, M. Suzuki, and S. Makino. Smile and laughter recognition using speech processing and face recognition from conversation video. In *Proceedings of th International Conference on Cyberworlds*, pages 437–444, 2005.

[87] Y. Ivanov, C. Stauffer, A. Bobick, and W.E.L. Grimson. Video surveillance of interactions. In *Proceedings of the Workshop on Visual Surveillance at Computer Vision and Pattern Recognition*, 1999.

[88] A. Jaimes, K. Omura, T. Nagamine, and K. Hirata. Memory cues for meeting video retrieval. In *Proceedings of Workshop on Continuous Archival and Retrieval of Personal Experiences*, pages 74–85, 2004.

[89] D. Keltner and P. Ekman. Facial expression of emotion. In M. Lewis and J.M. Haviland-Jones, editors, *Handbook of Emotions*, pages 236–249. 2000.

[90] D. Keltner and J. Haidt. Social functions of emotions at four levels of analysis. *Cognition and Emotion*, 13(5):505–521, 1999.

[91] D. Keltner and A.M. Kring. Emotion, social function, and psychopathology. *Review of General Psychology*, 2(3):320–342, 1998.

[92] L. Kennedy and D. Ellis. Laughter detection in meetings. In *Proceedings of the NIST Meeting Recognition Workshop*, 2004.

[93] M. Kimura and I. Daibo. Interactional synchrony in conversations about emotional episodes: A measurement by the between-participants pseudosynchrony experimental paradigm. *Journal of Nonverbal Behavior*, 30(3):115–126, 2006.

[94] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[95] A. Kleinsmith, R. De Silva, and N. Bianchi-Berthouze. Cross-cultural differences in recognizing affect from body posture. *Interacting with Computers*, 18(6):1371–1389, 2006.

[96] M.L. Knapp and J.A. Hall. *Nonverbal Communication in Human Interaction*. Harcourt Brace College Publishers, 1972.

[97] W.W. Kong and S. Ranganath. Automatic hand trajectory segmentation and phoneme transcription for sign language. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[98] Z. Kunda. *Social Cognition*. MIT Press, 1999.

[99] J.L. Lakin, V.E. Jefferis, C.M. Cheng, and T.L. Chartrand. The Chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior*, 27(3):145–162, 2003.

[100] L. Lee and W.E.L. Grimson. Gait analysis for recognition and classification. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 148–155, 2002.

[101] T.K. Leung, M.C. Burl, and P. Perona. Finding faces in cluttered scenes using random labeled graphmatching. In *Proceedings of the International Conference on Computer Vision*, pages 637–644, 1995.

[102] X. Li, S.J. Maybank, S. Yan, D. Tao, and D. Xu. Gait components and their application to gender recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 38(2):145–155, 2008.

[103] G. Littlewort, M.S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006.

[104] G.C. Littlewort, M.S. Bartlett, and K. Lee. Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 15–21, 2007.

[105] Y. Liu, E. Shriberg, A. Stolcke, and M. Harper. Comparing HMM, maximum entropy, and conditional random fields for disfluency detection. In *Proceeedings of the European Conference on Speech Communication and Technology*, 2005.

[106] D.F. Lott and R. Sommer. Seating arrangements and status. *Journal of Personality and Social Psychology*, 7(1):90–95, 1967.

[107] L. Lu, H.J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 10(7):504–516, 2002.

[108] S. Lucey, A.B. Ashraf, and J. Cohn. Investigating spontaneous facial action recognition through AAM representations of the face. In K. Delac and M. Grgic, editors, *Handbook of Face Recognition*, pages 275–286. I-Tech Education and Publishing, 2007.

[109] L.Z. McArthur and R.M. Baron. Toward an ecological theory of social perception. *Psychological Review*, 90(3):215–238, 1983.

[110] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling human interaction in meetings. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 748–751, 2003.

[111] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):305–317, 2005.

[112] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University Of Chicago Press, 1996.

[113] A. Mehrabian and S.R. Ferris. Inference of attitude from nonverbal communication in two channels. *Journal of Counseling Psychology*, 31(3):248–252, 1967.

[114] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proceedings of the European Conference on Computer Vision*, pages 69–81, 2004.

[115] T.B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[116] T.B Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, 2006.

[117] Y. Moh, P. Nguyen, and J.C. Junqua. Towards domain independent speaker clustering. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 85–88, 2003.

[118] S. Möller and R. Schönweiler. Analysis of infant cries for the early detection of hearing impairment. *Speech Communication*, 28(3):175–193, 1999.

[119] P.R. Montague, G.S. Berns, J.D. Cohen, S.M. McClure, G. Pagnoni, M. Dhamala, M.C. Wiest, I. Karpov, R.D. King, N. Apple, and R.E. Fisher. Hyperscanning: Simultaneous fMRI during linked social interactions. *Neuroimage*, 16(4):1159–1164, 2002.

[120] B.C.J. Moore. *An introduction to the psychology of hearing*. Academic Press, 1982.

[121] N. Morgan, E. Fosler, and N. Mirghafori. Speech recognition using on-line estimation of speaking rate. In *Proceedings of Eurospeech*, pages 2079–2082, 1997.

[122] N. Morgan and E. Fosler-Lussier. Combining multiple estimators of speaking rate. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 729–732, 1998.

[123] D. Morris. *Peoplewatching*. Vintage, 2007.

[124] S. Mota and R.W. Picard. Automated posture analysis for detecting learners interest level. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 49–56, 2003.

[125] S. Mukhopadhyay and B. Smith. Passive capture and structuring of lectures. *Proceedings of the ACM International Conference on Multimedia*, pages 477–487, 1999.

[126] K.P. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California Berkeley, 2002.

[127] C. Nass and K.M. Lee. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3):171–181, 2001.

[128] C. Nass and J. Steuer. Computers and social actors. *Human Communication Research*, 19(4):504–527, 1993.

[129] I. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions On Systems, Man, and Cybernetics - Part B*, 36(3):710–719, 2006.

[130] N. Oliver, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding*, 96(2):163–180, 2004.

[131] A. J. O'Toole, T. Price, T. Vetter, J.C. Bartlett, and V. Blanz. 3D shape and 2D surface textures of human faces: the role of averages in attractiveness and age. *Image and Vision Computing*, 18(1):9–19, 1999.

[132] S. Oviatt. User-centered modeling and evaluation of multimodal interfaces. *Proceedings of the IEEE*, 91:1457–1468, 2003.

[133] S. Oviatt, C. Darves, and R. Coulston. Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *ACM Transactions on Computer-Human Interaction*, 11(3):300–328, 2004.

[134] P. Pal, A.N. Iyer, and R.E. Yantorno. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 721–724, 2006.

[135] M. Pantic and M.S. Bartlett. Machine analysis of facial expressions. In K. Delac and M. Grgic, editors, *Handbook of Face Recognition*, pages 377–416. I-Tech Education and Publishing, 2007.

[136] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, 36(2):433–449, 2006.

[137] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human computing and machine understanding of human behavior: A survey. In *Lecture Notes in Artificial Intelligence*, volume 4451, pages 47–71. Springer Verlag, 2007.

[138] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human-centred intelligent human-computer interaction (HCI^2): How far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, 1(2):168–187, 2008.

[139] M. Pantic and L.J.M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[140] M. Pantic and L.J.M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[141] C. Pelachaud, V. Carofiglio, B. De Carolis, F. de Rosis, and I. Poggi. Embodied contextual agent in information delivering application. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 758–765, 2002.

[142] A. Pentland. Social dynamics: Signals and behavior. In *International Conference on Developmental Learning*, 2004.

[143] A. Pentland. Socially aware computation and communication. *IEEE Computer*, 38(3):33–40, 2005.

[144] A. Pentland. Automatic mapping and modeling of human networks. *Physica A*, 378:59–67, 2007.

[145] A. Pentland. Social Signal Processing. *IEEE Signal Processing Magazine*, 24(4):108–111, 2007.

[146] A. Pentland. *Honest signals: how they shape our world*. MIT Press, 2008.

[147] S. Petridis and M. Pantic. Audiovisual discrimination between laughter and speech. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5117–5121, 2008.

[148] S. Petridis and M. Pantic. Audiovisual laughter detection based on temporal features. In *Proceedings of IEEE International Conference on Multimodal Interfaces*, pages 37–44, 2008.

[149] T. Pfau and G. Ruske. Estimating the speaking rate by vowel detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 945–948, 1998.

[150] J.O. Pickles. *An introduction to the physiology of hearing*. Academic Press, 1982.

[151] F.E. Pollick, V. Lestou, J. Ryu, and S.B. Cho. Estimating the efficiency of recognizing gender and affect from biological motion. *Vision Research*, 42:2345–2355, 2002.

[152] F.E. Pollick, H.M. Paterson, A. Bruderlin, and A.J. Sanford. Perceiving affect from arm movement. *Cognition*, 82(2):51–61, 2001.

[153] R. Poppe. Vision-based human motion analysis: an overview. *Computer Vision and Image Understanding*, 108:4–18, 2007.

[154] G. Psathas. *Conversation Analysis - The study of talk-in-interaction*. Sage Publications, 1995.

[155] L. Rabiner and M. Sambur. Voiced-unvoiced-silence detection using the Itakura LPC distance measure. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 323–326, 1977.

[156] L.R. Rabiner and R.W. Schafer. *Digital processing of speech signals*. Prentice-Hall Englewood Cliffs, NJ, 1978.

[157] D.A. Reynolds, W. Campbell, T.T. Gleason, C. Quillen, D. Sturim, P. Torres-Carrasquillo, and A. Adami. The 2004 MIT Lincoln laboratory speaker recognition system. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 177–180, 2005.

[158] V.P. Richmond and J.C. McCroskey. *Nonverbal Behaviors in interpersonal relations*. Allyn and Bacon, 1995.

[159] R. Rienks and D. Heylen. Dominance Detection in Meetings Using Easily Obtainable Features. In *Lecture Notes in Computer Science*, volume 3869, pages 76–86. Springer, 2006.

[160] R. Rienks, D. Zhang, and D. Gatica-Perez. Detection and application of influence rankings in small group meetings. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 257–264, 2006.

[161] H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[162] J.A. Russell, J.A. Bachorowski, and J.M. Fernandez-Dols. Facial and vocal expressions of emotion. *Annual Reviews in Psychology*, 54(1):329–349, 2003.

[163] J.A. Russell and J.M. Fernandez-Dols, editors. *he Psychology of Facial Expression*. Cambridge University Press, 1997.

[164] N. Russo. Connotation of seating arrangements. *Cornell Journal of Social Relations*, 2(1):37–44, 1967.

[165] M.A. Sayette, D.W. Smith, M.J. Breiner, and G.T. Wilson. The effect of alcohol on emotional response to a social stressor. *Journal of Studies on Alcohol*, 53(6):541–545, 1992.

[166] A.E. Scheflen. The significance of posture in communication systems. *Psychiatry*, 27:316–331, 1964.

[167] K.R. Scherer. *Personality markers in speech*. Cambridge University Press, 1979.

[168] K.R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256, 2003.

[169] H. Schneiderman and T. Kanade. A statistical model for 3D object detection applied to faces and cars. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 746–751, 2000.

[170] B. Schuller, R. Müeller, B. Höernler, A. Höethker, H. Konosu, and G. Rigoll. Audiovisual recognition of spontaneous interest within conversations. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 30–37, 2007.

[171] U. Segerstrale and P. Molnar, editors. *Nonverbal communication: where nature meets culture*. Lawrence Erlbaum Associates, 1997.

[172] A. Sepheri, Y. Yacoob, and L. Davis. Employing the hand as an interface device. *Journal of Multimedia*, 1(7):18–29, 2006.

[173] E. Shriberg. Phonetic consequences of speech disfluency. *Proceedings of the International Congress of Phonetic Sciences*, 1:619–622, 1999.

[174] E. Shriberg, R. Bates, and A. Stolcke. A prosody-only decision-tree model for disfluency detection. In *Proceedings of Eurospeech*, pages 2383–2386, 1997.

[175] E. Shriberg, A. Stolcke, and D. Baron. Observations of overlap: findings and implications for automatic processing of multiparty conversation. In *Proceedings of Eurospeech*, pages 1359–1362, 2001.

[176] P.E. Shrout and D.W. Fiske. Nonverbal behaviors and social evaluation. *Journal of Personality*, 49(2):115–128, 1981.

[177] K. Sjölander and J. Beskow. Wavesurfer-an open source speech tool. In *Proceedings of International Conference on Spoken Language Processing*, pages 464–467, 2000.

[178] A.F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, 2006.

[179] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding*, 104(2-3):210–220, 2006.

[180] L. Smith-Lovin and C. Brody. Interruptions in group discussions: the effects of gender and group composition. *American Sociological Review*, 54(3):424–435, 1989.

[181] K.K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

[182] E. L. Thorndike. Intelligence and its use. *Harper's Magazine*, 140:227–235, 1920.

[183] C. Thurau. Behavior histograms for action recognition and human detection. In *Lecture Notes in Computer Science*, volume 4814, pages 271–284. Springer Verlag, 2007.

[184] C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2008.

[185] Y. Tian, T. Kanade, and J.F. Cohn. Facial expression analysis. In S.Z. Li and A.K. Jain, editors, *Handbook of Face Recognition*, pages 247–276. 2005.

[186] H.L. Tischler. *Introduction to Sociology.* Harcourt Brace College Publishers, 1990.

[187] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1683–1699, 2007.

[188] D. Tran, A. Sorokin, and D.A. Forsyth. Human activity recognition with metric learning. In *Proceedings of the European Conference on Computer Vision*, 2008.

[189] S.E. Tranter and D.A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565, 2006.

[190] H.C. Triandis. *Culture and social behavior*. McGraw-Hill, 1994.

[191] K.P. Truong and D.A. Leeuwen. Automatic detection of laughter. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 485–488, 2005.

[192] K.P. Truong and D.A. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007.

[193] L.Q. Uddin, M. Iacoboni, C. Lange, and J.P. Keenan. The self and social cognition: the role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences*, 11(4):153–157, 2007.

[194] A. Utsumi and N. Tetsutani. Human detection using geometrical pixel value structures. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 34–39, 2002.

[195] M.F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 38–45, 2007.

[196] M.F. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition Workshop*, pages 149–150, 2006.

[197] M.F. Valstar, M. Pantic, Z. Ambadar, and J.F. Cohn. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 162–170, 2006.

[198] J. Van den Stock, R. Righart, and B. de Gelder. Body expressions influence recognition of emotions in the face and voice. *Emotion*, 7(3):487–494, 2007.

[199] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 301–308, 2001.

[200] A. Vinciarelli. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, 9(9):1215–1226, 2007.

[201] A. Vinciarelli and J.-M. Odobez. Application of information retrieval technologies to presentation slides. *IEEE Transactions on Multimedia*, 8(5):981–995, 2006.

[202] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social Signal Processing: State-of-the-art and future perspectives of an emerging domain. In *Proceedings of the ACM International Conference on Multimedia*, pages 1061–1070, 2008.

[203] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social signals, their function, and automatic analysis: A survey. In *Proceedings of the ACM International Conference on Multimodal Interfaces*, pages 61–68, 2008.

[204] P. Viola and M. Jones. Robust real-time face detection. *Computer Vision*, 57(2):137–154, 2004.

[205] A. Waibel, T. Schultz, M. Bett, M. Denecke, R. Malkin, I. Rogina, and R. Stiefelhagen. SMaRT: the Smart Meeting Room task at ISL. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 752–755, 2003.

[206] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36(3):585–601, 2003.

[207] P. Wang and Q. Ji. Multi-view face detection under complex scene based on combined SVMs. In *Proceedings of International Conference on Pattern Recognition*, pages 179–182, 2004.

[208] R.M. Warner and D.B. Sugarman. Attributions of personality based on physical appearance, speech, and handwriting. *Journal of Personality and Social Psychology*, 50(4):792–799, 1986.

[209] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[210] C.Y. Weng, W.T. Chu, and J.L. Wu. Movie analysis based on roles social network. In *Proceedings of IEEE International Conference on Multimedia and Expo*, pages 1403–1406, 2007.

[211] J. Whitehill and J.R. Movellan. Personalized facial attractiveness prediction. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[212] A.C.C. Williams. Facial expression of pain: An evolutionary account. *Behavioral and Brain Sciences*, 25(4):439–455, 2003.

[213] Y. Wu and T.S. Huang. Vision-based gesture recognition: A review. In *Proceedings of the International Gesture Workshop*, pages 103–109, 1999.

[214] Y. Yacoob and L. Davis. Detection and analysis of hair. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1164–1169, 2006.

[215] M.H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[216] J. Yao and J.-M. Odobez. Fast human detection from videos using covariance features. In *Proceedings of European Conference on Computer Vision Visual Surveillance Workshop*, 2008.

[217] G. Yule. *Pragmatics*. Oxford University Press, 1996.

[218] M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 28–34, 2006.

[219] B. Zellner. Pauses and the temporal structure of speech. In E. Keller, editor, *Fundamentals of speech synthesis and speech recognition*, pages 41–62. John Wiley & Sons, 1994.

[220] Z. Zeng, Y. Fu, G.I. Roisman, Z. Wen, Y. Hu, and T.S. Huang. Spontaneous emotional facial expression detection. *Journal of Multimedia*, 1(5):1–8, 2006.

[221] Z. Zeng, M. Pantic, G.I. Roisman, and T.H. Huang. A survey of affect recognition methods: audio, visual and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[222] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):699–714, 2005.

[223] Q. Zhu, S. Avidan, M.C. Yeh, and K.T. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *Proceedings of Computer Vision and Pattern Recognition*, pages 1491–1498, 2006.

On Space-Time Interest Points*

Ivan Laptev and Tony Lindeberg

Computational Vision and Active Perception Laboratory (CVAP),
Department of Numerical Analysis and Computer Science,
KTH, SE-100 44 Stockholm, Sweden

Email: {laptev, tony}@nada.kth.se

Technical report ISRN KTH/NA/P–03/12–SE

Shortened version in Proc. ICCV'03
Nice, France, October 2003, pp. 432–439.

Earlier version also in Proc. Scale Space '03
Isle of Skye, Scotland, June 2003, pp. 372–387.

Abstract

Local image features or interest points provide compact and abstract representations of patterns in an image. In this paper, we extend the notion of spatial interest points into the spatio-temporal domain and show how the resulting features capture interesting events in video and can be used for a compact representation and for interpretation of video data.

To detect spatio-temporal events, we build on the idea of the Harris and Förstner interest point operators and detect local structures in space-time where the image values have significant local variations in both space and time. We estimate the spatio-temporal extents of the detected events by maximizing a normalized spatio-temporal Laplacian operator over spatial and temporal scales. To represent the detected events, we then compute local, spatio-temporal, scale-invariant N-jets and classify each event with respect to its jet descriptor. For the problem of human motion analysis, we illustrate how a video representation in terms of local space-time features allows for detection of walking people in scenes with occlusions and dynamic cluttered backgrounds.

*The support from the Swedish Research Council and from the Royal Swedish Academy of Sciences as well as the Knut and Alice Wallenberg Foundation is gratefully acknowledged.

1 Introduction

Analyzing and interpreting video is a growing topic in computer vision and its applications. Video data contains information about changes in the environment and is highly important for many visual tasks including navigation, surveillance and video indexing.

Traditional approaches for motion analysis mainly involve the computation of optic flow (Barron, Fleet and Beauchemin, 1994) or feature tracking (Smith and Brady, 1995; Blake and Isard, 1998). Although very effective for many tasks, both of these techniques have limitations. Optic flow approaches mostly capture first-order motion and may fail when the motion has sudden changes. Interesting solutions to this problem have been proposed (Niyogi, 1995; Fleet, Black and Jepson, 1998; Hoey and Little, 2000). Feature trackers often assume a constant appearance of image patches over time and may hence fail when the appearance changes, for example, in situations when two objects in the image merge or split. Model-based solutions for this problem have been presented by (Black and Jepson, 1998).

Image structures in video are not restricted to constant velocity and/or constant appearance over time. On the contrary, many interesting events in video are characterized by strong variations in the data along both the spatial and the temporal dimensions. For example, consider a scene with a person entering a room, applauding hand gestures, a car crash or a water splash; see also the illustrations in figure 1.

More generally, points with non-constant motion correspond to accelerating local image structures that may correspond to accelerating objects in the world. Hence, such points can be expected to contain information about the forces acting in the physical environment and changing its structure.

In the spatial domain, points with a significant local variation of image intensities have been extensively investigated in the past (Förstner and Gülch, 1987; Harris and Stephens, 1988; Lindeberg, 1998; Schmid, Mohr and Bauckhage, 2000). Such image points are

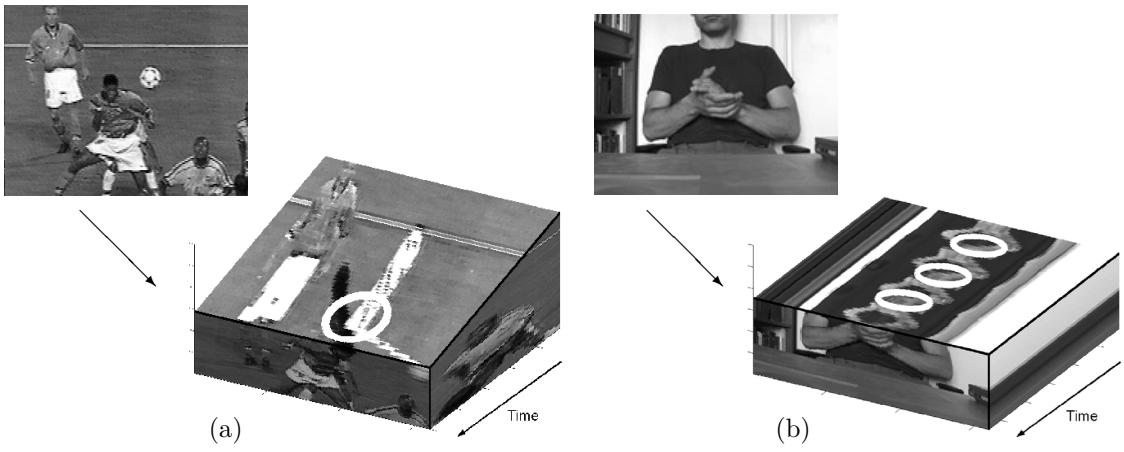


Figure 1: Result of detecting the strongest spatio-temporal interest points in a football sequence with a player heading the ball (a) and in a hand clapping sequence (b). From the temporal slices of space-time volumes shown here, it is evident that the detected events correspond to neighborhoods with high spatio-temporal variation in the image data or "space-time corners".

frequently referred to as "interest points" and are attractive due to their high information contents and relative stability with respect to perspective transformations of the data. Highly successful applications of interest points have been presented for image indexing (Schmid and Mohr, 1997), stereo matching (Tuytelaars and Van Gool, 2000; Mikolajczyk and Schmid, 2002; Tell and Carlsson, 2002), optic flow estimation and tracking (Smith and Brady, 1995; Bretzner and Lindeberg, 1998), and object recognition (Lowe, 1999; Hall, de Verdiere and Crowley, 2000; Fergus, Perona and Zisserman, 2003; Wallraven, Caputo and Graf, 2003).

In this paper, we extend the notion of interest points into the spatio-temporal domain and show that the resulting local space-time features often correspond to interesting events in video data (see figure 1). In particular, we aim at a direct scheme for event detection and interpretation that does not require feature tracking nor computation of optic flow. In the considered sample application we show that the proposed type of features can be used for sparse coding of video information that in turn can be used for interpreting video scenes such as human motion in situations with complex and non-stationary background.

To detect spatio-temporal interest points, we build on the idea of the Harris and Förstner interest point operators (Harris and Stephens, 1988; Förstner and Gülch, 1987) and describe the detection method in section 2. As events often have characteristic extents in both space and time (Koenderink, 1988; Lindeberg and Fagerström, 1996; Florack, 1997; Lindeberg, 1997; Chomat, Martin and Crowley, 2000b; Zelnik-Manor and Irani, 2001), we investigate the behavior of interest points in spatio-temporal scale-space and adapt both the spatial and the temporal scales of the detected features in section 3. In section 4, we show how the neighborhoods of interest points can be described in terms of spatio-temporal derivatives and then be used to distinguish different events in video. In section 5, we consider a video representation in terms of classified spatio-temporal interest points and demonstrate how this representation can be efficient for the task of video registration. In particular, we present an approach for detecting walking people in complex scenes with occlusions and dynamic background. Finally, section 6 concludes the paper with the summary and discussion.

2 Spatio-temporal interest point detection

2.1 Interest points in the spatial domain

In the spatial domain, we can model an image $f^{sp} : \mathbb{R}^2 \mapsto \mathbb{R}$ by its linear scale-space representation (Witkin, 1983; Koenderink and van Doorn, 1992; Lindeberg, 1994; Florack, 1997) $L^t \colon \mathbb{R}^2 \times \mathbb{R}_+ \mapsto \mathbb{R}$

$$L^{sp}(x, y; \sigma_l^2) = g^{sp}(x, y; \sigma_l^2) * f^{sp}(x, y), \tag{1}$$

defined by the convolution of f^{sp} with Gaussian kernels of variance σ_l^2

$$g^{sp}(x, y; \sigma_l^2) = \frac{1}{2\pi\sigma_l^2} \exp(-(x^2 + y^2)/2\sigma_l^2). \tag{2}$$

The idea of the Harris interest point detector is to find spatial locations where f^{sp} has significant changes in both directions. For a given scale of observation σ_l^2, such points can be found using a second moment matrix integrated over a Gaussian window with

variance σ_i^2 (Förstner and Gülch, 1987; Bigün, Granlund and Wiklund, 1991; Lindeberg and Garding, 1997):

$$\begin{aligned} \mu^{sp}(\cdot;\, \sigma_l^2, \sigma_i^2) &= g^{sp}(\cdot;\, \sigma_i^2) * \big((\nabla L(\cdot;\, \sigma_l^2))(\nabla L(\cdot;\, \sigma_l^2))^T\big) \\ &= g^{sp}(\cdot;\, \sigma_i^2) * \begin{pmatrix} (L_x^{sp})^2 & L_x^{sp} L_y^{sp} \\ L_x^{sp} L_y^{sp} & (L_y^{sp})^2 \end{pmatrix} \end{aligned} \tag{3}$$

where $'*'$ denotes the convolution operator, and L_x^{sp} and L_y^{sp} are Gaussian derivatives computed at local scale σ_l^2 according to $L_x^{sp} = \partial_x(g^{sp}(\cdot;\, \sigma_l^2) * f^{sp}(\cdot))$ and $L_y^{sp} = \partial_y(g^{sp}(\cdot;\, \sigma_l^2) * f^{sp}(\cdot))$. The second moment descriptor can be thought of as the covariance matrix of a two-dimensional distribution of image orientations in the local neighborhood of a point. Hence, the eigenvalues λ_1, λ_2, $(\lambda_1 \leq \lambda_2)$ of μ^{sp} constitute descriptors of variations in f^{sp} along the two image directions. Specifically, two significantly large values of λ_1, λ_2 indicate the presence of an interest point. To detect such points, Harris and Stephens (1988) proposed to detect positive maxima of the corner function

$$H^{sp} = \det(\mu^{sp}) - k\, \mathrm{trace}^2(\mu^{sp}) = \lambda_1 \lambda_2 - k(\lambda_1 + \lambda_2)^2. \tag{4}$$

At the positions of the interest points, the ratio of the eigenvalues $\alpha = \lambda_2/\lambda_1$ has to be high. From (4) it follows that for positive local maxima of H^{sp}, the ratio α has to satisfy $k \leq \alpha/(1+\alpha)^2$. Hence, if we set $k = 0.25$, the positive maxima of H will only correspond to ideally isotropic interest points with $\alpha = 1$, i.e. $\lambda_1 = \lambda_2$. Lower values of k allow us to detect interest points with more elongated shape, corresponding to higher values of α. A commonly used value of k in the literature is $k = 0.04$ corresponding to the detection of points with $\alpha < 23$.

The result of detecting Harris interest points in an outdoor image sequence of a walking person is presented at the bottom row of figure 8.

2.2 Interest points in the spatio-temporal domain

In this section, we develop an operator that responds to events in temporal image sequences at specific locations and with specific extents in space-time. The idea is to extend the notion of interest points in the spatial domain by requiring the image values in local spatio-temporal volumes to have large variations along both the spatial and the temporal directions. Points with such properties will correspond to spatial interest points with a distinct location in time corresponding to a local spatio-temporal neighborhood with non-constant motion.

To model a spatio-temporal image sequence, we use a function $f\colon \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ and construct its linear scale-space representation $L\colon \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}_+^2 \mapsto \mathbb{R}$ by convolution of f with an anisotropic Gaussian kernel[1] with distinct spatial variance σ_l^2 and temporal variance τ_l^2

$$L(\cdot;\, \sigma_l^2, \tau_l^2) = g(\cdot;\, \sigma_l^2, \tau_l^2) * f(\cdot), \tag{5}$$

[1] In general, convolution with a Gaussian kernel in the temporal domain violates causality constraints, since the temporal image data is available only for the past. For real-time implementation, time-causal scale-space filters thus have to be used (Koenderink, 1988; Lindeberg and Fagerström, 1996; Florack, 1997; Lindeberg, 2002). In this paper, however, we simplify this part of the investigation and assume that the data is available for a sufficiently long period of time and that the image sequence can be convolved with a Gaussian kernel over both space and time.

where the spatio-temporal separable Gaussian kernel is defined as

$$g(x, y, t; \sigma_l^2, \tau_l^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}} \exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/2\tau_l^2). \tag{6}$$

Using a separate scale parameter for the temporal domain is essential, since the spatial and the temporal extents of events are in general independent. Moreover, as will be illustrated in section 2.3, events detected using our interest point operator depend on both the spatial and the temporal scales of observation and, hence, require separate treatment of the corresponding scale parameters σ_l^2 and τ_l^2.

Similar to the spatial domain, we consider a spatio-temporal second-moment matrix, which is a 3-by-3 matrix composed of first order spatial and temporal derivatives averaged using a Gaussian weighting function $g(\cdot; \sigma_i^2, \tau_i^2)$

$$\mu = g(\cdot; \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}, \tag{7}$$

where we here relate the integration scales σ_i^2 and τ_i^2 to the local scales σ_l^2 and τ_l^2 according to $\sigma_i^2 = s\sigma_l^2$ and $\tau_i^2 = s\tau_l^2$. The first-order derivatives are defined as

$$L_x(\cdot; \sigma_l^2, \tau_l^2) = \partial_x(g * f), \quad L_y(\cdot; \sigma_l^2, \tau_l^2) = \partial_y(g * f), \quad L_t(\cdot; \sigma_l^2, \tau_l^2) = \partial_t(g * f).$$

To detect interest points, we search for regions in f having significant eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of μ. Among different approaches to find such regions, we propose here to extend the Harris corner function (4) defined for the spatial domain into the spatio-temporal domain by combining the determinant and the trace of μ as follows:

$$H = \det(\mu) - k \operatorname{trace}^3(\mu) = \lambda_1 \lambda_2 \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3. \tag{8}$$

To show how positive local maxima of H correspond to points with high values of $\lambda_1, \lambda_2, \lambda_3$ ($\lambda_1 \leq \lambda_2 \leq \lambda_3$), we define the ratios $\alpha = \lambda_2/\lambda_1$ and $\beta = \lambda_3/\lambda_1$ and re-write H as

$$H = \lambda_1^3(\alpha\beta - k(1 + \alpha + \beta)^3).$$

From the requirement $H \geq 0$, we get $k \leq \alpha\beta/(1 + \alpha + \beta)^3$ and it follows that k assumes its maximum possible value $k = 1/27$ when $\alpha = \beta = 1$. For sufficiently large values of k, positive local maxima of H correspond to points with high variation in the image values along both the spatial and the temporal directions. In particular, if we set the maximum value of α, β to 23 as in the spatial domain, the value of k to be used in H (8) will then be $k \approx 0.005$. Thus, spatio-temporal interest points of f can be found by detecting local positive spatio-temporal maxima in H.

2.3 Experimental results for synthetic data

In this section, we illustrate the detection of spatio-temporal interest points on synthetic image sequences. For clarity of presentation, we show the spatio-temporal data as 3-D space-time plots, where the original signal is represented by a threshold surface, while the detected interest points are illustrated by ellipsoids with positions corresponding to the

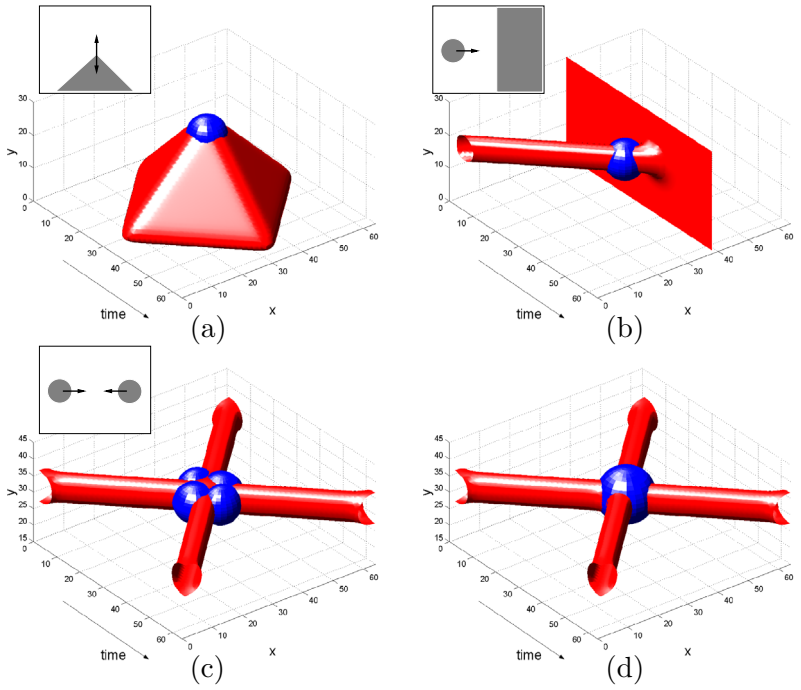


Figure 2: Results of detecting spatio-temporal interest points on synthetic image sequences. (a): A moving corner; (b) A merge of a ball and a wall; (c): Collision of two balls with interest points detected at scales $\sigma_l^2 = 8$ and $\tau_l^2 = 8$; (d): the same signal as in (c) but with the interest points detected at coarser scales $\sigma_l^2 = 16$ and $\tau_l^2 = 16$.

space-time location of the interest point and the length of the semi-axes proportional to the local scale parameters σ_l and τ_l used in the computation of H.

Figure 2a shows a sequence with a moving corner. The interest point is detected at the moment in time when the motion of the corner changes direction. This type of event occurs frequently in natural sequences, such as sequences of articulated motion. Note that according to the definition of spatio-temporal interest points, image structures with constant motion do not give rise to responses of the detector. Other typical types of events that can be detected by the proposed method are splits and unifications of image structures. In figure 2b, the interest point is detected at the moment and the position corresponding to the collision of a ball and a wall. Similarly, interest points are detected at the moment of collision and bouncing of two balls as shown in figure 2c-d. Note, that different types of events are detected depending on the scale of observation.

To further emphasize the importance of the spatial and the temporal scales of observation, let us consider an oscillating signal with different spatial and temporal frequencies defined by $f(x, y, t) = \mathrm{sgn}(y - \sin(x^4)\sin(t^4))$, where $\mathrm{sgn}(u) = 1$ for $u > 0$ and $\mathrm{sgn}(u) = -1$ for $u < 0$ (see figure 3). As can be seen from the illustration, the result of detecting the strongest interest points highly depends on the choice of the scale parameters σ_l^2 and τ_l^2. We can observe that space-time structures with long temporal extents are detected for large values of τ_l^2 while short events are preferred by the detector with small values of

τ_l^2. Similarly, the spatial extent of the events is related to the value of the spatial scale parameter σ_l^2.

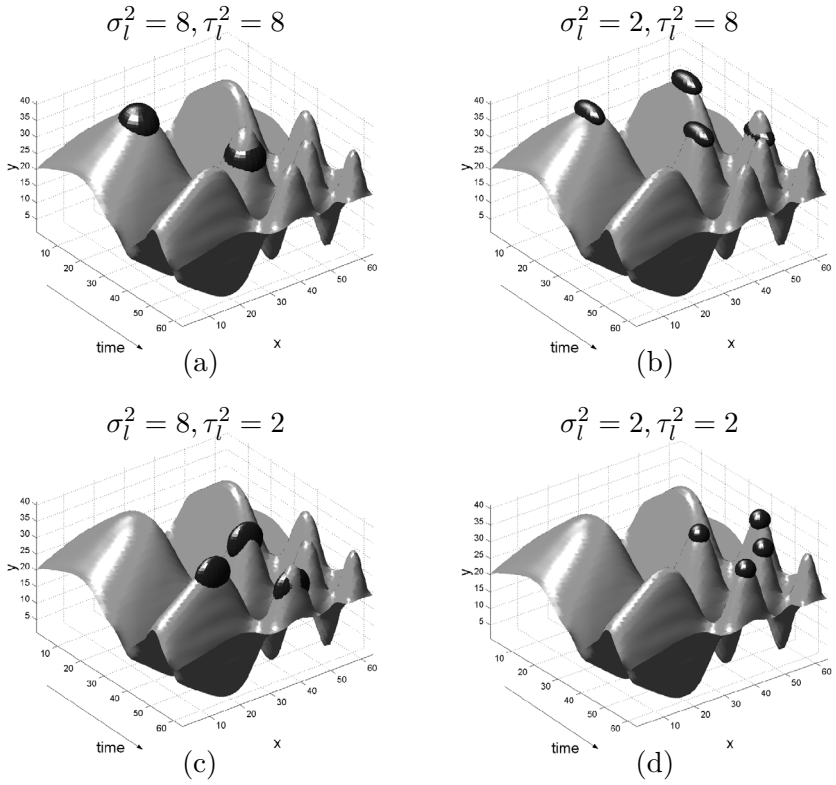


Figure 3: Results of detecting interest point at different spatial and temporal scales for a synthetic sequence with impulses having varying extents in space and time. The extents of the detected events roughly corresponds to the scale parameters σ_l^2 and τ_l^2 used for computing H (8).

From the presented examples, we can conclude that a correct selection of temporal and spatial scales is crucial when capturing events with different spatial and temporal extents. Moreover, estimating the spatio-temporal extents of events is important for their further interpretation. In the next section, we will present a mechanism for simultaneous estimation of spatio-temporal scales. This mechanism will then be combined with the interest point detector resulting in scale-adapted interest points in section 3.2.

3 Spatio-temporal scale adaptation

3.1 Scale selection in space-time

During recent years, the problem of automatic scale selection has been addressed in several different ways, based on the maximization of normalized derivative expressions over scale, or the behavior of entropy measures or error measures over scales (see Lindeberg and Bretzner (2003) for a review). To estimate the spatio-temporal extent of an event in

space-time, we follow works on local scale selection proposed in the spatial domain by Lindeberg (1998) as well as in the temporal domain (Lindeberg, 1997). The idea is to define a differential operator that assumes simultaneous extrema over spatial and temporal scales that are characteristic for an event with a particular spatio-temporal location.

For the purpose of analysis, we will first study a prototype event represented by a spatio-temporal Gaussian blob

$$f(x, y, t;\, \sigma_0^2, \tau_0^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}} \exp(-(x^2 + y^2)/2\sigma_0^2 - t^2/2\tau_0^2)$$

with spatial variance σ_0^2 and temporal variance τ_0^2 (see figure 4a). Using the semi-group property of the Gaussian kernel, it follows that the scale-space representation of f is

$$L(\cdot;\, \sigma^2, \tau^2) = g(\cdot;\, \sigma^2, \tau^2) * f(\cdot;\, \sigma_0^2, \tau_0^2) = g(\cdot;\, \sigma_0^2 + \sigma^2, \tau_0^2 + \tau^2).$$

To recover the spatio-temporal extent (σ_0, τ_0) of f, we consider second-order derivatives of L normalized by the scale parameters as follows

$$L_{xx,norm} = \sigma^{2a} \tau^{2b} L_{xx}, \quad L_{yy,norm} = \sigma^{2a} \tau^{2b} L_{yy}, \quad L_{tt,norm} = \sigma^{2c} \tau^{2d} L_{tt}. \tag{9}$$

All of these entities assume local extrema over space and time at the center of the blob f. Moreover, depending on the parameters a, b and c, d, they also assume local extrema over scales at certain spatial and temporal scales, $\tilde{\sigma}^2$ and $\tilde{\tau}^2$.

The idea of scale selection we follow here is to determine the parameters a, b, c, d such that $L_{xx,norm}$, $L_{yy,norm}$ and $L_{tt,norm}$ assume extrema at scales $\tilde{\sigma}^2 = \sigma_0^2$ and $\tilde{\tau}^2 = \tau_0^2$. To find such extrema, we differentiate the expressions in (9) with respect to the spatial and the temporal scale parameters. For the spatial derivatives we obtain the following expressions at the center of the blob

$$\frac{\partial}{\partial \sigma^2} \Big[L_{xx,norm}(0, 0, 0;\, \sigma^2, \tau^2) \Big] = -\frac{a\sigma^2 - 2\sigma^2 + a\sigma_0^2}{\sqrt{(2\pi)^3 (\sigma_0^2 + \sigma^2)^6 (\tau_0^2 + \tau^2)}}\, \sigma^{2(a-1)} \tau^{2b} \tag{10}$$

$$\frac{\partial}{\partial \tau^2} \Big[L_{xx,norm}(0, 0, 0;\, \sigma^2, \tau^2) \Big] = -\frac{2b\tau_0^2 + 2b\tau^2 - \tau^2}{\sqrt{2^5 \pi^3 (\sigma_0^2 + \sigma^2)^4 (\tau_0^2 + \tau^2)^3}}\, \tau^{2(b-1)} \sigma^{2a}. \tag{11}$$

By setting these expressions to zero, we obtain the following simple relations for a and b

$$a\sigma^2 - 2\sigma^2 + a\sigma_0^2 = 0, \quad 2b\tau_0^2 + 2b\tau^2 - \tau^2 = 0$$

which after substituting $\sigma^2 = \sigma_0^2$ and $\tau^2 = \tau_0^2$ lead to $a = 1$ and $b = 1/4$. Similarly, differentiating the second-order temporal derivative

$$\frac{\partial}{\partial \sigma^2} \Big[L_{tt,norm}(0, 0, 0;\, \sigma^2, \tau^2) \Big] = -\frac{c\sigma^2 - \sigma^2 + c\sigma_0^2}{\sqrt{(2\pi)^3 (\sigma_0^2 + \sigma^2)^4 (\tau_0^2 + \tau^2)^3}}\, \sigma^{2(c-1)} \tau^{2d} \tag{12}$$

$$\frac{\partial}{\partial \tau^2} \Big[L_{tt,norm}(0, 0, 0;\, \sigma^2, \tau^2) \Big] = -\frac{2d\tau_0^2 + 2d\tau^2 - 3\tau^2}{\sqrt{2^5 \pi^3 (\sigma_0^2 + \sigma^2)^2 (\tau_0^2 + \tau^2)^5}}\, \tau^{2(d-1)} \sigma^{2c} \tag{13}$$

leads to the expressions

$$c\sigma^2 - 2\sigma^2 + c\sigma_0^2 = 0, \quad 2d\tau_0^2 + 2d\tau^2 - \tau^2 = 0$$

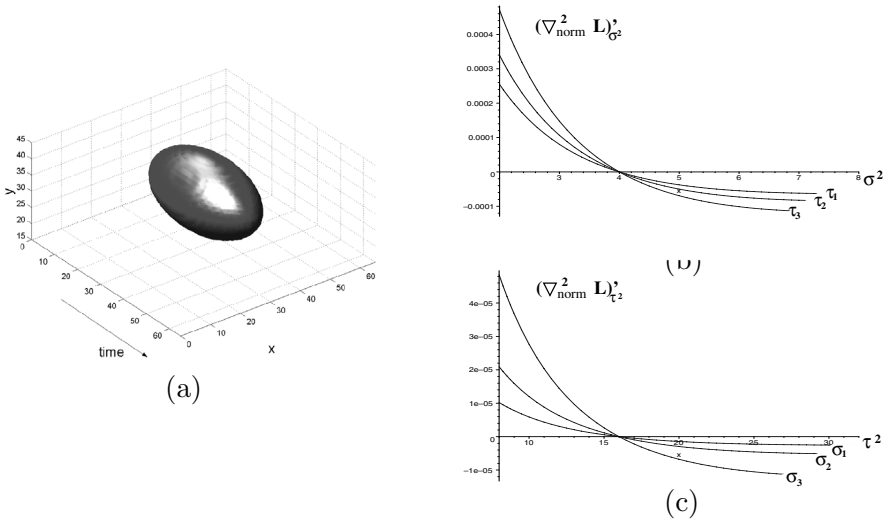


Figure 4: (a): A Spatio-temporal Gaussian blob with spatial variance $\sigma_0^2 = 4$ and temporal variance $\tau_0^2 = 16$; (b)-(c) derivatives of $\nabla^2_{norm} L$ with respect to scales. The zero-crossings of $(\nabla^2_{norm} L)'_{\sigma^2}$ and $(\nabla^2_{norm} L)'_{\tau^2}$ indicate extrema of $\nabla^2_{norm} L$ at scales corresponding to the spatial and the temporal extents of the blob.

which after substituting $\sigma^2 = \sigma_0^2$ and $\tau^2 = \tau_0^2$ result in $c = 1/2$ and $d = 3/4$.

The normalization of derivatives in (9) guarantees that all these partial derivative expressions assume local space-time-scale extrema at the center of the blob f and at scales corresponding to the spatial and the temporal extents of f, i.e. $\sigma = \sigma_0$ and $\tau = \tau_0$. From the sum of these partial derivatives, we then define a normalized spatio-temporal Laplace operator according to

$$\begin{aligned} \nabla^2_{norm} L &= L_{xx,norm} + L_{yy,norm} + L_{tt,norm} \\ &= \sigma^2 \tau^{1/2} (L_{xx} + L_{yy}) + \sigma \tau^{3/2} L_{tt}. \end{aligned} \tag{14}$$

Figures 4b-c show derivatives of this operator with respect to the scale parameters evaluated at the center of a spatio-temporal blob with spatial variance $\sigma_0^2 = 4$ and temporal variance $\tau_0^2 = 16$. The zero-crossings of the curves verify that $\nabla^2_{norm} L$ assumes extrema at the scales $\sigma^2 = \sigma_0^2$ and $\tau^2 = \tau_0^2$. Hence, the spatio-temporal extent of the Gaussian prototype can be estimated by finding the extrema of $\nabla^2_{norm} L$ over both spatial and temporal scales. In the following section, we will use this operator for estimating the extents of other spatio-temporal structures, in analogy with previous work of using the normalized Laplacian operator as a general tool for estimating the spatial extent of image structures in the spatial domain.

3.2 Scale-adapted space-time interest points

Local scale estimation using the normalized Laplace operator has shown to be very useful in the spatial domain (Lindeberg, 1998; Almansa and Lindeberg, 2000; Chomat, de Verdiere, Hall and Crowley, 2000a). In particular, Mikolajczyk and Schmid (2001) combined the

Harris interest point operator with the normalized Laplace operator and derived a scale-invariant Harris-Laplace interest point detector. The idea is to find points in scale-space that are both spatial maxima of the Harris function H^{sp} (4) and extrema over scale of the scale-normalized Laplace operator in space.

Here, we extend this idea and detect interest points that are simultaneous maxima of the spatio-temporal corner function H (8) over space and time (x, y, t) as well as extrema of the normalized spatio-temporal Laplace operator $\nabla^2_{norm} L$ (14) over scales (σ^2, τ^2). One way of detecting such points is to compute space-time maxima of H for each spatio-temporal scale level and then to select points that maximize $(\nabla^2_{norm} L)^2$ at the corresponding scale. This approach, however, requires dense sampling over the scale parameters and is therefore computationally expensive.

An alternative we follow here, is to detect interest points for a set of sparsely distributed scale values and then to track these points in the spatio-temporal scale-time-space towards the extrema of $\nabla^2_{norm} L$. We do this by iteratively updating the scale and the position of the interest points by (i) selecting the neighboring spatio-temporal scale that maximizes $(\nabla^2_{norm} L)^2$ and (ii) re-detecting the space-time location of the interest point at a new scale. Thus, instead of performing a simultaneous maximization of H and $\nabla^2_{norm} L$ over five dimensions $(x, y, t, \sigma^2, \tau^2)$, we implement the detection of local maxima by splitting the space-time dimensions (x, y, t) and scale dimensions (σ^2, τ^2) and iteratively optimizing over the subspaces until the convergence has been reached.[2] The corresponding algorithm is presented in figure 5.

The result of scale-adaptation of interest points for the spatio-temporal pattern in figure 3 is shown in figure 6. As can be seen, the chosen scales of the adapted interest points match the spatio-temporal extents of the corresponding structures in the pattern.

It should be noted, however, that the presented algorithm has been developed for processing pre-recorded video sequences. In real-time situations, when using causal scale-space representation based on recursive temporal filters (Lindeberg and Fagerström, 1996; Lindeberg, 2002), only a fixed set of discrete temporal scales is available at any moment. In that case an approximate estimate of temporal scale can still be found by choosing interest points that maximize $(\nabla^2_{norm} L)^2$ in a local neighborhood of the spatio-temporal scale-space; see also (Lindeberg, 1997) for a treatment of automatic scale selection for time-causal scale-spaces.

3.3 Experiments

In this section, we investigate the performance of the proposed scale-adapted spatio-temporal interest point detector applied to real image sequences. In the first example, we consider a sequence of a walking person with non-constant image velocities due to the oscillating motion of the legs. As can be seen in figure 7, the spatio-temporal image pattern gives rise to stable interest points. Note that the detected interest points reflect well-localized events in both space and time, corresponding to specific space-time structures of the leg. From the space-time plot in figure 7(a), we can also observe how the

[2] For the experiments presented in this paper, with image sequences of spatial resolution 160×120 pixels and temporal sampling frequency 25Hz (totally up to 200 frames per sequence), we initialized the detection of interest points using combinations of spatial scales $\sigma_l^2 = [2, 4, 8]$ and temporal scales $\sigma_l^2 = [2, 4, 8]$, while using $s = 2$ for the ratio between the integration and the local scale when computing the second-moment matrix.

1. Detect interest points $p_j = (x_j, y_j, t_j, \sigma_{l,j}^2, \tau_{l,j}^2)$, $j = 1..N$ as maxima of H (8) over space and time using sparsely selected combinations of initial spatial scales $\sigma_l^2 = \sigma_{l,1}^2, .., \sigma_{l,n}^2$ and temporal scales $\tau_l^2 = \tau_{l,1}^2, .., \tau_{l,m}^2$ as well as integration scales $\sigma_i^2 = s\sigma_l^2$ and $\tau_i^2 = s\tau_l^2$.

2. **for** each interest point p_j **do**

3. Compute $\nabla_{norm}^2 L$ at position (x_j, y_j, t_j) and combinations of neighboring scales $(\tilde{\sigma}_{i,j}^2, \tilde{\tau}_{i,j}^2)$ where $\tilde{\sigma}_{i,j}^2 = 2^\delta \sigma_{i,j}^2$, $\tilde{\tau}_{i,j}^2 = 2^\delta \tau_{i,j}^2$, and $\delta = -0.25, 0, 0.25$

5. Choose the combination of integration scales $(\tilde{\sigma}_{i,j}^2, \tilde{\tau}_{i,j}^2)$ that maximizes $(\nabla_{norm}^2 L)^2$

6. **if** $\tilde{\sigma}_{i,j}^2 \neq \sigma_{i,j}^2$ or $\tilde{\tau}_{i,j}^2 \neq \tau_{i,j}^2$

Re-detect interest point $\tilde{p}_j = (\tilde{x}_j, \tilde{y}_j, \tilde{t}_j, \tilde{\sigma}_{l,j}^2, \tilde{\tau}_{l,j}^2)$ using integration scales $\tilde{\sigma}_{i,j}^2 = \tilde{\sigma}_{i,j}^2$, $\tilde{\tau}_{i,j}^2 = \tilde{\tau}_{i,j}^2$, local scales $\tilde{\sigma}_{l,j}^2 = \frac{1}{s}\tilde{\sigma}_{i,j}^2$, $\tilde{\tau}_{l,j}^2 = \frac{1}{s}\tilde{\tau}_{i,j}^2$ and position $(\tilde{x}_j, \tilde{y}_j, \tilde{t}_j)$ that is closest to (x_j, y_j, t_j);
set $p_j := \tilde{p}_j$ and **goto 3**

7. **end**

Figure 5: Algorithm for scale adaption of spatio-temporal interest points.

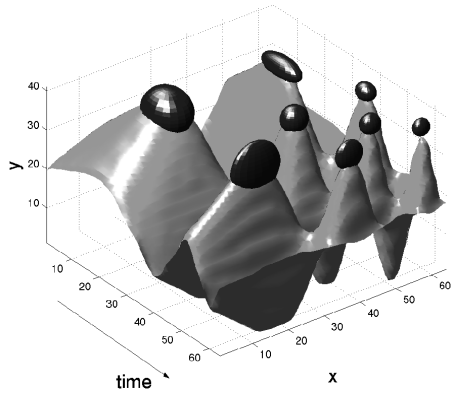
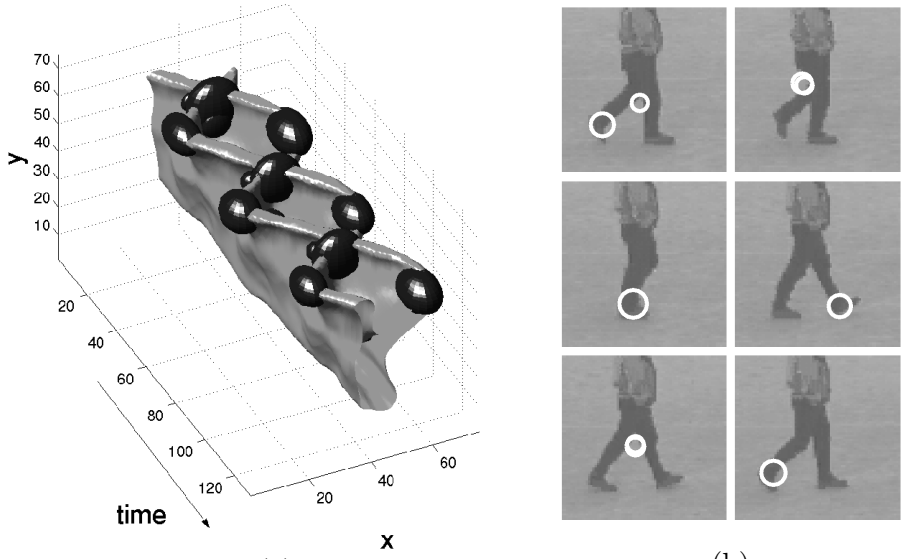


Figure 6: The result of scale-adaptation of spatio-temporal interest points computed from a space-time pattern of the form $f(x, y, t) = \text{sgn}(y - \sin(x^4) * \sin(t^4))$. The interest points are illustrated as ellipsoids showing the selected spatio-temporal scales overlayed on a surface plot of the intensity landscape.



(a) (b)

Figure 7: Results of detecting spatio-temporal interest points from the motion of the legs of a walking person. (a): 3-D plot with a thresholded level surface of a leg pattern (here shown upside down to simplify interpretation) and the detected interest points illustrated by ellipsoids; (b): spatio-temporal interest points overlayed on single frames in the original sequence.

selected spatial and temporal scales of the detected features roughly match the spatio-temporal extents of the corresponding image structures.

The top rows of figure 8 show interest points detected in an outdoor sequence with a walking person and a zooming camera. The changing values of the selected spatial scales (illustrated by the size of the circles) illustrate the invariance of the method with respect to spatial scale changes of the image structures. Note that besides events in the leg pattern, the detector finds spurious points due to the non-constant motion of the coat and the arms. Image structures with constant motion in the background, however, do not result in the response of the detector. The pure spatial interest operator[3] on the contrary gives strong responses in the static background as shown at the bottom row of figure 8

The third example explicitly illustrates how the proposed method is able to estimate the temporal extent of detected events. Figure 9 shows a person making hand-waving gestures with a high frequency on the left and a low frequency on the right. The distinct interest points are detected at the moments and at the spatial positions where the palm of a hand changes its direction of motion. Whereas the spatial scale of the detected interest points remains constant, the selected temporal scale depends on the frequency of the wave pattern. The high frequency pattern results in short events and gives rise to interest points with small temporal extent (see figure 9a). Correspondingly, hand motions with low frequency result in interest points with long temporal extent as shown in figure 9b.

[3]Here, we used the scale-adapted Harris interest point detector (Mikolajczyk and Schmid, 2001) that detects maxima of H^{sp} (4) in space and extrema of normalized Laplacian operator over scales (Lindeberg, 1998).

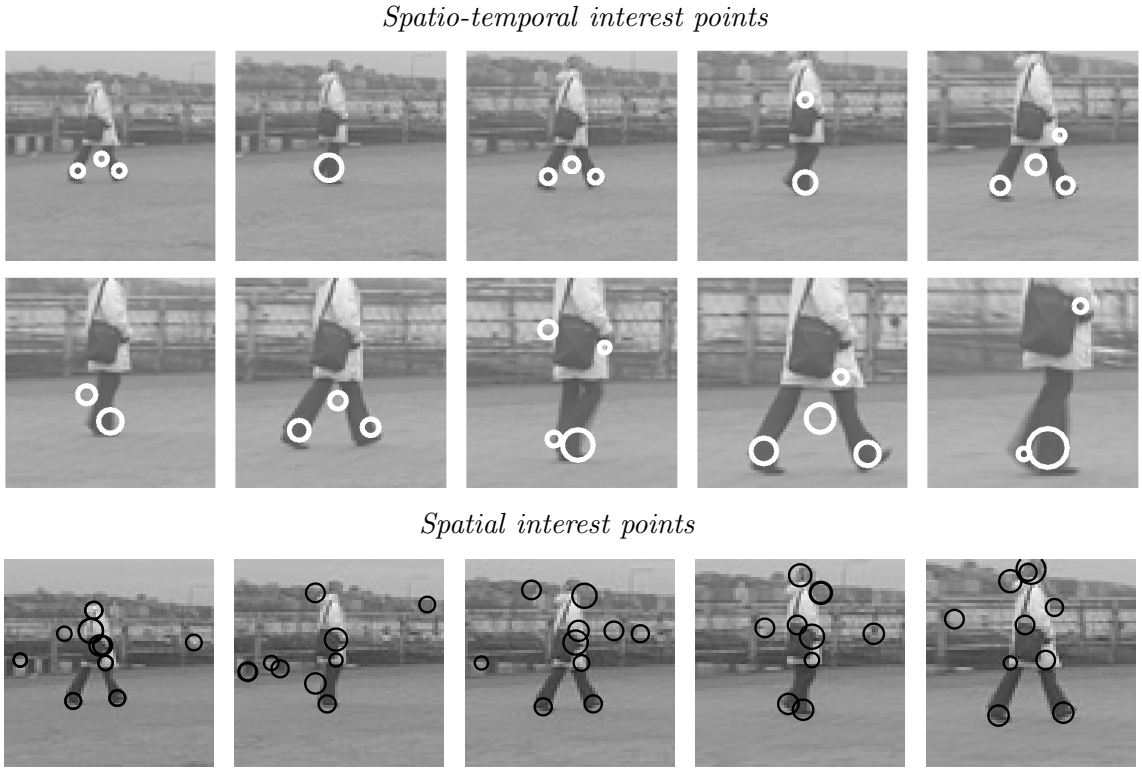


Figure 8: Top: Results of spatio-temporal interest point detection for a zoom-in sequence of a walking person. The spatial scale of the detected points (corresponding to the size of circles) matches the increasing spatial extent of the image structures and verifies the invariance of the interest points with respect to changes in spatial scale. Bottom: Pure spatial interest point detector (here, Harris-Laplace) selects both moving and stationary points and is less restrictive.

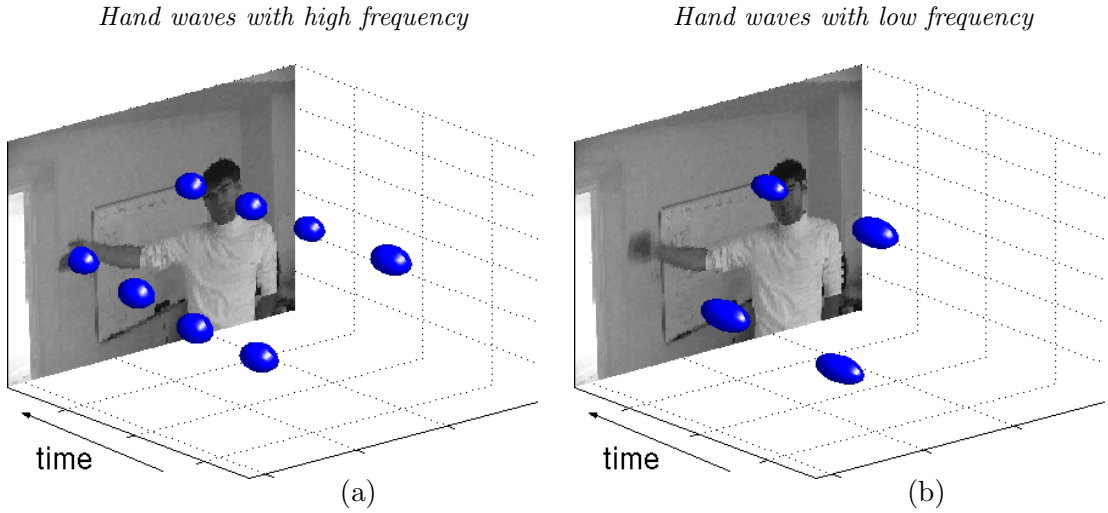


Figure 9: Result of interest point detection for a sequence with waving hand gestures: (a) Interest points for hand movements with high frequency; (b) Interest points for hand movements with low frequency.

4 Classification of events

The detected interest points have significant variations of image values in a local spatio-temporal neighborhood. To differentiate events from each other and from noise, one approach is to compare local neighborhoods and assign points with similar neighborhoods to the same class of events. A similar approach has proven to be highly successful in the spatial domain for the task of image representation (Malik, Belongie, Shi and Leung, 1999) indexing (Schmid and Mohr, 1997) and recognition (Hall et al., 2000; Weber, Welling and Perona, 2000; Leung and Malik, 2001). In the spatio-temporal domain, local descriptors have been previously used by (Chomat et al., 2000b) and others.

To describe a spatio-temporal neighborhood, we consider normalized spatio-temporal Gaussian derivatives defined as

$$L_{x^m y^n t^k} = \sigma^{m+n} \tau^k (\partial_{x^m y^n t^k} g) * f, \tag{15}$$

computed at the scales used for detecting the corresponding interest points. The normalization with respect to scale parameters guarantees the invariance of the derivative responses with respect to image scalings in both the spatial domain and the temporal domain. Using derivatives, we define event descriptors from the third order local jet[4] (Koenderink and van Doorn, 1987) computed at spatio-temporal scales determined from the detection scales of the corresponding interest points

$$j = (L_x, L_y, L_t, L_{xx}, ..., L_{ttt})\Big|_{\sigma^2=\tilde{\sigma}_i^2, \tau^2=\tilde{\tau}_i^2} \tag{16}$$

To compare two events, we compute the Mahalanobis distance between their descriptors as

$$d^2(j_1, j_2) = (j_1 - j_2)\Sigma^{-1}(j_1 - j_2)^T, \tag{17}$$

where Σ is a covariance matrix corresponding to the typical distribution of interest points in training data.

To detect similar events in the data, we apply k-means clustering (Duda, Hart and Stork, 2001) in the space of point descriptors and detect groups of points with similar spatio-temporal neighborhoods. Thus clustering of spatio-temporal neighborhoods is similar to the idea of textons (Malik et al., 1999) used to describe image texture as well as to detect object parts for spatial recognition (Weber et al., 2000). Given training sequences with periodic motion, we can expect repeating events to give rise to populated clusters. On the contrary, sporadic interest points can be expected to be sparsely distributed over the descriptor space giving rise to weakly populated clusters. To test this idea we applied k-means clustering with $k = 15$ to the sequence of a walking person in the upper row of figure 11. We found out that the four most densely populated clusters $c1, ..., c4$ indeed corresponded to stable interest points of the gait pattern. Local spatio-temporal neighborhoods of these points are shown in figure 10, where we can confirm the similarity of patterns inside the clusters and their difference between clusters.

To represent characteristic repetitive events in video, we compute cluster means $m_i = \frac{1}{n_i} \sum_{k=1}^{n_i} j_k$ for each significant cluster c_i consisting of n_i points. Then, in order to classify

[4]Note that our representation is currently not invariant with respect to planar image rotations. Such invariance could be added by considering steerable derivatives or rotationally invariant operators in space.

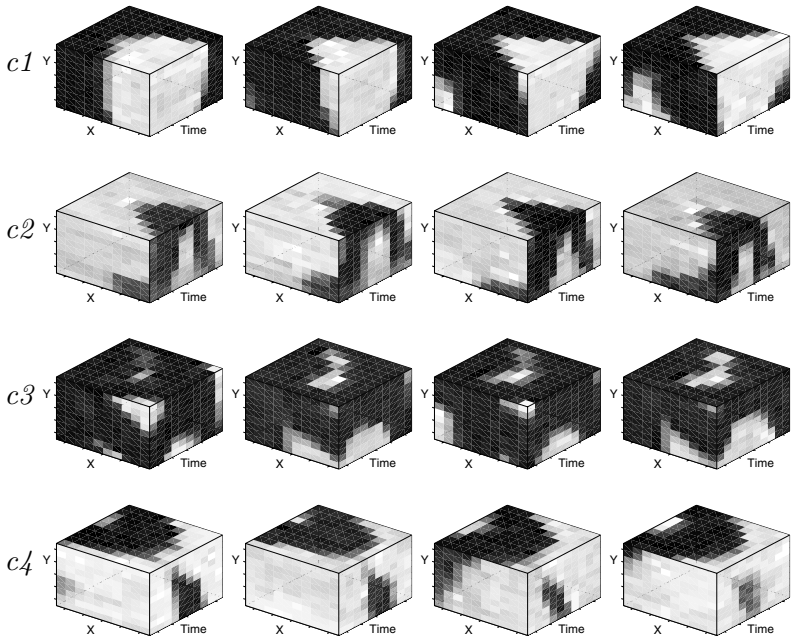


Figure 10: Local spatio-temporal neighborhoods of interest points corresponding to the first four most populated clusters obtained from a sequence of walking person.

an event on an unseen sequence, we assign the detected point to the cluster c_i that minimizes the distance $d(m_i, j_0)$ (17) between the jet of the interest point j_0 and the cluster mean m_i. If the distance is above a threshold, the point is classified as background. An application of this classification scheme is demonstrated in the second row of figure 11. As can be seen, most of the points corresponding to the gait pattern are correctly classified, while the other interest points are discarded. Observe that the person in the second sequence of figure 11 undergoes significant size changes in the image. Due to the scale-invariance of the interest points as well as their jet responses, the size transformations do not effect neither the result of event detection nor the result of classification.

5 Application to video interpretation

In this section, we illustrate how a sparse representation of video sequences by classified spatio-temporal interest points can be used for video interpretation. We consider the problem of detecting walking people and estimating their poses when viewed from the side in outdoor scenes. Such a task is complicated, since the variations in appearance of people together with the variations in the background may lead to ambiguous interpretations. Human motion is a strong cue that has been used to resolve this ambiguity in a number of previous works. Some of the works rely on pure spatial image features while using sophisticated body models and tracking schemes to constrain the interpretation (Baumberg and Hogg, 1996; Bregler and Malik, 1998; Sidenbladh, Black and Fleet, 2000). Other approaches use spatio-temporal image cues such as optical flow (Black, Yacoob, Jepson and

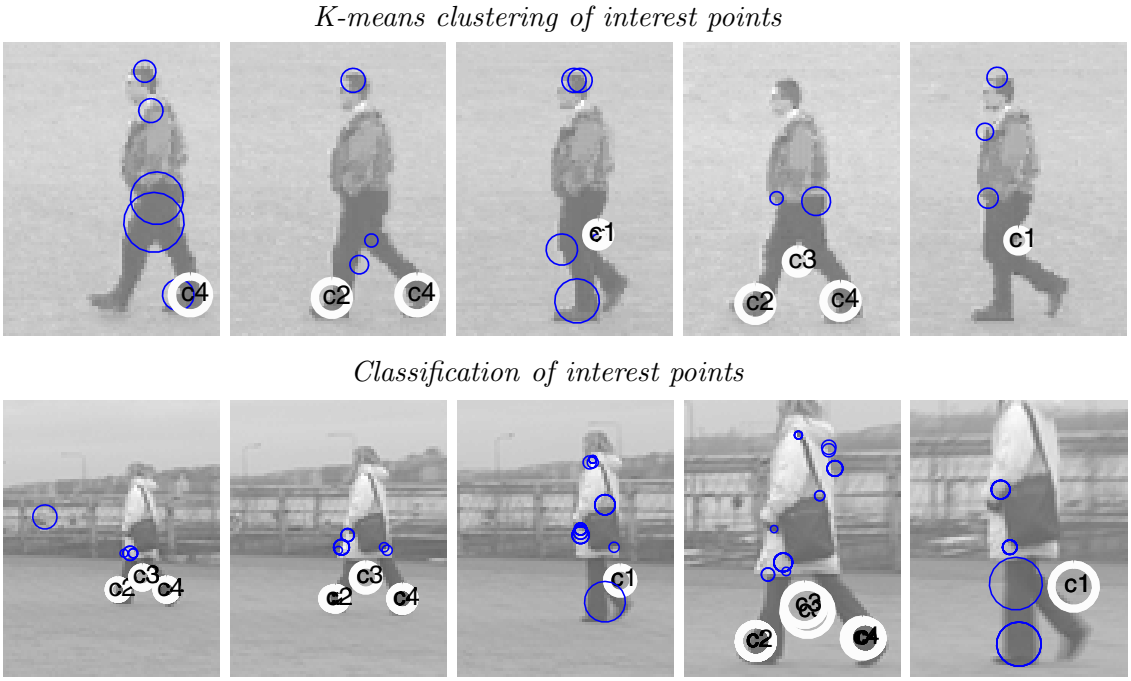


Figure 11: Interest points detected for sequences of walking persons. First row: the result of clustering spatio-temporal interest points in training data. The labelled points correspond to the four most populated clusters; Second row: the result of classifying interest points with respect to the clusters found in the first sequence.

Fleet, 1997) or motion templates (Baumberg and Hogg, 1996). The work of Niyogi and Adelson (1994) concerns the structure of the spatio-temporal gait pattern and is closer to ours.

The idea of the following approach is to represent both the model and the data using local and discriminative spatio-temporal features and to match the model by matching its features to the correspondent features of the data inside a spatio-temporal window (see figure 12).

5.1 Walking model

To obtain a model of a walking person, we consider the upper sequence in figure 11 and manually select a time interval (t_0, t_0+T) corresponding to the period T of the gait pattern. Then, given n features $f_i^m = (x_i^m, y_i^m, t_i^m, \sigma_i^m, \tau_i^m, c_i^m)$, $i = 1, ..., n$ (m stands for model) defined by the positions (x_i^m, y_i^m, t_i^m), scales (σ_i^m, τ_i^m) and classes c_i^m of interest points detected in the selected time interval, i.e. $t_i^m \in (t_0, t_0+T)$, we define the walking model by a set of periodically repeating features $M = \{f_i + (0, 0, kT, 0, 0, 0, 0) | i = 1, ..., n, k \in \mathcal{Z}\}$. Furthermore, to account for variations of the position and the size of a person in the image, we introduce a state for the model determined by the vector $X = (x, y, \theta, s, \xi, v_x, v_y, v_s)$. The components of X describe the position of the person in the image (x, y), his size s, the frequency of the gait ξ, the phase of the gait cycle θ at the current time moment as well as the temporal variations (v_x, v_y, v_s) of (x, y, s); v_x and v_y describe the velocity in

the image domain, while v_s describes how fast size changes occur. Given the state X, the parameters of each model feature $f \in M$ transform according to

$$\begin{aligned} \tilde{x}^m &= x + sx^m + \xi v_x(t^m + \theta) + s\xi x^m v_s(t^m + \theta) \\ \tilde{y}^m &= y + sy^m + \xi v_y(t^m + \theta) + s\xi y^m v_s(t^m + \theta) \\ \tilde{t}^m &= \xi(t^m + \theta) \\ \tilde{\sigma}^m &= s\sigma^m + v_s s\sigma^m(t^m + \theta) \\ \tilde{\tau}^m &= \xi\tau^m \\ \tilde{c}^m &= c^m \end{aligned} \tag{18}$$

It follows that this type of scheme is able to handle translations and uniform rescalings in the image domain as well as uniform rescalings in the temporal domain. Hence, it allows for matching of patterns with different image velocities as well as with different frequencies over time.

To estimate the boundary of the person, we extract silhouettes $S = \{x^s, y^s, \theta^s | \theta^s = 1, ..., T\}$ on the model sequence (see figure 11) one for each frame corresponding to the discrete value of the phase parameter θ. The silhouette is used here only for visualization purpose and allows us to approximate the boundary of the person in the current frame using the model state X and a set of points $\{(x^s, y^s, \theta^s) \in S | \theta^s = \theta\}$ transformed according to $\tilde{x}^s = sx^s + x$, $\tilde{y}^s = sy^s + y$.

5.2 Model matching

Given a model state X, a current time t_0, a length of the time window t_w, and a set of data features $D = \{f^d = (x^d, y^d, t^d, \sigma^d, \tau^d, c^d) | t^d \in (t_0, t_0 - t_w)\}$ detected from the the recent time window of the data sequence, the match between the model and the data is defined by a weighted sum of distances h between the model features f_i^m and the data features f_j^d

$$\mathcal{H}(\tilde{M}(X), D, t_0) = \sum_i^n h(\tilde{f}_i^m, f_j^d) e^{-(\tilde{t}_i^m - t_0)^2/\xi}, \tag{19}$$

where $\tilde{M}(X)$ is a set of n model features in the time window $(t_0, t_0 - t_w)$ transformed according to (18), i.e. $\tilde{M} = \{\tilde{f}^m | t^m \in (t_0, t_0 - t_w)\}$, $f_j^d \in D$ is a data feature minimizing the distance h for a given f_i^m and ξ is the variance of the exponential weighting function that gives more importance to recent features.

The distance h between two features of the same class is defined as a Euclidean distance between two points in space-time, where the spatial and the temporal dimensions are weighted with respect to a parameter ν as well as by the extents of the features in space-time

$$h^2(f^m, f^d) = (1 - \nu)\frac{(x^m - x^d)^2 + (y^m - y^d)^2}{(\sigma^m)^2} + \nu\frac{(t^m - t^d)^2}{(\tau^m)^2}. \tag{20}$$

Here, the distance between features of different classes is regarded as infinite. Alternatively, one could measure the feature distance by taking into account their descriptors and distances from several of the nearest cluster means.

To find the best match between the model and the data, we search for the model state $\tilde{X}$ that minimizes $\mathcal{H}$ in (19)

$$\tilde{X} = \text{argmin}_X \, \mathcal{H}(\tilde{M}(X), D, t_0) \tag{21}$$

using a standard Gauss-Newton optimization method. The result of such an optimization for a sequence with data features in figure 12(a) is illustrated in figure 12(b). Here, the match between the model and the data features was searched over a time window corresponding to three periods of the gait pattern or approximately 2 seconds of video. As can be seen from figure 12(c), the overlaps between the model features and the data features confirm the match between the model and the data. Moreover, the model silhouette transformed according to $\tilde{X}$ matches with the contours of the person in the current frame and confirms a reasonable estimate of the model parameters.

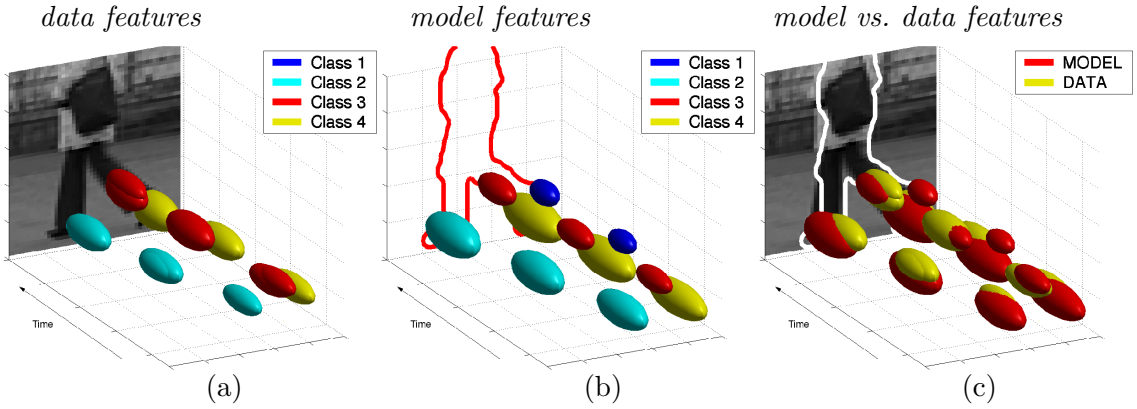


Figure 12: Matching of spatio-temporal data features with model features: (a) Features detected from the data sequence over a time interval corresponding to three periods of the gait cycle; (b) Model features minimizing the distance to the features in (a); (c) Model features and data features overlaid. The estimated silhouette overlayed on the current frame confirms the correctness of the method.

5.3 Results

Figure 13 presents results of the described approach applied to two outdoor sequences. The first sequence illustrates the invariance of the method with respect to size variations of the person in the image plane. The second sequence shows the successful detection and pose estimation of a person despite the presence of a complex non-stationary background and occlusions. Note that these results have been obtained by re-initializing model parameters before optimization at each frame. Hence, the approach is highly stable and could be improved further by tracking the model parameters $\tilde{X}$ over time.

The need for careful initialization and/or simple background are frequent obstacles in previous approaches for human motion analysis. The success of our method is due to the low ambiguity and simplicity of the matching scheme originating from the distinct and stable nature of the spatio-temporal features. In this respect, we want to propose direct detection of spatio-temporal events as an interesting alternative when representing and interpreting video data.

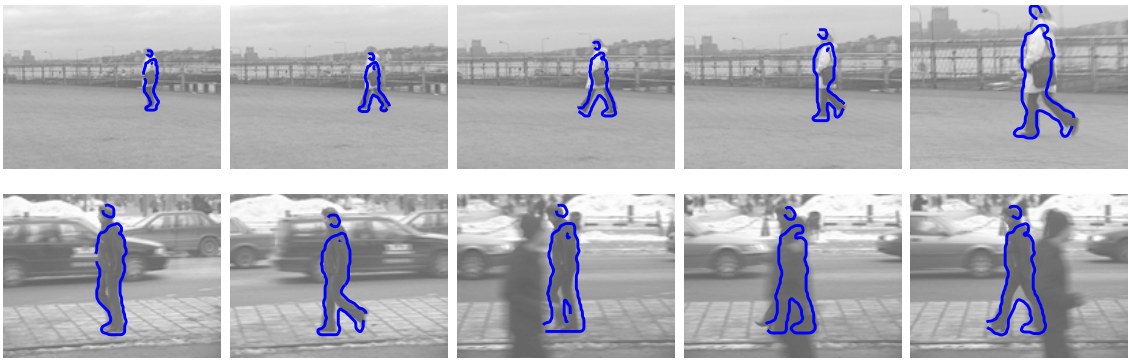


Figure 13: The result of matching a spatio-temporal walking model to sequences of outdoor scenes.

6 Summary

We have described an interest point detector that finds local image features in space-time characterized by a high variation of the image values in space and non-constant motion over time. From the presented examples, it follows that many of the detected points indeed correspond to meaningful events. Moreover, we propose local maximization of the normalized spatio-temporal Laplacian operator as a general tool for scale selection in space-time. Using this mechanism, we estimated characteristic spatio-temporal extents of detected events and computed their scale-invariant spatio-temporal descriptors.

Using scale-adapted descriptors in terms of N-jets we then addressed the problem of event classification and illustrated how classified spatio-temporal interest points constitute distinct and stable descriptors of events in video, which can be used for video representation and interpretation. In particular, we have shown how a video representation by spatio-temporal interest points enables detection and pose estimation of walking people in the presence of occlusions and highly cluttered and dynamic background. Note that this result was obtained using a standard optimization method without careful manual initialization or tracking.

In future work, we plan to extend application of interest points to the field of motion-based recognition (Schüldt, Laptev and Caputo, 2004). Moreover, as the current scheme of event detection is not invariant under Galilean transformations, future work should investigate the possibilities of including such invariance and making the approach independent of the relative camera motion (Laptev and Lindeberg, 2002; Laptev and Lindeberg, 2004). Another extension should consider the invariance of spatio-temporal descriptors with respect to the direction of motion, changes in image contrast and rotations. Finally, other types of space-time interest operators will be considered and investigated (Lindeberg, Akbarzadeh and Laptev, 2004).

7 Acknowledgments

We thank Anastasiya Syromyatnikova, Josephine Sullivan and Carsten Rother for their help in obtaining video data for the experiments.

References

Almansa, A. and Lindeberg, T. (2000). Fingerprint enhancement by shape adaptation of scale-space operators with automatic scale-selection, *IEEE Transactions on Image Processing* **9**(12): 2027–2042.

Barron, J., Fleet, D. and Beauchemin, S. (1994). Performance of optical flow techniques, *International Journal of Computer Vision* **12**(1): 43–77.

Baumberg, A. M. and Hogg, D. (1996). Generating spatiotemporal models from examples, *Image and Vision Computing* **14**(8): 525–532.

Bigün, J., Granlund, G. and Wiklund, J. (1991). Multidimensional orientation estimation with applications to texture analysis and optical flow, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**(8): 775–790.

Black, M. and Jepson, A. (1998). Eigentracking: Robust matching and tracking of articulated objects using view-based representation, *International Journal of Computer Vision* **26**(1): 63–84.

Black, M., Yacoob, Y., Jepson, A. and Fleet, D. (1997). Learning parameterized models of image motion, *Proc. Computer Vision and Pattern Recognition*, pp. 561–567.

Blake, A. and Isard, M. (1998). Condensation – conditional density propagation for visual tracking, *International Journal of Computer Vision* **29**(1): 5–28.

Bregler, C. and Malik, J. (1998). Tracking people with twists and exponential maps, *Proc. Computer Vision and Pattern Recognition*, Santa Barbara, CA, pp. 8–15.

Bretzner, L. and Lindeberg, T. (1998). Feature tracking with automatic selection of spatial scales, *Computer Vision and Image Understanding* **71**(3): 385–392.

Chomat, O., de Verdiere, V., Hall, D. and Crowley, J. (2000a). Local scale selection for Gaussian based description techniques, *Proc. Sixth European Conference on Computer Vision*, Vol. 1842 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Dublin, Ireland, pp. I:117–133.

Chomat, O., Martin, J. and Crowley, J. (2000b). A probabilistic sensor for the perception and recognition of activities, *Proc. Sixth European Conference on Computer Vision*, Vol. 1842 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Dublin, Ireland, pp. I:487–503.

Duda, R., Hart, P. and Stork, D. (2001). *Pattern Classification*, Wiley.

Fergus, R., Perona, P. and Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning, *Proc. Computer Vision and Pattern Recognition*, Santa Barbara, CA, pp. II:264–271.

Fleet, D., Black, M. and Jepson, A. (1998). Motion feature detection using steerable flow fields, *Proc. Computer Vision and Pattern Recognition*, Santa Barbara, CA, pp. 274–281.

Florack, L. M. J. (1997). *Image Structure*, Kluwer Academic Publishers, Dordrecht, Netherlands.

Förstner, W. A. and Gülch, E. (1987). A fast operator for detection and precise location of distinct points, corners and centers of circular features, *Proc. Intercommission Workshop of the Int. Soc. for Photogrammetry and Remote Sensing*, Interlaken, Switzerland.

Hall, D., de Verdiere, V. and Crowley, J. (2000). Object recognition using coloured receptive fields, *Proc. Sixth European Conference on Computer Vision*, Vol. 1842 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Dublin, Ireland, pp. I:164–177.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector, *Alvey Vision Conference*, pp. 147–152.

Hoey, J. and Little, J. (2000). Representation and recognition of complex human motion, *Proc. Computer Vision and Pattern Recognition*, Hilton Head, SC, pp. I:752–759.

Koenderink, J. and van Doorn, A. (1987). Representation of local geometry in the visual system, *Biological Cybernetics* **55**: 367–375.

Koenderink, J. J. (1988). Scale-time, *Biological Cybernetics* **58**: 159–162.

Koenderink, J. J. and van Doorn, A. J. (1992). Generic neighborhood operators, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**(6): 597–605.

Laptev, I. and Lindeberg, T. (2002). Velocity-adaptation of spatio-temporal receptive fields for direct recognition of activities: An experimental study, *in* D. Suter (ed.), *Proc. ECCV'02 Workshop on Statistical Methods in Video Processing*, Copenhagen, Denmark, pp. 61–66.

Laptev, I. and Lindeberg, T. (2004). Velocity adaptation of space-time interest points, *in preparation*.

Leung, T. and Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons, *International Journal of Computer Vision* **43**(1): 29–44.

Lindeberg, T. (1994). *Scale-Space Theory in Computer Vision*, Kluwer Academic Publishers, Boston.

Lindeberg, T. (1997). On automatic selection of temporal scales in time-causal scale-space, *AFPAC'97: Algebraic Frames for the Perception-Action Cycle*, Vol. 1315 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, pp. 94–113.

Lindeberg, T. (1998). Feature detection with automatic scale selection, *International Journal of Computer Vision* **30**(2): 77–116.

Lindeberg, T. (2002). Time-recursive velocity-adapted spatio-temporal scale-space filters, *Proc. Seventh European Conference on Computer Vision*, Vol. 2350 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Copenhagen, Denmark, pp. I:52–67.

Lindeberg, T., Akbarzadeh, A. and Laptev, I. (2004). Galilean-corrected spatio-temporal interest operators, *in preparation*.

Lindeberg, T. and Bretzner, L. (2003). Real-time scale selection in hybrid multi-scale representations, *in* L. Griffin and M. Lillholm (eds), *Scale-Space'03*, Vol. 2695 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, pp. 148–163.

Lindeberg, T. and Fagerström, D. (1996). Scale-space with causal time direction, *Proc. Fourth European Conference on Computer Vision*, Vol. 1064 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Cambridge, UK, pp. I:229–240.

Lindeberg, T. and Garding, J. (1997). Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure, *Image and Vision Computing* **15**(6): 415–434.

Lowe, D. (1999). Object recognition from local scale-invariant features, *Proc. Seventh International Conference on Computer Vision*, Corfu, Greece, pp. 1150–1157.

Malik, J., Belongie, S., Shi, J. and Leung, T. (1999). Textons, contours and regions: Cue integration in image segmentation, *Proc. Seventh International Conference on Computer Vision*, Corfu, Greece, pp. 918–925.

Mikolajczyk, K. and Schmid, C. (2001). Indexing based on scale invariant interest points, *Proc. Eighth International Conference on Computer Vision*, Vancouver, Canada, pp. I:525–531.

Mikolajczyk, K. and Schmid, C. (2002). An affine invariant interest point detector, *Proc. Seventh European Conference on Computer Vision*, Vol. 2350 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Copenhagen, Denmark, pp. I:128–142.

Niyogi, S. A. (1995). Detecting kinetic occlusion, *Proc. Fifth International Conference on Computer Vision*, Cambridge, MA, pp. 1044–1049.

Niyogi, S. and Adelson, H. (1994). Analyzing and recognizing walking figures in XYT, *Proc. Computer Vision and Pattern Recognition*, pp. 469–474.

Schmid, C. and Mohr, R. (1997). Local grayvalue invariants for image retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**(5): 530–535.

Schmid, C., Mohr, R. and Bauckhage, C. (2000). Evaluation of interest point detectors, *International Journal of Computer Vision* **37**(2): 151–172.

Schüldt, C., Laptev, I. and Caputo, B. (2004). Recognizing human actions: a local SVM approach, *in preparation*.

Sidenbladh, H., Black, M. and Fleet, D. (2000). Stochastic tracking of 3D human figures using 2D image motion, *Proc. Sixth European Conference on Computer Vision*, Vol. 1843 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Dublin, Ireland, pp. II:702–718.

Smith, S. and Brady, J. (1995). ASSET-2: Real-time motion segmentation and shape tracking, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**(8): 814–820.

Tell, D. and Carlsson, S. (2002). Combining topology and appearance for wide baseline matching, *Proc. Seventh European Conference on Computer Vision*, Vol. 2350 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Copenhagen, Denmark, pp. I:68–83.

Tuytelaars, T. and Van Gool, L. (2000). Wide baseline stereo matching based on local, affinely invariant regions, *British Machine Vision Conference*, pp. 412–425.

Wallraven, C., Caputo, B. and Graf, A. (2003). Recognition with local features: the kernel recipe, *Proc. Ninth International Conference on Computer Vision*, Nice, France, pp. 257–264.

Weber, M., Welling, M. and Perona, P. (2000). Unsupervised learning of models for visual object class recognition, *Proc. Sixth European Conference on Computer Vision*, Vol. 1842 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Dublin, Ireland, pp. I:18–32.

Witkin, A. P. (1983). Scale-space filtering, *Proc. 8th Int. Joint Conf. Art. Intell.*, Karlsruhe, Germany, pp. 1019–1022.

Zelnik-Manor, L. and Irani, M. (2001). Event-based analysis of video, *Proc. Computer Vision and Pattern Recognition*, Kauai Marriott, Hawaii, pp. II:123–130.