Image and Vision Computing 31 (2013) 421–433

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

A review of motion analysis methods for human Nonverbal Communication Computing☆

Dimitris Metaxas *, Shaoting Zhang

Center for Computational Biomedicine Imaging and Modeling (CBIM), Department of Computer Science, Rutgers University, Piscataway, NJ, USA

ARTICLE INFO

Article history:
Received 6 December 2012
Received in revised form 14 March 2013
Accepted 29 March 2013

Keywords:
Nonverbal Communication Computing
Motion analysis
Face tracking
Facial expression recognition
Gesture recognition
Group activity analysis

ABSTRACT

Human Nonverbal Communication Computing aims to investigate how people exploit nonverbal aspects of their communication to coordinate their activities and social relationships. Nonverbal behavior plays important roles in message production and processing, relational communication, social interaction and networks, deception and impression management, and emotional expression. This is a fundamental yet challenging research topic. To effectively analyze Nonverbal Communication Computing, motion analysis methods have been widely investigated and employed. In this paper, we introduce the concept and applications of Nonverbal Communication Computing and also review some of the motion analysis methods employed in this area. They include face tracking, expression recognition, body reconstruction, and group activity analysis. In addition, we also discuss some open problems and the future directions of this area.

1. Introduction

Understanding how people exploit nonverbal aspects of their communication to coordinate their activities and social relationships is a fundamental scientific challenge. Deeper insights into nonverbal communication can have a profound impact on how we link theories of perception, learning, cognition and action to models of interactions and groups at the social level. Models of nonverbal behaviors in interaction are essential for collaboration tools, human–computer and virtual interaction and other assistive technologies designed to support people in real-world activities. This knowledge is also useful to develop models of the deficits of specific populations, such as autistic children, and interventions that bring them into fuller participation in communities. In general, nonverbal communication research offers high-level principles that might explain how people organize, display, adapt and understand such behaviors for communicative purposes and social goals. However, the specifics are generally not fully understood, nor is the way to translate these principles into algorithms and computer-aided communication technologies such as intelligent agents.

To model such complex dynamic processes effectively, novel computer vision and learning algorithms are needed that take into account both the heterogeneity and the dynamicity intrinsic to behavior data. As one of the most active research areas in computer vision, human motion analysis has become a widely-used tool in this area. It uses image sequences to detect and track people, and also to interpret human activities. Emerging automated methods for analyzing motion [1] have been studied and developed to enable tracking diverse human movements precisely and robustly as well as correlating multiple people's movements in interaction. Some of the applications of using motion analysis methods for Nonverbal Communication Computing include deception detection, expression recognition, sign language recognition, behavior analysis, and group activity recognition. In the following we illustrate several examples of Nonverbal Communication Computing.

Fig. 1 shows an example of deception detection during interactions using an automated motion analysis system [2]. This work investigates how degree of the interactional synchrony can signal whether an interactant is truthful or deceptive. This automated, data-driven and unobtrusive framework consists of several motion analysis methods such as face tracking, gesture detection, facial expression recognition and interactional synchrony estimation. It is able to automatically track gestures and analyze expressions of both the target interviewee and the interviewer, extract normalized meaningful synchrony features and learn classification models for deception detection. The analysis results show that these features reliably capture simultaneous synchrony. The relationship between synchrony and deception is shown to be correlated and complex.

The second example is to use an automated motion analysis system to recognize facial expressions of emotions and fatigue from sleep loss in spaceflight [3]. Specifically, this research project aims to develop non-obtrusive objective means of detecting and mitigating cognitive performance deficits, stress, fatigue, anxiety and depression for the operational setting of spaceflight. To do so, a computational model-based tracker and an emotion recognizer of the human face have been

☆ This paper has been recommended for acceptance by Matti Pietikainen.
* Corresponding author.

0262-8856/$ – see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.imavis.2013.03.005

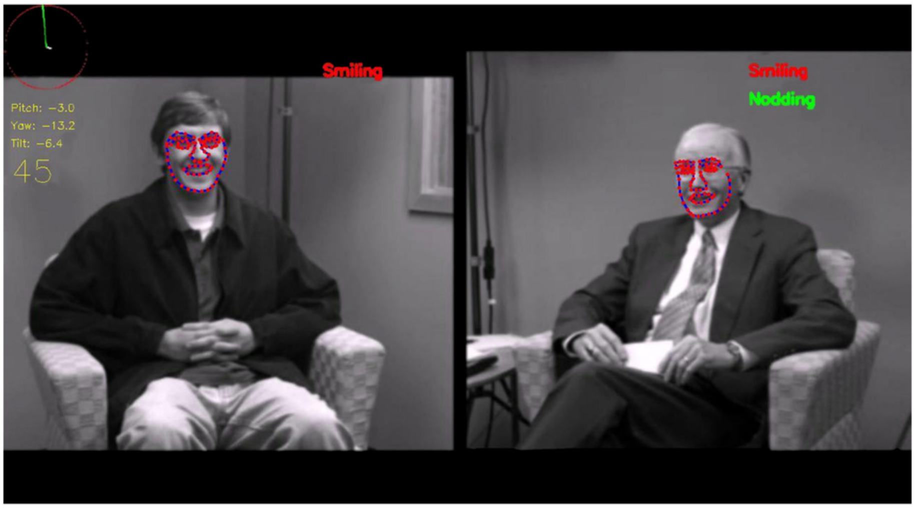

Fig. 1. An example of the automated analysis for human Nonverbal Communication Computing [2]. Sample snapshots from tracked facial data showing an interviewee (left) and an interviewer (right). Red dots represent tracked facial landmarks (eyes, eyebrows, etc.), while ellipse in top left corner depicts the estimated 3D head pose of the subject; top right corners show the detected expressions and head gestures for subject and interviewer.

developed to reliably identify when astronauts are displaying various negative emotional expressions and ocular signs of fatigue from sleep loss during space flight. Fig. 2 shows an illustration of using this system to recognize the facial expression of emotion. This subject had an emotion of sadness induced by guided recollection of negative memories. The system scored the video clip for a 2 min period. Sad was the predominant selection for the frames in the clip. This agreed with the human ratings of sadness as the dominant emotional expression during this period, as well as with the emotion induced.

The third application is an automated detection of non-manual grammatical markings in American Sign Language (ASL) [4], as shown in Fig. 3. Facial expressions and head gestures convey important linguistic information, including cues to the locations of word and phrase boundaries, emphasis on particular sentence parts, affective/emotional state, and attitude. They can offer backchannel information, regulate turn-taking, and provide indicators of speaker confidence, uncertainty, or deception. Using a robust face tracker and 3D warping [5] to extract and combine geometric and appearance features, this system can effectively recognize the eyebrows and the head gestures, as well as their temporal phases. After detecting the linguistically relevant portion of the eyebrow and periodic head gestures, it further leverages this information to improve the detection of non-manual grammatical markers in ASL.

Besides using the non-manual grammatical markings, the hand movement information can also be employed for the discrimination between fingerspelling and continuous signs in American Sign Language. In ASL, fingerspelled words are articulated using one hand (usually the dominant one) in a specific area of the signing space (in front of and slightly above the signer's shoulder). However,

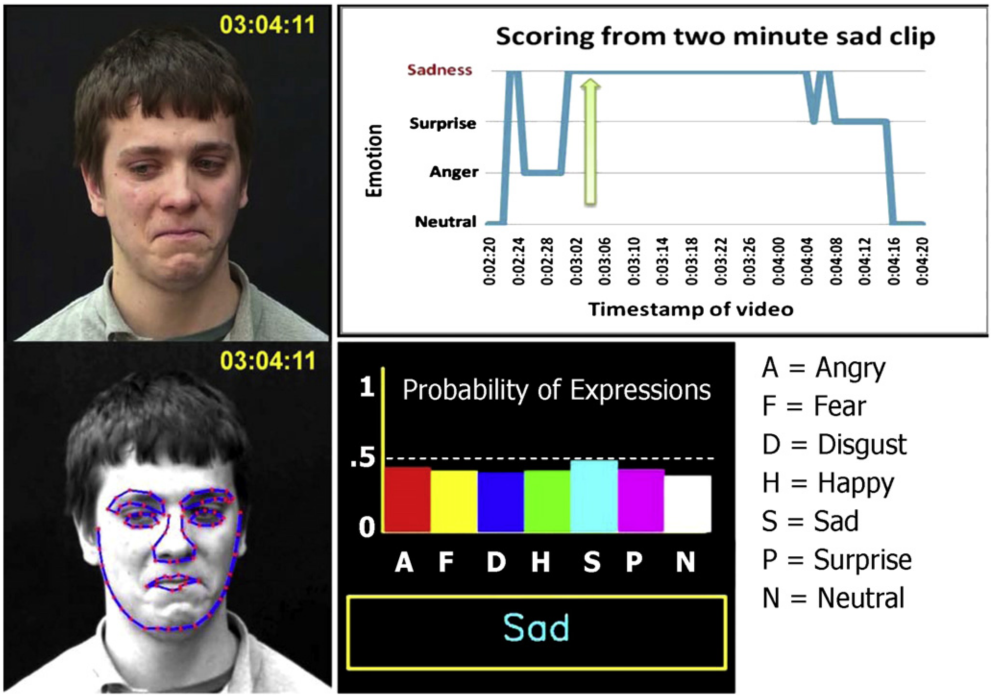

Fig. 2. A system of recognizing a specific facial expression of emotion [3]. The system scored the videos clip for a 2 min period. The graph (lower right) shows the probabilities (on the Y axis) for each of seven emotional expressions (X axis) for this specific video frame. The upward arrow in the upper right graph indicates the time at which the frame occurred that was scored by the system (lower right) as well as all results over the 2 min clip.

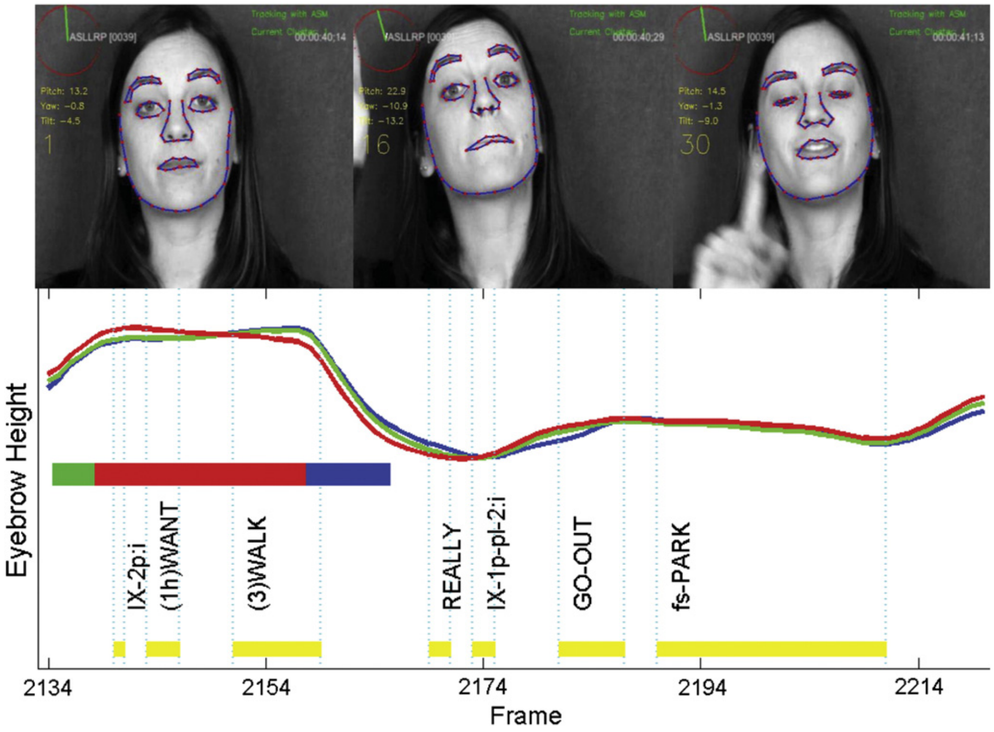

Fig. 3. An example of using an automated system to detect the nonmanual grammatical markings in ASL [4]. In this sentence (meaning 'If you want to walk, the two of us could go out to the park,'), the first clause is marked with a typical non-manual expression for conditional modality, which includes raised eyebrows. Inner, middle, and outer eyebrow heights are shown by the blue, green, and red curves; the green, red, and blue bars identify the temporal phases of the eyebrow movement: onset, apex, and offset, respectively. The yellow bars identify the durations of the manual signs, as glossed.

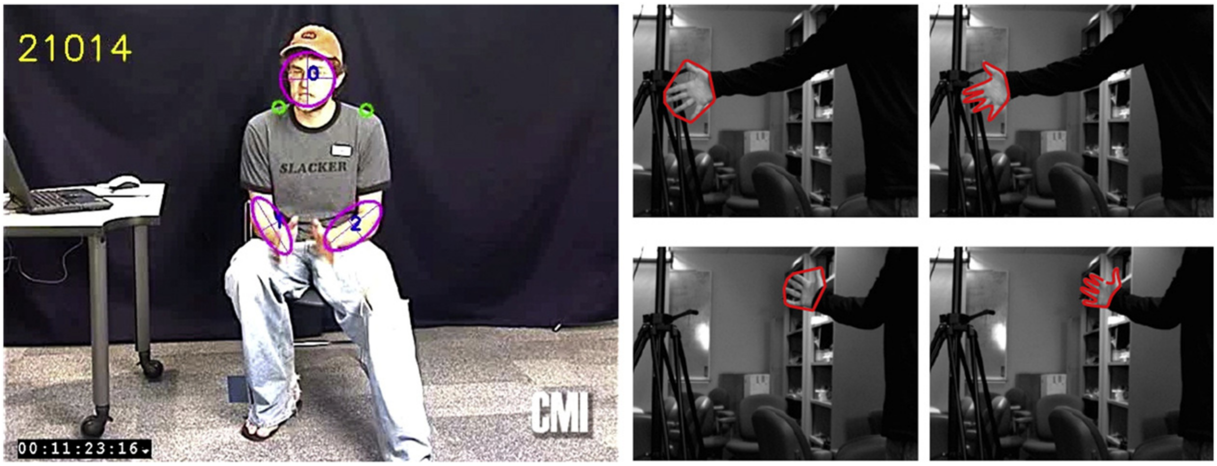

Fig. 4. Left: Demonstration of the skin blob tracking (magenta ellipses) and shrug detection (green circles) [6]. For each frame, the program records blob velocity, blob size and orientation. Right: Hand tracking results by coupling dynamically the discrete and continuous trackers [7].

automatic identification of fingerspelling portions within a fluent stream of signing is non-trivial, as many of the same handshapes that are used as letters are also used in the formation of other types of signs. Fig. 4 shows examples of gesture recognition (Left) and hand tracking results (Right) by coupling dynamically the discrete and continuous trackers [7], using videos of native ASL signers collected and annotated at Boston University as part of the American Sign Language Linguistic Research Project (http://www.bu.edu/asllrp/), in conjunction with the National Center for Sign Language and Gesture Resources.[1] Since this method robustly handles articulations, rotations, abrupt movements and cluttered background, it can accurately discriminate between fingerspelling and non-fingerspelled signs in ASL.

The above examples demonstrate that motion analysis methods such as face tracking are critical to Nonverbal Communication Computing. In this paper, we focus on reviewing the research in the area of human Nonverbal Communication Computing, and especially the motion analysis tools developed to address this problem. We discuss methods to analyze Nonverbal Communication Computing in multiple scales, including face, head, full body and group activities. The remainder of this paper is organized as follows. Section 2 reviews relevant work in human motion analysis and Nonverbal Communication Computing, and also introduces our recent achievements. Section 2.1 covers

[1] The video data and annotations associated with this project are available to the research community from http://www.bu.edu/asllrp/cslgr/.

face tracking methods; Section 2.2 discusses expression recognition; body reconstruction is presented in Section 2.3; and human activity recognition is introduced in Section 2.4. Section 3 summarizes this paper and discusses future directions and open problems.

2. Application domains and developed methodologies

Research in the area of human Nonverbal Communication Computing can be categorized in two main categories: a) highly structured such as American Sign Language (ASL) and, b) less structured which includes application domains such as detection of deception, emotional expressions, stress, and impairments with respect to cognitive and social skills. Both of them rely on robust motion analysis methods such as tracking, reconstruction and recognition. In the following, we will present the motion analysis methods needed for this line of work and several examples to demonstrate the complexity of the problems.

2.1. Face tracking

One of the most important cues for Nonverbal Communication Computing comes from facial motions. Thus accurately tracking head movements and facial actions is very important and has attracted much attention in computer vision and graphics community. Early work typically focused on either rigid head tracking with no facial expression [8,9], or recognizing expressions of a roughly stationary head [10,11]. In contrast, contemporary face tracking systems need to track facial features (e.g. eye corners, nose-tip etc.) under both head motion and varying expressions. A series of face models and tracking algorithms have been developed in recent years. We will introduce face tracking methods based on parametric models, statistical models (Active Shape Models, Constrained Local Models and Active Appearance Models), as well as face tracking from range data.

2.1.1. 3D morphable model-based methods

Parametric face models were first explored to track facial features. Black and Yacoob [12,13] explored the use of local parameterized models and image motion for recovering and recognizing non-rigid and articulated motion of human faces. De Carlo and Metaxas [14,15] described the 3D shape of the face as a polygon mesh, and applied optical flow as a non-holonomic constraint solved by using the least square method. Pighin et al. [16] proposed to use a linear combination of 3D texture-mapped models, each corresponding to a particular basic facial expression. They used a scattered data interpolation technique to deform the face mesh to fit the subject's face from photographs.

The parametric face models have to be carefully designed beforehand, with a set of parameters controlling the deformations driven by elastic forces or image motion. Since the models cannot exactly represent anatomical structures of bones and muscles, unrealistic shapes may be generated. An alternative approach is to learn 3D morphable models from a group of face shapes and textures [17,18], which are usually acquired by high accuracy 3D scans. These 3D face models can represent a wide variety of faces and facial motions. On the other hand, it is computationally expensive and unable to run in real time.

2.1.2. Active shape model-based methods

The Active Shape Models (ASMs) [19] learn statistical distributions of 2D feature points, which allow shapes to vary only in ways seen in a training set. Kanaujia and Metaxas [20] built a real-time face tracking system based on ASM. They trained a mixture of ASMs for pre-aligned faces of different clusters, each corresponding to a different pose, as shown in Fig. 5. The target shape is fitted by first searching the local features along the normal direction, followed by constraining the global shape using the most probable cluster.

2D ASM based methods are also combined with 3D face models, which govern the overall shape, orientation and location. Vogler et al. [21] developed a framework to integrate both 2D ASM and 3D deformable models, which allows robust tracking of faces and estimation of both rigid and non-rigid motions. The displacements between the actual projected model points and the identified correspondences are defined as image forces to update the deformation parameters. Yang et al. [22,23] built a face tracker by combining statistical models of both 2D and 3D faces. Shape fitting was performed by minimizing both feature displacement errors and subspace energy terms with temporal smoothness constraints.

Given the limited number of training samples, traditional statistical shape models may overfit and generalize poorly for new samples. Instead of building models on the entire face, Huang et al. [24] built separate ASM models for face components to preserve local shape deformations. They applied Markov Network to provide global geometry constraints. Some recent research work enhanced the ASM fitting by using sparse displacement errors [25–28]. These models are more robust to outliers and partial occlusions.

2.1.3. Constrained local model-based methods

The constrained local models (CLMs) are extensions of ASM, and they use an independent set of local detectors for landmark detection [29]. CLM fitting is generally posed as the search for the point distribution model (PDM) parameters $\mathbf{p}$, and it jointly minimizes the misalignment error over all landmarks: $Q(\mathbf{p}) = R(\mathbf{p}) + \sum_{i=1}^{n} D_i(x_i)$, where $R(\mathbf{p})$ measures the distance of the current shape from the shape distribution, which is often modeled as Gaussian [30] and Gaussian mixture model (GMM) [31]. $D_i(x_i)$ measures the misalignment of the ith landmark at position x_i. Examples of the misalignment error functions include the Mahalanobis distance for local patch appearances [19], as well as the boosted Harr-like feature based classifiers [29].

As the local landmark detectors are learned from small image regions with limited structures, the maximum responses may not coincide with the correct landmark locations. Some recently proposed methods try to alleviate this problem. Wang et al. [32] proposed a convex quadratic function to fit to the negative log of the response

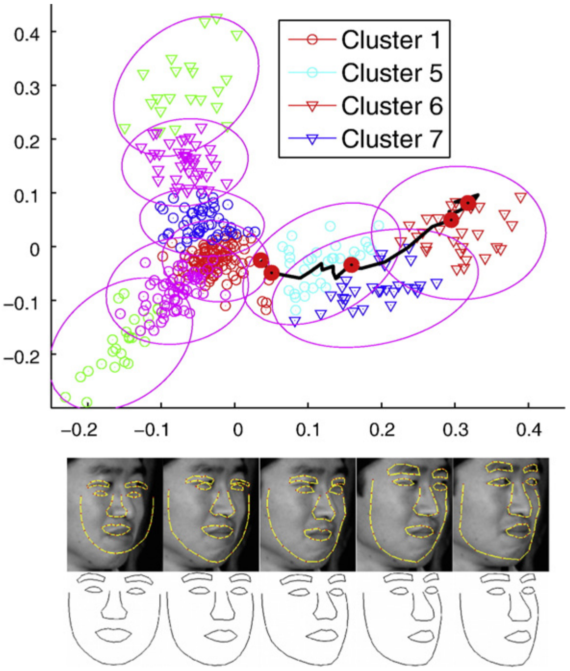

Fig. 5. Top: The face shape manifold is approximated by piecewise linear sub-regions. Bottom: This method [20] searches across multiple clusters to find the best local linear model.

map, from which the mean and covariance of the approximating density can be inferred. Zhou et al. [33] used the summed-squared-difference as a measure of landmark fit, and applied Laplace's approximation to find the covariance estimate. Saragih et al. [34,35] proposed an optimization strategy where a nonparametric representation of the landmark distributions is maximized within a hierarchy of smoothed estimates. The resulting update equations are reminiscent of mean-shift but with a subspace constraint placed on the shape's variability.

2.1.4. Active appearance model-based methods

The ASM and CLM introduced above only model statistical distributions of the shapes. In contrast, AAM decouples the shape and texture of the deformable object, and is able to generate a variety of photo-realistic instances [36]. Fitting an AAM to an image consists of minimizing the error between the input image and the closest model instance, i.e. solving a nonlinear optimization problem. Matthews and Baker [37] suggested to reformulate AAM model fitting as an image alignment problem, which can be efficiently solved by Lucas–Kanade inverse compositional algorithm [38]. The proposed method avoids updating texture parameters and turns out to be the fastest fitting algorithm for AAM.

AAM has been successfully used for real time face tracking. To deal with pose variations, Cootes et al. [39] proposed view based AAM models, which are a combination of a few 2D models. Sung et al. [40] combined AAM with a cylinder head model, where the global head motion parameters obtained from the cylinder model are used as the cues of the AAM parameters for a good fitting or re-initialization. Xiao et al. [41,42] proposed a real time face tracking algorithm by combining 2D + 3D AAM models. Zhou et al. [43] introduced temporal matching constraints to enforce inter-frame coherence in AAM fitting. A comprehensive review is provided by Gao et al. [44].

2.1.5. Face tracking from range data

The face tracking systems using optical cameras suffer from bad lighting conditions, which significantly alter the appearance of feature and cast shadows on faces. In contrast, face tracking systems by using range data are more robust in such conditions. Structured light stereo methods have been applied for capturing depth maps of moving faces [45,46]. Zhang et al. [47] developed a 3D face tracking system by employing synchronized video cameras and structured light projectors to capture streams of images from multiple viewpoints, and the 3D shapes were matched to a template by using both depth error and shape regularization.

Following the recent development of inexpensive depth cameras, there is rapidly growing interest in exploring depth information in vision systems. Fanelli et al. [48] developed a random forest algorithm to estimate head orientations from the range data. Cai et al. [49] developed a maximum likelihood solution to track face shapes from the noisy input depth data. Weise et al. [50] developed a realtime system to reconstruct 3D head shapes from the range data, and used them to generate face animations. Baltrusaitis et al. [51] extended the Constrained Local Model approach to use depth information alongside intensity for facial feature point tracking. Microsoft also published its official Face Tracking SDK that is able to track facial landmark and detect head pose and face expressions in realtime by using a Kinect camera [52].

2.2. Facial expression recognition

Based on the tracked face region, we are able to analyze facial expressions. Facial expression recognition has attracted much attention since as early as 1970s, and it has still been widely investigated in the past decade [53–59], for there remain a lot of opening issues due to the complexity and variety of facial expressions.

The previous works of automatic facial expression recognition can be categorized into two main categories: image based methods [60–62] and video-based methods [63,64,53]. The image based methods take only mug shots as observations which capture characteristic images at the apex of the expressions, and recognize expressions according to appearance features [61,65–67,60,53]. For examples, Gabor features were used in [60] and demonstrated to be more robust in low-resolution facial expression recognition [62]. However, it is computationally expensive to convolve face images with multi-banks of Gabor filters in order to extract multi-scale and orientational coefficients. Haar features were employed in [68] and the Haar + Adaboost method is proved to operate at least two orders of magnitude faster than Gabor + SVM method with a comparable recognition rate. Local Binary Pattern features are used to efficiently represent the facial images [69]. In some cases, it is sufficient to do expression recognition based on the information on a single static image. However, a natural facial expression is dynamic, which evolves over time from the onset, the apex, to the offset. The image based methods ignore such dynamic characteristics, so they cannot perform well in most real world settings. In [53], it states that spontaneous deliberately displayed facial behavior has differences both in utilized facial muscles and their dynamics. Psychological researches have also demonstrated that besides the categories of expression, facial expression dynamics is important to decipher its meaning [70]. Therefore, the video-based methods become much popular in recent years [13,71,53], which aim to analyze the dynamics of facial expression for recognition.

For the video-based methods, how to extract and represent the dynamics of facial expression is a key issue. The typical approaches track facial key points, and analyze their motion and geometric variation of facial appearance [72,73]. These approaches highly depend on the facial key point detection and tracking, which should be invariant to occlusions like glasses and facial hair as long as these do not entirely occlude facial key points. On the other hand, they are easily influenced by illumination. Some researchers assume that the dynamics of facial expression are embedded in a manifold subspace, and such manifold subspace can be learned for facial expression recognition [61,74,75]. However, how to decide the dimension of manifold is still an open problem. In [76], Zhao and Peitikainen proposed Volume Local Binary Pattern (VLBP) and LBP-TOP (LBP from three orthogonal planes) descriptors to capture the dynamics of facial expression, which take the video as a volumetric data in the spatio-temporal domain. The volume feature has the advantage of coupling temporal dynamics with spatial appearance tightly. Similar volume features have also been introduced to action recognition [77], video-based face recognition [78], and pedestrian detection [79]. The performance of volume features suffers from the varying speed at which the facial expressions or actions performed by different people at various situations. One solution is to employ a dynamic time warping preprocessing step, but in literature few work discussed this. The other approach is to make the volume feature detectors robust to such variations. For instance, in [77] the sequences are aligned at the start of the motion but diverge at the end of the sequences, and their work automatically learns to ignore the noisy tail ends of the sequences.

Most of the existing 2D intensity image or video feature-based methods are suitable for the analysis of facial expressions under a small range of head motions. Some attempts have been made to produce pose invariant facial expression classifiers. However, most of these attempts have only considered yaw variations of up to 45°, where the whole face is still visible [80]. They do not consider views greater than 45°, when part of the face is occluded. In order to deal with the inherent pose and illumination variations, 3D and 4D (dynamic 3D) recordings are increasingly used in expression analysis research [81]. Zafeiriou and Yin [82] also present a brief overview on this direction. The first systematic effort to collect 3D facial data for facial expression recognition resulted in the creation of BU-3DFE dataset [83], and they also collected a high-resolution 3D dynamic facial expression database BU-4DFE (3D + time) [84] two years later. The majority of systems developed have attempted the recognition of expressions from static 3D facial expression data [85–89], however, more recent methods employ dynamic 3D facial expression data for this purpose [90,91,88,92]. Most methods in the field of facial expressions in 3D are all based on databases of acted, exaggerated expressions of the six basic emotions, although these are significantly different from natural facial expressions occurring in everyday life.

Furthermore, focus is now shifting towards the recognition of spontaneous facial micro-expressions very recently [56,93–97]. Facial micro-expressions are rapid involuntary facial expressions which reveal suppressed affect. In contrast to the large number of facial expression recognition publications, only a few studies have been done on recognizing micro-expressions. Michael et al. [93] proposed a method for automated deception detection using body movement and extracted motion profiles to capture micro-expressions. In [56], they show how temporal interpolation model together with the first comprehensive spontaneous micro-expression corpus enable them to accurately recognize these very short expressions. Shreve et al. [94] used strain patterns as a feature descriptor for spotting micro-expressions in videos. While more recently Wu et al. [95] used GentleSVM, which is a combination of Gentleboost algorithm and SVM classifier, for spotting and recognizing micro-expressions. However, the biggest obstacle of micro-expression recognition to date has been the lack of a suitable database. In [97], they present a novel Spontaneous Micro-expression Database, which is available online to foster the research in this branch.

We have begun our work on facial expression from synthesizing the 3D facial expressions [98]. In this work, the deformable mesh was used to track the facial motions, and the novel expressions can be synthesized after the facial motion was mapped into low dimensional space. The synthesis work was later extended in [99] for visual interactions. We also perform expression classification in the real data [20], based on our facial tracker. Fig. 6 shows an example of estimating facial expression. The facial motion is estimated by tracking the landmarks on the faces, and the shape information is also integrated into expression analysis. In order to further analyze the facial expression in the video, the encoded dynamic features, which contain both spatial and temporal information, were developed, and boosting method was applied to handle the large dimension problem [100,101]. In order to handle the time resolution problem, the dynamic binary pattern was further proposed [102,103]. Besides the expression classification, the continuous change of expression also plays a key role in lots of applications. Therefore, we further proposed the ranking model to estimate the expression intensity [104]. Comparing with the previous methods, it was the first time that the intensity order was exploited into a learning phase, and this method achieved state-of-the-art performance.

2.3. Full body reconstruction and pose estimation

In addition to face modeling and analysis, whole body motions and gestures are also important factors for Nonverbal Communication Computing. Many applications of Nonverbal Communication Computing, such as the recognition of ASL, need to combine the nonmanual markers (e.g., facial expression) with body movements and gestures to improve the recognition accuracy. Therefore, we have included the discussion of full body reconstruction and 3D pose estimation.

A series of methods have been developed to reconstruct 3D body gesture from monocular video sequences [105–110]. The general framework for 3D pose recovery from monocular sequence has been inspired by the gaining popularity of part-based methods for the problem of 2D human pose alignment in the images. There exists extensive literature on part-based models for the detection and localization of 2D human body parts in images. A few recent works on 2D human pose estimation are [111–117]. Most of these approaches focus on either improving feature extraction to improve part detection confidence or learning efficient priors to model plausible spatial configurations of parts in 2D. Prominent among them is Yang and Ramanan [112] which further enhanced the pictorial structure framework by modeling contextual co-occurrence relations between different part configurations. Poselet based approach [114] uses 3D human pose dataset to improve part detectors and uses Hough transform to vote for the 2D pose configuration. However, all the above approaches focus on estimating 2D pose which is significantly difficult to constrain using standard anthropometric priors.

A much richer literature is in the domain of 3D human pose recovery from monocular images. Several generative [118–121] as well as discriminative [122–125] methods have been proposed for 3D human pose prediction. One major challenge of resolving 3D-pose is that the inverse mapping from observations to (3D pose) states is

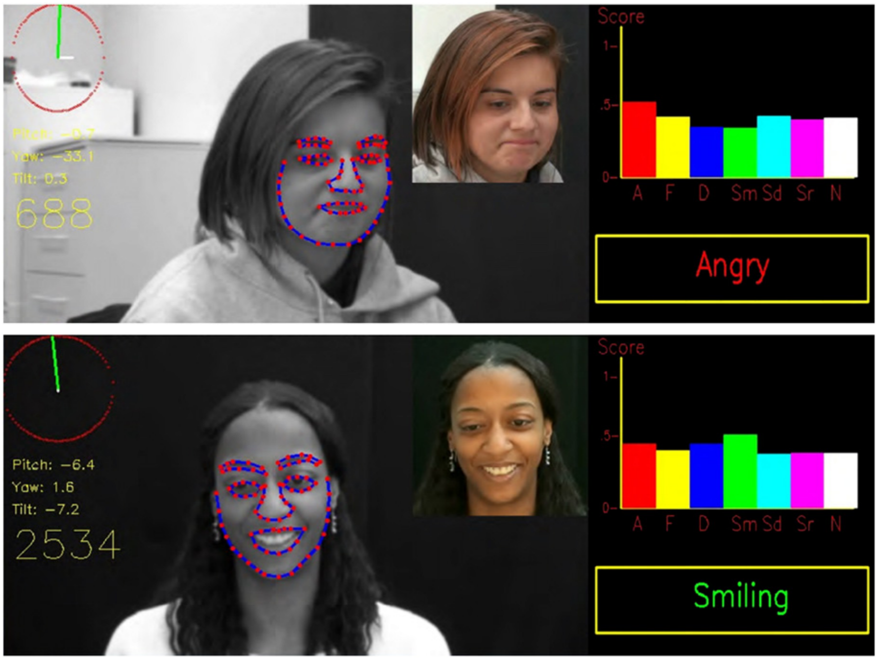

Fig. 6. Sample processed frames showing tracked landmarks, estimated head pose (top left corner) and predicted facial expression scores.

multi-valued and cannot be functionally or globally approximated [123]. Therefore, these methods primarily focus on resolving ambiguities due to multi-valuedness [122,123,125], and are based on coarse, global feature encodings [124,123,125] such as silhouettes that are often noisy and cannot resolve depth ambiguities in images. Lee and Cohen [126] used local likelihood distribution for body parts to locally refine 3D pose. However, their part appearance modeling is based on skin pixel extraction that has limited application to detecting body parts. More recently the work by Serra et al. [127] investigates 3D pose estimation using 2D part-based models. Off-the-shelf articulated 2D pose detector is used to generate initial hypotheses for the 3D pose estimation framework. Recent studies in feature extraction and depth estimation from a single image [128,129] have also shown that image encodings at multiple scale space can be used to estimate depth.

We have developed a discriminative approach — Bayesian Mixture of Experts (BME) [123]. BME models multivalued image-to-pose relations using several experts. Predictions from these experts are combined in a probabilistic Gaussian mixture, with centers at predicted values. However, the predicted 3D pose is sensitive to image ambiguities and lacks true geometric or kinematic constraints. Therefore, we can use the BME model as an efficient (but approximate) prior to a search space in the greedy optimization framework that we have developed. Fig. 7 shows examples of BME predictions. The new framework involves using both global shape cues and local alignment features to search for the optimal 3D pose that best matches a given 2D observation, as shown in Fig. 8. Our 3D search framework introduces a new method of 2D to 3D pose reconstruction that is flexible and can be applied to cluttered real-world images. We combine discriminative and generative approaches that include both local/part-based fitting, as well as global shape, into a single framework. The discriminative initialization can provide efficient robust approximation(s) while the generative 3D model will enable us to resolve ambiguities due to depth and occlusion.

2.4. Activity recognition

Besides analyzing Nonverbal Communication Computing at the level of individual persons, many researchers have also investigated the group activities employing motion analysis and/or machine learning methods [131–139]. Modeling group activities plays an important role in video surveillance and smart camera systems, and there are many promising applications. For examples, automated recognition and classification of videos enable more efficient video searching, e.g. finding tackles in soccer matches, handshakes in news footage or typical dance moves in music videos. It is also important for automatic surveillance, e.g. monitoring shopping malls. Another example is to support aging in places for the elderly in smart homes. Interaction applications like human–computer interactions also benefit from the advances in automatic human action recognition. Various abnormal activities have been studied, including restricted-area access detection [140], car counting [141], detection of people carrying cases [142], abandoned objects [143], group activity detection [144,145], social network modeling [146], monitoring vehicles [147], scene analysis [148] and so on. Fig. 9 shows two sample frames from the BEHAVE dataset [130].

In recent years, a lot of algorithms have been proposed to improve the performance of action/activity analysis. Many of them focus on finding better image representation and features extracted from the image sequences. Ideally, these should generalize over small variations in person appearance, background, viewpoint and action types. At the same time, the representations must be sufficiently rich for robust action classification. Using local descriptors or patches is a popular way to represent human actions. A video sequence is then represented by a collection of independent patches. Accurate localization and background subtraction are not required. The local representations are invariant to changes in viewpoint, person appearance and partial occlusions. Space–time interest points are the locations in space and time where sudden changes of movement occur in the video. Laptev and Lindeberg [149] extended the Harris corner detector [150] to 3D. Space–time interest points are those points where the local neighborhood has a significant variation in both the spatial and the temporal domains. Dollár et al. [151] used dense sampling instead of sparse interest points for feature representation. This method applies Gabor filtering on the spatial and temporal dimensions individually. In addition to intensity and motion cues, Rapantzikos et al. [152] also incorporated color information.

After local interest point detection, local descriptors are applied to summarize an image/video patch. The spatial and temporal size of a patch is usually determined by the scale of the interest point. Schuldt et al. [153] calculated patches of normalized derivatives in space and time. Niebles et al. [154] took the same approach but apply smoothing before reducing the dimensionality using PCA. Dollar et al. [151] tested with both image gradients and optical flow (refer to Mikolajczyk et al. [155] for a detailed survey on features). How to model the relationship among local features is also very important. One solution is to build grids over spatial/temporal domain. Ikizler and Duygulu [156] sampled oriented rectangular patches and bin them into a grid. Zhao and Elgammal [157] used local descriptors around interest points in a histogram with different levels of granularity. Nowozin et al. [158] used a temporal instead of a spatial grid. Another way is to exploit correlations between local descriptors to construct higher-level descriptors. Scovanner et al. [159] constructed a word co-occurrence matrix for a reduced codebook size. Liu et al. [160] used a combination of the space–time features and spin images to represent the correlations of features.

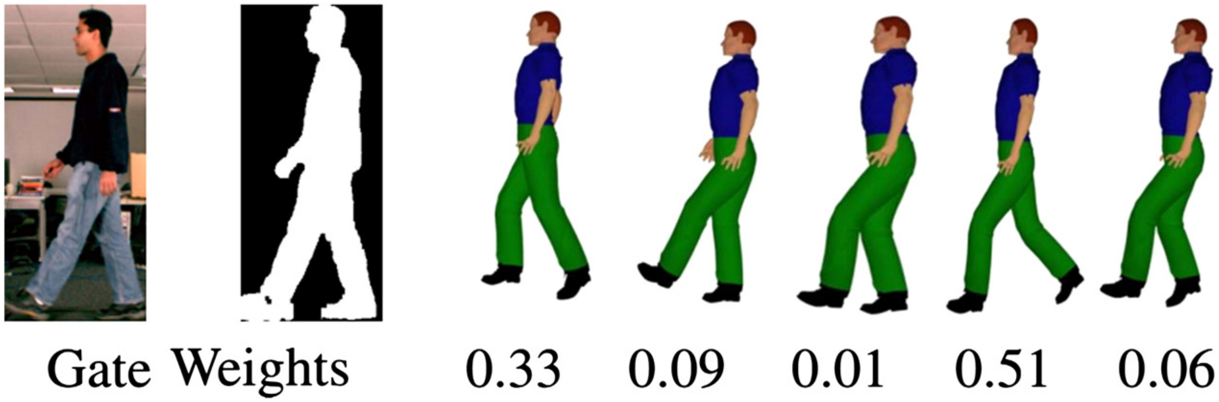

Fig. 7. Bayesian Mixture of Experts (BME) predictions from the top 5 experts [123].

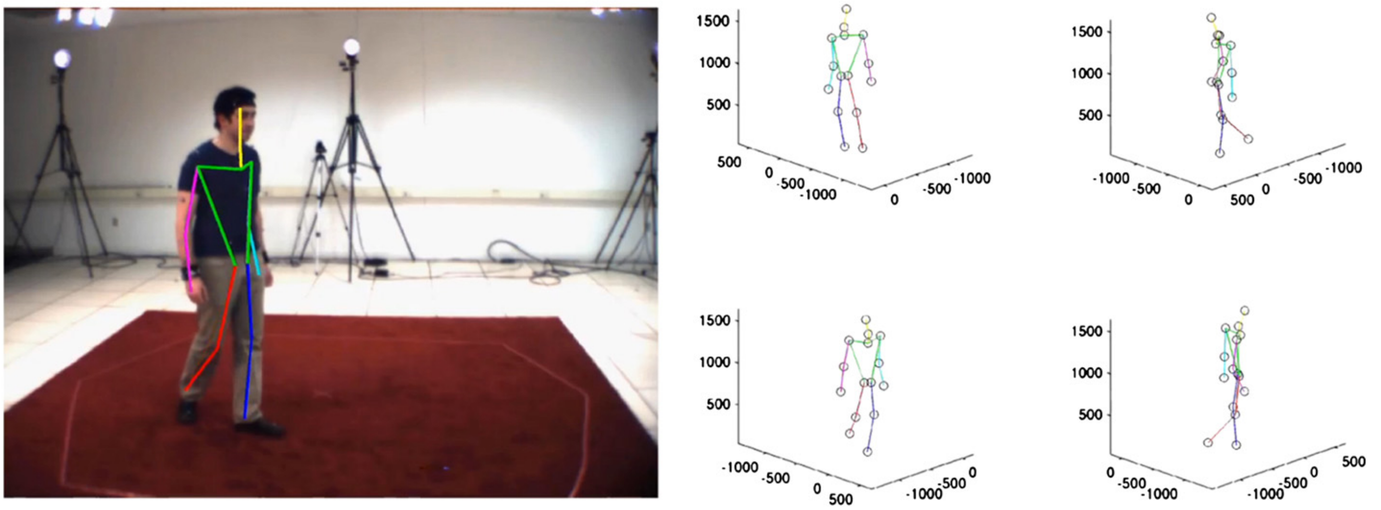

Fig. 8. An example of search for the optimal 3D pose that best matches a given 2D observation.

These algorithms have been successfully applied to action recognition problems. They focus on single action with one person (*hand-waving*, *running*… [153]) or pair-wise action recognition (*answer phone* [161], *horse riding* [162]). However, they do not consider interactions among multiple people. For most of the surveillance systems in public area, it is also important to identify group activities. Events like fighting or escaping often involve multiple people and their interactions. Several algorithms for group activity modeling have been proposed in recent years. Different features are used for group activity: human body/body parts [163,164], optical flow [165] and detecting moving regions [166]. Recently, Zhou et al. [144] and Ni et al. [145] used trajectory analysis to describe different group activities.

Modeling social behaviors of people is an important branch to represent group activity, and it has been widely used in evacuation dynamics, traffic analysis and graphics. Pedestrian behaviors have been studied from a crowd perspective, with macroscopic models for crowd density and velocity. On the other end, microscopic models deal with individual pedestrians. A popular model is the Social Force Model [167]. In the Social Force Model, pedestrians react to energy potentials caused by other pedestrians and static obstacles through a repulsive force, while trying to keep a desired speed and motion direction. Helbing and Molnar [167] originally introduced this concept to investigate people movement dynamics. It is also applied to the simulation of crowd behavior [168], virtual reality and studies in computer graphics for creating realistic animations of the crowd [169].

Social behavior analysis has also attracted much attention in the computer vision community. Ali and Shah [170] used the cellular automaton model to track in extremely crowded situations. Antonini et al. [171] proposed a variant of Discrete Choice Model to build a probability distribution over pedestrian positions in next time step. Scovanner and Tappen [172] modeled pedestrians' dynamics and motions as a continuous optimization problem. Pellegrini et al. [173] proposed a Linear Trajectory Avoidance (LTA) method to track multiple targets. Predictions of velocities are computed by the minimization of energy potentials. Recently, Mehran et al. [174] proposed a method to model behaviors among a group of people. It represents the abnormal patterns in a local region based on moving particles. Wu et al. [175] used chaotic invariants of Lagrangian Particle Trajectories to model abnormal patterns in crowded scenes. They have been successfully used in crowded scene modeling. We have also proposed a method named as Interaction Energy Potential to model such interactions [176]. It is based on the relationship between the current state of a person and his/her reactions. Specifically, the relationship between the current state of a subject and the corresponding reaction is explored to model the normal/abnormal patterns. The framework will learn and recognize abnormal events in different environmental contexts.

Although a significant amount of progress on activity recognition has been achieved, there are still many open problems. First, accurate segmentation and tracking are still a challenging task, which are

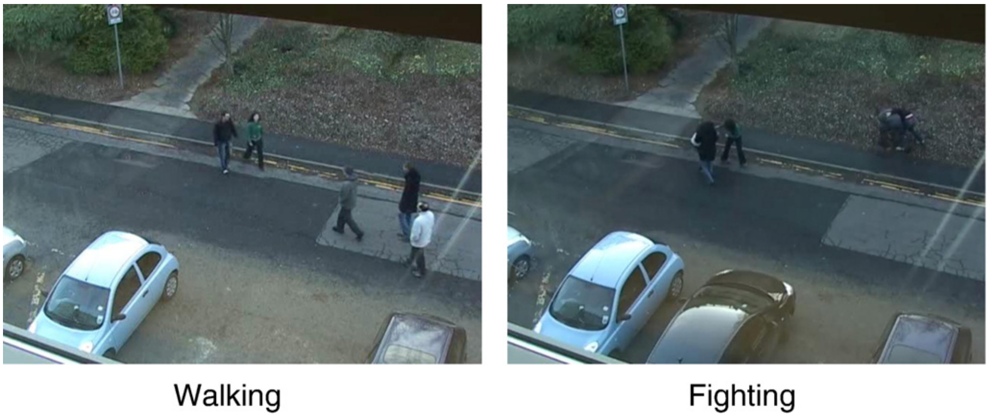

Fig. 9. Two samples from the BEHAVE dataset [130].

caused by the poor lighting, crowded environments, noisy images, and camera movements. Therefore, developing robust segmentation and tracking methods is always important. Furthermore, most public databases are still based on experimental settings. There is still a big gap between the research and practice. Thus it is necessary to validate current approaches on real-world applications. New applications are also encouraged, such as monitoring doctors and patients in the hospital environment or other health-care facilities.

3. Future work and open problems

Research in Nonverbal Communication Computing and motion analysis is maturing, and there are many exciting methods and applications that need to be addressed. Most current systems suffer from the lack of accuracy and robustness, which is a major obstacle when dealing with large data and complex motions. Analytic methods should employ robust tracking and statistical learning methods to identify the important motion parameters that describe behavior. In this section, we discuss the future work and open problems in two main directions: 1) robust motion analysis methods using 3D deformable models, and 2) fusion of domain knowledge and multiple cues.

3.1. 3D deformable models for motion analysis

Most of the above-mentioned motion analysis methods are based on 2D models. Traditional 2D methods are not able to handle large off plane pose changes (e.g., rotations or head tilts) and occlusions. The reason is that objects they track have a 3D shape and a 2D solution tracks the projection on the plane only. Therefore, when the rotation is large (e.g., head shaking), the 2D shape of the face may be degenerated to a thin region.

To deal with such problems, one can train multiple 2D models at different rotations, and switch among multiple models during motion analysis. However, it significantly increases the computational complexity, and the rotation space is actually infinite (it is a continuous variable). Therefore the solution is not as accurate as that from a continuous 3D model. A 2D tracking solution has significant problems with large occlusions since it only tracks 2D projections; for example a hand in front of a face is ambiguous in terms of how far it is from the face in 3D, touching or not. Using 3D models (e.g., a 3D face mask) [178,5,179,178], we can parametrically represent 3D rotations and relative depth, which we can estimate during tracking. Therefore a 3D model can deal seamlessly with occlusions and non-planar movement, as opposed to 2D approaches. In addition, in the case of deformations, it can also deal with 3D deformations that a 2D solution cannot. Fig. 10 shows an example of using a 3D deformable model. It is able to handle occlusion and large rotations. Fig. 11 shows an example of tracking hand in 3D. It is able to robustly track a sequence of hand rotation and finger movements (such accurate hand tracking results can be employed for deception detection [6]).

In addition, traditional 2D approaches provide the shape (in the form of 2D contours) of the face, eyes, eyebrows, nose, and other subparts, which are often used for recognition. They work well when people face the camera. However, they are challenged with non-frontal facial poses and often are not robust to head shaking, head tilting and large rotations. By contrast, with a 3D model based tracker, one can track inherently 3D parameters which are continuous and not discrete, and can also obtain improved recognition results for both the head pose and related facial deformations. An additional significant benefit of a 3D approach is that it can naturally normalize the tracking pose and facial estimation parameters, which is a requirement for the recognition of pose and expression. In a 2D approach the normalization process is as accurate as in a 3D approach since the perspective distortions are nonlinear [180,181].

Although 3D approaches can achieve promising performance, its main challenge is the computational efficiency. 3D approaches usually have significantly more degrees of freedom and thus estimating 3D parameters is more complex. Therefore, developing real-time 3D solutions is a very important research topic.

3.2. Fusion of domain knowledge and multiple cues

To further improve the performance of Nonverbal Communication Computing, researchers have proposed solutions such as the incorporation of nonverbal coding systems and domain knowledge in motion analysis and behavior interpretation. They include kinesics, proxemics, and linguistic knowledge (e.g., in ASL). Possible future contexts will range from highly structured (e.g., interviews and ASL) to little structured (e.g., casual conversations), from face-to-face to mediated contexts, social interactions and social network-based interactions, and will include both normal and impaired communication. These extensions will allow the research of Nonverbal Communication Computing to evolve beyond initial foundational science and proof of concept, to include and solve problems in applied contexts. For example, from linguistic knowledge, we know that changes in eyebrow configuration, in combination with head gestures and other facial expressions, are used to signal essential grammatical information in signed languages. Therefore, we propose methods to recognize the components of eyebrow and periodic head gestures, and successfully improve the detection of non-manual grammatical markings in ASL [4]. The main challenge of these approaches is how to effectively incorporate domain knowledge into traditional models, and how to efficiently solve them. The composite prior models are promising solutions because of their flexibility in modeling prior knowledge and their computational efficiency [182,183].

Finally, some of the clusterings of nonverbal behaviors observed to correlate with specific constructions can be decomposed into components with their own semantic contributions, physical realizations, and linguistic distributions. Therefore, combining multiple nonverbal behaviors can potentially advance the performance of Nonverbal Communication Computing. In addition, combining nonverbal behavior analytics with other behavioral cues will allow eventually the comprehensive study of human behavior and human–computer interaction. A major challenge of these approaches is to effectively fuse these cues, or features. Since these features may have redundant information, sparse methods, especially group sparsity, are potential solutions to fuse them [184–186]. These fusion methods can automatically discover a sparse subset of multiple features and improve the recognition accuracy and efficiency.

3.3. Summary

To summarize, although a large amount of work has been done in the area of Nonverbal Communication Computing, there are still many open problems and new promising applications to explore. Other interesting topics and open problems for future research include large-scale data analysis, physics-based modeling, and robust learning. For example, most datasets used for ASL recognition [187,188] are not large-scale. However, to be able to process massive data in real world application, it is necessary to provide ready access to large-scale, high-quality multimodal corpora for several signed and spoken languages, with linguistic annotation, fine-grained computational analysis, and tools for data visualization. This has been constrained by difficulties inherent in collecting, annotating, and analyzing large quantities of video data. Therefore, new protocols should be developed for collection, analysis, storage, and dissemination of high-quality audio/video language corpora larger in scale and more diverse in content than ASL datasets now available. New promising applications are also encouraged. For example, researchers have started to analyze neurological diseases, such as Parkinson's disease and schizophrenia, by coupling motion analysis methods and brain activity. The computational methods developed in Nonverbal

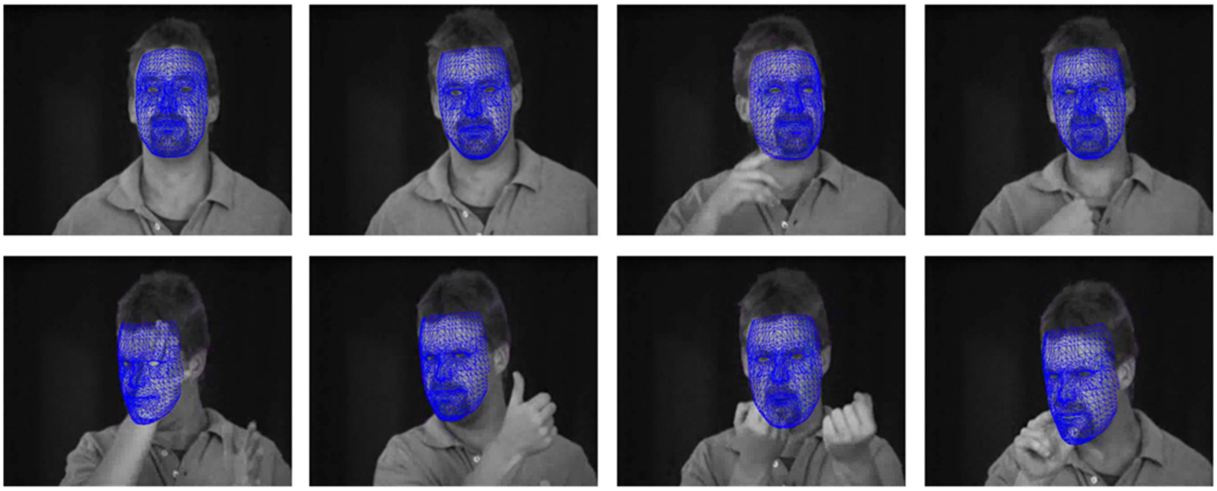

Fig. 10. Tracking results using a 3D deformable model. It is robust to occlusion and large rotations.

Communication Computing can be combined with brain activity analysis to transform the screening and treatments of these neurological spectral disorders. Such applications will have significant scientific and societal impact.

Acknowledgments

The authors would like to thank all the reviewers for their constructive suggestions. We also would like to thank our long time collaborators Judee Burgoon (UA), David Dinges (UPENN), and Carol Neidle (BU). Metaxas would like to thank his previous PhD students Ioannis Kakadiaris (Univ of Houston), Atul Kanaujia (Object Video), Siome Goldenstein (UNICAMP, Brazil), Dimitris Samaras (SUNY Stony Brook), Christian Vogler (Gallaudet) and Gabriel Tsechpenakis (IUPUI) for their seminal work in this area. Finally, we would like to thank our current CBIM students and colleagues Fei Yang, Xinyi Cui, Mark Dilsizian, Christophe Restif and Qiong Hu for their help and valuable suggestions. This work is supported in part by NSF (award numbers 0964597, 1064965, 1059281), NASA-NSBRI-NBTS01601, NASA-NSBRI-NBTS004, and ONR-N000140910104.

References

[1] L. Wang, W. Hu, T. Tan, Recent developments in human motion analysis, Pattern Recognit. 36 (3) (2003) 585–601.

[2] X. Yu, S. Zhang, Z. Yan, F. Yang, J. Huang, N. Dunbar, M. Jensen, J. Burgoon, D. Metaxas, Is interactional dissynchrony a clue to deception: Insights from automated analysis of nonverbal visual cues, The Hawaii International Conference on System Sciences (HICSS), 2013.

[3] N. Michael, F. Yang, D. Metaxas, D. Dinges, Development of optical computer recognition (ocr) for monitoring stress and emotions in space, 18th IAA Humans in Space Symposium, 2011.

[4] J. Liu, B. Liu, S. Zhang, F. Yang, P. Yang, D. Metaxas, C. Neidle, Recognizing eyebrow and periodic head gestures using CRFs for non-manual grammatical marker detection in ASL, International Conference on Automatic Face and Gesture Recognition, 2013.

[5] F. Yang, J. Wang, E. Shechtman, L. Bourdev, D. Metaxas, Expression flow for 3D-aware face component transfer, ACM Trans. Graph. (Proc. SIGGRAPH) 27 (3) (2011) 60.

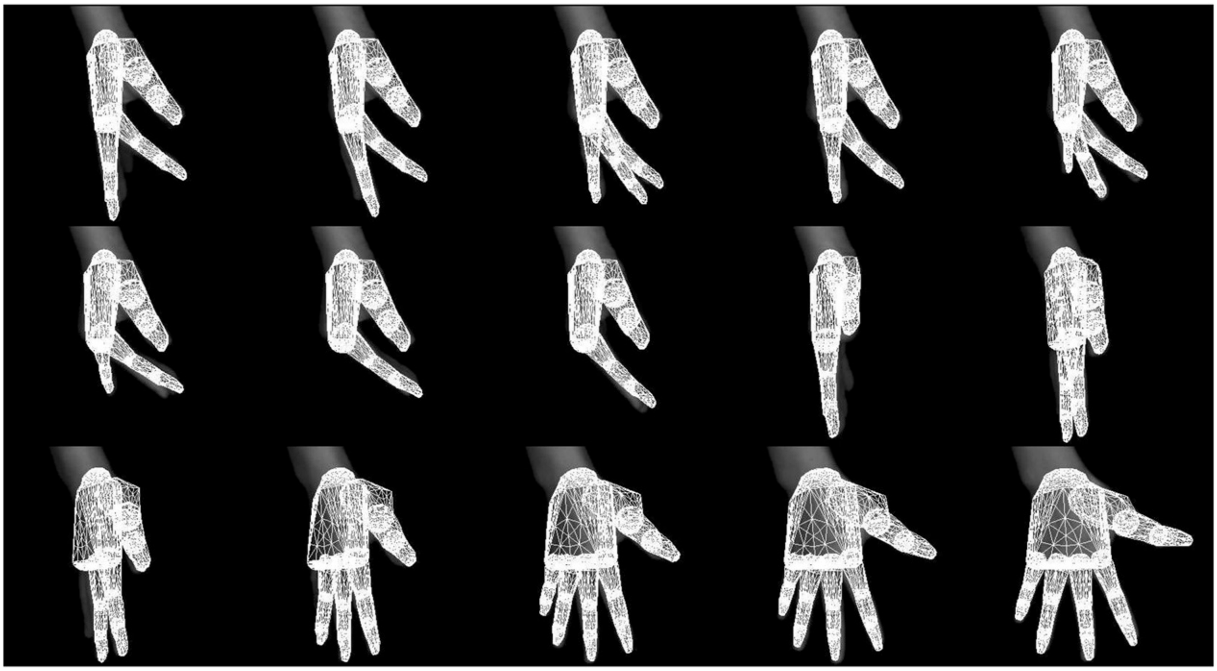

Fig. 11. 3D hand tracking results using a deformable model framework [177]. It is able to track a sequence of hand rotation and finger movements.

[6] S. Lu, G. Tsechpenakis, D.N. Metaxas, M.L. Jensen, J. Kruse, Blob analysis of the head and hands: a method for deception detection, The 38th Annual Hawaii International Conference on System Sciences, IEEE, 2005.

[7] G. Tsechpenakis, D. Metaxas, C. Neidle, Learning-based dynamic coupling of discrete and continuous trackers, Comput. Vis. Image Underst. 104 (2) (2006) 140–156.

[8] A. Azarbayejani, B. Horowitz, A. Pentland, Recursive estimation of structure and motion using relative orientation constraints, Computer Vision and Pattern Recognition, 1993. Proceedings CVPR'93., 1993 IEEE Computer Society Conference on, IEEE, 1993, pp. 294–299.

[9] A. Azarbayejani, T. Starner, B. Horowitz, A. Pentland, Visually controlled graphics, IEEE Trans. Pattern Anal. Mach. Intell. 15 (6) (1993) 602–605.

[10] A. Yuille, P. Hallinan, Active vision, Ch. Deformable Templates, MIT Press, 1993. 21–38.

[11] D. Terzopoulos, K. Waters, Analysis and synthesis of facial image sequences using physical and anatomical models, IEEE Trans. Pattern Anal. Mach. Intell. 15 (6) (1993) 569–579.

[12] M.J. Black, Y. Yacoob, Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion, Computer Vision, 1995. Proceedings., Fifth International Conference on, IEEE, 1995, pp. 374–381.

[13] M.J. Black, Y. Yacoob, Recognizing facial expressions in image sequences using local parameterized models of image motion, Int. J. Comput. Vis. 25 (1) (1997) 23–48.

[14] D. DeCarlo, D. Metaxas, The integration of optical flow and deformable models with applications to human face shape and motion estimation, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 1996, pp. 231–238.

[15] D. Decarlo, D. Metaxas, Optical flow constraints on deformable models with applications to face tracking, Int. J. Comput. Vis. 38 (2) (2000) 99–127.

[16] F. Pighin, R. Szeliski, D.H. Salesin, Resynthesizing facial animation through 3d model-based tracking, ICCV 1999, vol. 1, IEEE, 1999, pp. 143–150.

[17] V. Blanz, T. Vetter, A morphable model for the synthesis of 3d faces, Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, ACM Press/Addison-Wesley Publishing Co, 1999, pp. 187–194.

[18] V. Blanz, C. Basso, T. Poggio, T. Vetter, Reanimating faces in images and video, Computer Graphics Forum, vol. 22, Wiley Online Library, 2003, pp. 641–650.

[19] T. Cootes, C. Taylor, D. Cooper, J. Graham, et al., Active shape models-their training and application, Comput. Vis. Image Underst. 61 (1) (1995) 38–59.

[20] A. Kanaujia, D. Metaxas, Recognizing facial expressions by tracking feature shapes, 18th International Conference on Pattern Recognition, vol. 2, IEEE, 2006, pp. 33–38.

[21] C. Vogler, Z. Li, A. Kanaujia, S. Goldenstein, D. Metaxas, The best of both worlds: combining 3D deformable models with active shape models, IEEE 11th International Conference on Computer Vision, IEEE, 2007, pp. 1–7.

[22] F. Yang, J. Wang, E. Shechtman, L. Bourdev, D. Metaxas, Expression flow for 3D-aware face component transfer, ACM Trans. Graph. (Proc. SIGGRAPH) 27 (3) (2011) 60.

[23] F. Yang, L. Bourdev, E. Shechtman, J. Wang, D. Metaxas, Facial expression editing in video using a temporally-smooth factorization, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 861–868.

[24] Y. Huang, Q. Liu, D. Metaxas, A component based deformable model for generalized face alignment, IEEE 11th International Conference on Computer Vision, IEEE, 2007, pp. 1–8.

[25] F. Yang, J. Huang, D. Metaxas, Sparse shape registration for occluded facial feature localization, 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, IEEE, 2011, pp. 272–277.

[26] S. Zhang, Y. Zhan, M. Dewan, J. Huang, D. Metaxas, X. Zhou, Sparse shape composition: a new framework for shape prior modeling, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 1025–1032.

[27] S. Zhang, Y. Zhan, M. Dewan, J. Huang, D. Metaxas, X. Zhou, Towards robust and effective shape modeling: sparse shape composition, Med. Image Anal. (2012) 265–277.

[28] S. Zhang, Y. Zhan, D.N. Metaxas, Deformable segmentation via sparse representation and dictionary learning, Med. Image Anal. 16 (7) (2012) 1385–1396.

[29] D. Cristinacce, T.F. Cootes, Feature detection and tracking with constrained local models, BMVC, 2006, 2006, pp. 929–938.

[30] C. Basso, T. Vetter, V. Blanz, Regularized 3D morphable models, Higher-Level Knowledge in 3D Modeling and Motion Analysis, 2003. HLK 2003. First IEEE International Workshop on, IEEE, 2003, pp. 3–10.

[31] L. Gu, T. Kanade, A generative shape regularization model for robust face alignment, ECCV (1), 2008, pp. 413–426.

[32] Y. Wang, S. Lucey, J.F. Cohn, Enforcing convexity for improved alignment with constrained local models, CVPR, 2008, 2008.

[33] X.S. Zhou, D. Comaniciu, A. Gupta, An information fusion framework for robust shape tracking, IEEE Trans. Pattern Anal. Mach. Intell. 27 (1) (2005) 115–129.

[34] J.M. Saragih, S. Lucey, J.F. Cohn, Face alignment through subspace constrained mean-shifts, ICCV, 2009, 2009, pp. 1034–1041.

[35] J.M. Saragih, S. Lucey, J.F. Cohn, Deformable model fitting by regularized landmark mean-shift, Int. J. Comput. Vis. 91 (2) (2011) 200–215.

[36] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.

[37] I. Matthews, S. Baker, Active appearance models revisited, Int. J. Comput. Vis. 60 (2) (2004) 135–164.

[38] S. Baker, I. Matthews, Lucas–Kanade 20 years on: a unifying framework, Int. J. Comput. Vis. 56 (3) (2004) 221–255.

[39] T.F. Cootes, G. Wheeler, K. Walker, C. Taylor, View-based active appearance models, Image Vis. Comput. 20 (9) (2002) 657–664.

[40] J. Sung, T. Kanade, D. Kim, Pose robust face tracking by combining active appearance models and cylinder head models, Int. J. Comput. Vis. 80 (2) (2008) 260–274.

[41] J. Xiao, S. Baker, I. Matthews, T. Kanade, Real-time combined 2D + 3D active appearance models, CVPR, 2, 2004, pp. 535–542.

[42] I. Matthews, J. Xiao, S. Baker, 2D vs. 3D deformable face models: representational power, construction, and real-time fitting, Int. J. Comput. Vis. 75 (1) (2007) 93–113.

[43] M. Zhou, L. Liang, J. Sun, Y. Wang, AAM based face tracking with temporal matching and face segmentation, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 701–708.

[44] X. Gao, Y. Su, X. Li, D. Tao, A review of active appearance models, IEEE Trans. Syst. Man Cybern. C Appl. Rev. 40 (2) (2010) 145–158.

[45] M. Proesmans, L. Van Gool, A. Oosterlinck, One-shot active 3D shape acquisition, Pattern Recognition, 1996, Proceedings of the 13th International Conference on, vol. 3, IEEE, 1996, pp. 336–340.

[46] P.S. Huang, C. Zhang, F.-P. Chiang, High-speed 3-D shape measurement based on digital fringe projection, Opt. Eng. 42 (1) (2003) 163–168.

[47] L. Zhang, N. Snavely, B. Curless, S.M. Seitz, Spacetime faces: high-resolution capture for modeling and animation, ACM Annual Conference on Computer Graphics, 2004, pp. 548–558.

[48] G. Fanelli, T. Weise, J. Gall, L.J.V. Gool, Real time head pose estimation from consumer depth cameras, Proc. DAGM-Symposium, 2011, pp. 101–110.

[49] Q. Cai, D. Gallup, C. Zhang, Z. Zhang, 3D deformable face tracking with a commodity depth camera, Proc. ECCV, 2010, pp. 229–242.

[50] T. Weise, S. Bouaziz, H. Li, M. Pauly, Realtime performance-based facial animation, ACM Trans. Graph. 30 (4) (2011) 77.

[51] T. Baltrusaitis, P. Robinson, L. Morency, 3D constrained local model for rigid and non-rigid facial tracking, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2610–2617.

[52] Z. Zhang, Microsoft kinect sensor and its effect, IEEE Multimedia 19 (2) (2012) 4–10.

[53] Z. Zeng, M. Pantic, G. Roisman, T. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58.

[54] P. Yang, Q. Liu, D. Metaxas, Dynamic soft encoded patterns for facial event analysis, Comput. Vis. Image Underst. 115 (3) (2011) 456–465.

[55] X. Huang, G. Zhao, W. Zheng, M. Pietikainen, Spatiotemporal local monogenic binary patterns for facial expression recognition, IEEE Signal Process. Lett. 19 (5) (2012) 243–246.

[56] T. Pfister, X. Li, G. Zhao, M. Pietikainen, Recognising spontaneous facial micro-expressions, IEEE International Conference on Computer Vision, IEEE, 2011, pp. 1449–1456.

[57] G. Zhao, X. Huang, M. Taini, S. Li, M. Pietikäinen, Facial expression recognition from near-infrared videos, Image Vis. Comput.

[58] O. Rudovic, I. Patras, M. Pantic, Coupled Gaussian process regression for pose-invariant facial expression recognition, European Conference on Computer Vision, 2010, pp. 350–363.

[59] G. Zhao, M. Pietikäinen, Boosted multi-resolution spatiotemporal descriptors for facial expression recognition, Pattern Recognit. Lett. 30 (12) (2009) 1117–1127.

[60] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, J. Movellan, Dynamics of facial expression extracted automatically from video, Image Vis. Comput. 24 (6) (2006) 615–625.

[61] C. Shan, S. Gong, P. McOwan, Conditional mutual information based boosting for facial expression recognition, British Machine Vision Conference, vol. 1, 2005.

[62] Y. Tian, Evaluation of face resolution for expression analysis, Conference on Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04, IEEE, 2004, pp. 82–88.

[63] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognit. 36 (1) (2003) 259–275.

[64] M. Pantic, L. Rothkrantz, Automatic analysis of facial expressions: the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1424–1445.

[65] M. Bartlett, G. Littlewort, I. Fasel, J. Movellan, Real time face detection and facial expression recognition: development and applications to human computer interaction, Conference on Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03, vol. 5, IEEE, 2003, pp. 53–60.

[66] M. Pantic, L. Rothkrantz, Facial action recognition for facial expression analysis from static face images, IEEE Trans. Syst. Man Cybern. B Cybern. 34 (3) (2004) 1449–1461.

[67] C. Shan, S. Gong, P. McOwan, Robust facial expression recognition using local binary patterns, IEEE International Conference on Image Processing (ICIP), vol. 2, IEEE, 2005, (II–370).

[68] J. Whitehill, C. Omlin, Haar features for FACS AU recognition, 7th International Conference on Automatic Face and Gesture Recognition, 2006. FGR 2006, IEEE, 2006, (5 pp.).

[69] X. Feng, M. Pietikainen, A. Hadid, Facial expression recognition with local binary patterns and linear programming, Pattern Recognition and Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii, 15 (2), 2005, p. 546.

[70] J. Bassili, Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face, J. Pers. Soc. Psychol. 37 (11) (1979) 2049.

[71] Y. Yacoob, L. Davis, Computing spatio-temporal representations of human faces, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 1994, pp. 70–75.

[72] H. Gu, Q. Ji, Facial event classification with task oriented dynamic bayesian network, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE, 2004, (II–870).

[73] M. Valstar, I. Patras, M. Pantic, Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data, IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops, IEEE, 2005, pp. 76–83.

[74] Y. Chang, C. Hu, M. Turk, Manifold of facial expression, IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003, pp. 28–35.

[75] C. Lee, A. Elgammal, Facial expression analysis using nonlinear decomposable generative models, Analysis and Modelling of Faces and Gestures, 2005, pp. 17–31.

[76] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.

[77] Y. Ke, R. Sukthankar, M. Hebert, Efficient visual event detection using volumetric features, IEEE International Conference on Computer Vision, vol. 1, IEEE, 2005, pp. 166–173.

[78] A. Hadid, M. Pietikäinen, S. Li, Learning personal specific facial dynamics for face recognition from videos, Analysis and Modeling of Faces and Gestures, 2007, pp. 1–15.

[79] X. Cui, Y. Liu, S. Shan, X. Chen, W. Gao, 3D haar-like features for pedestrian detection, IEEE International Conference on Multimedia and Expo, IEEE, 2007, pp. 1263–1266.

[80] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, Y. Sato, Pose-invariant facial expression recognition using variable-intensity templates, Int. J. Comput. Vis. 83 (2) (2009) 178–194.

[81] G. Sandbach, S. Zafeiriou, M. Pantic, L. Yin, Static and dynamic 3d facial expression recognition: A comprehensive survey, Image Vis. Comput.

[82] S. Zafeiriou, L. Yin, 3D facial behaviour analysis and understanding, Image Vis. Comput. 30 (10) (2012) 681–682.

[83] L. Yin, X. Wei, Y. Sun, J. Wang, M. Rosato, A 3d facial expression database for facial behavior research, International Conference on Automatic Face and Gesture Recognition, IEEE, 2006, pp. 211–216.

[84] L. Yin, X. Chen, Y. Sun, T. Worm, M. Reale, A high-resolution 3D dynamic facial expression database, International Conference on Automatic Face and Gesture Recognition, IEEE, 2008, pp. 1–6.

[85] H. Soyel, H. Demirel, Facial expression recognition using 3D facial feature distances, Image Anal. Recognit. (2007) 831–838.

[86] H. Tang, T. Huang, 3D facial expression recognition based on properties of line segments connecting facial feature points, International Conference on Automatic Face and Gesture Recognition, IEEE, 2008, pp. 1–6.

[87] X. Li, Q. Ruan, Y. Ming, 3D facial expression recognition based on basic geometric features, International Conference on Signal Processing (ICSP), IEEE, 2010, pp. 1366–1369.

[88] A. Maalej, B. Amor, M. Daoudi, A. Srivastava, S. Berretti, et al., Local 3D shape analysis for facial expression recognition, 20th International Conference on Pattern Recognition, 2010, pp. 4129–4132.

[89] A. Savran, B. Sankur, M. Taha Bilge, Comparative evaluation of 3D vs. 2D modality for automatic detection of facial action units, Pattern Recognit. 45 (2) (2012) 767–782.

[90] Y. Sun, L. Yin, Facial expression recognition based on 3D dynamic range model sequences, European Conference on Computer Vision, 2008, pp. 58–71.

[91] F. Tsalakanidou, S. Malassiotis, Robust facial action recognition from real-time 3D streams, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2009, pp. 4–11.

[92] S. Canavan, Y. Sun, X. Zhang, L. Yin, A dynamic curvature based approach for facial activity analysis in 3D space, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2012, pp. 14–19.

[93] N. Michael, M. Dilsizian, D. Metaxas, J. Burgoon, Motion profiles for deception detection using visual cues, European Conference on Computer Vision, 2010, pp. 462–475.

[94] M. Shreve, S. Godavarthy, D. Goldgof, S. Sarkar, M. Shreve, S. Godavarthy, D. Goldgof, S. Sarkar, Macro- and micro-expression spotting in long videos using spatio-temporal strain, IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, IEEE, 2011, pp. 51–56.

[95] Q. Wu, X. Shen, X. Fu, The machine knows what you are hiding: an automatic micro-expression recognition system, Affect. Comput. Intell. Interact. (2011) 152–162.

[96] H. Dibeklioğlu, A. Salah, T. Gevers, Are you really smiling at me? Spontaneous versus posed enjoyment smiles, European Conference on Computer Vision, 2012, pp. 525–538.

[97] X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikäinen, A spontaneous micro-expression database: inducement, collection and baseline, IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, 2013.

[98] Y. Wang, X. Huang, C. Lee, S. Zhang, Z. Li, D. Samaras, D. Metaxas, A. Elgammal, P. Huang, High resolution acquisition, learning and transfer of dynamic 3-D facial expressions, Computer Graphics Forum, vol. 23, Wiley Online Library, 2004, pp. 677–686.

[99] C. Lee, A. Elgammal, D. Metaxas, Synthesis and control of high resolution facial expressions for visual interactions, IEEE International Conference on Multimedia and Expo, IEEE, 2006, pp. 65–68.

[100] P. Yang, Q. Liu, D. Metaxas, Boosting coded dynamic features for facial action units and facial expression recognition, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–6.

[101] P. Yang, Q. Liu, D. Metaxas, Boosting encoded dynamic features for facial expression recognition, Pattern Recognit. Lett. 30 (2) (2009) 132–139.

[102] P. Yang, Q. Liu, X. Cui, D. Metaxas, Facial expression recognition using encoded dynamic features, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.

[103] P. Yang, Q. Liu, D. Metaxas, Exploring facial expressions with compositional features, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 2638–2644.

[104] P. Yang, Q. Liu, D. Metaxas, RankBoost with l1 regularization for facial expression recognition and intensity estimation, IEEE 12th International Conference on Computer Vision, Ieee, 2009, pp. 1018–1025.

[105] S. Mitra, T. Acharya, Gesture recognition: a survey, IEEE Trans. Syst. Man Cybern. B Cybern. 37 (3) (2007) 311–324.

[106] N. Howe, M. Leventon, W. Freeman, Bayesian reconstruction of 3D human motion from single-camera video, Neural Information Processing Systems, vol. 1999, 1999, p. 1, (Cambridge, MA).

[107] R. Kehl, M. Bray, L. Van Gool, Full body tracking from multiple views using stochastic sampling, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, Ieee, 2005, pp. 129–136.

[108] R. Plankers, P. Fua, Articulated soft objects for video-based body modeling, IEEE International Conference on Computer Vision, vol. 1, IEEE, 2001, pp. 394–401.

[109] G. Loy, M. Eriksson, J. Sullivan, S. Carlsson, Monocular 3D reconstruction of human motion in long action sequences, The European Conference on Computer Vision, 2004, pp. 442–455.

[110] J. Alon, V. Athitsos, Q. Yuan, S. Sclaroff, A unified framework for gesture recognition and spatiotemporal gesture segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (9) (2009) 1685–1699.

[111] V. Ferrari, M. Marin-Jimenez, A. Zisserman, Progressive search space reduction for human pose estimation, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.

[112] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 1385–1392.

[113] B. Sapp, D. Weiss, B. Taskar, Parsing human motion with stretchable models, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 1281–1288.

[114] L. Bourdev, J. Malik, Poselets: body part detectors trained using 3D human pose annotations, International Conference on Computer Vision, IEEE, 2009, pp. 1365–1372.

[115] V. Singh, R. Nevatia, C. Huang, Efficient inference with multiple heterogeneous part detectors for human pose estimation, European Conference on Computer Vision, 2010, pp. 314–327.

[116] D. Park, D. Ramanan, N-best maximal decoders for part models, IEEE International Conference on Computer Vision, IEEE, 2011, pp. 2627–2634.

[117] M. Sun, S. Savarese, Articulated part-based model for joint object detection and pose estimation, IEEE International Conference on Computer Vision, IEEE, 2011, pp. 723–730.

[118] M. Isard, A. Blake, Condensation—conditional density propagation for visual tracking, Int. J. Comput. Vis. 29 (1) (1998) 5–28.

[119] J. Deutscher, A. Blake, I. Reid, Articulated body motion capture by annealed particle filtering, IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE, 2000, pp. 126–133.

[120] L. Sigal, S. Bhatia, S. Roth, M.J. Black, M. Isard, Tracking loose-limbed people, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, IEEE, 2004, (I–421).

[121] A.O. Balan, L. Sigal, M.J. Black, J.E. Davis, H.W. Haussecker, Detailed human shape and pose from images, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.

[122] R. Rosales, S. Sclaroff, Learning body pose via specialized maps, NIPS, vol. 1, 2002, p. 2.

[123] C. Sminchisescu, A. Kanaujia, Z. Li, D. Metaxas, Discriminative density propagation for 3D human motion estimation, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, IEEE, 2005, pp. 390–397.

[124] A. Agarwal, B. Triggs, 3D human pose from silhouettes by relevance vector regression, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE, 2004, (II–882).

[125] L. Sigal, A. Balan, M. Black, Combined discriminative and generative articulated pose and non-rigid shape estimation, Adv. Neural Inf. Process. Syst. 20 (2007) 1337–1344.

[126] M.W. Lee, I. Cohen, Proposal maps driven MCMC for estimating human body pose in static images, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE, 2004, (II–334).

[127] E. Simo-Serra, A. Ramisa, G. Alenya, C. Torras, F. Moreno-Noguer, Single image 3D human pose estimation from noisy observations, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2673–2680.

[128] L. Zelnik-Manor, On sifts and their scales, IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2012, pp. 1522–1528.

[129] B. Liu, S. Gould, D. Koller, B. Liu, S. Gould, D. Koller, Single image depth estimation from predicted semantic labels, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 1253–1260.

[130] Behave, http://homepages.inf.ed.ac.uk/rbf/behave/.

[131] R. Poppe, A survey on vision-based human action recognition, Image Vis. Comput. 28 (6) (2010) 976–990.

[132] J. Liu, Y. Yang, I. Saleemi, M. Shah, Learning semantic features for action recognition via diffusion maps, Comput. Vis. Image Underst. 116 (3) (2012) 361–377.

[133] V. Kellokumpu, G. Zhao, M. Pietikäinen, Recognition of human actions using texture descriptors, Mach. Vis. Appl. 22 (5) (2011) 767–780.

[134] S. Ali, M. Shah, Human action recognition in videos using kinematic features and multiple instance learning, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2) (2010) 288–303.

[135] J. Liu, M. Shah, B. Kuipers, S. Savarese, Cross-view action recognition via view knowledge transfer, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 3209–3216.
[136] V. Kellokumpu, G. Zhao, M. Pietikäinen, Human activity recognition using a dynamic texture based method, BMVC08, 2008. 1–10.
[137] A. Oikonomopoulos, M. Pantic, I. Patras, Sparse B-spline polynomial descriptors for human activity recognition, Image Vis. Comput. 27 (12) (2009) 1814–1825.
[138] R. Souvenir, J. Babbs, Learning the viewpoint manifold for action recognition, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–7.
[139] R. Souvenir, K. Parrigan, Viewpoint manifolds for action recognition, J. Image Video Process. 2009 (2009) 1.
[140] J. Konrad, Motion detection and estimation, Handbook of Image and Video Processing, 2nd Edition, 2010.
[141] N. Friedman, S. Russell, Image segmentation in video sequences: a probabilistic approach, Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 1997, pp. 175–181.
[142] I. Haritaoglu, D. Harwood, L. Davis, W^4: real-time surveillance of people and their activities, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 809–830.
[143] K. Smith, P. Quelhas, D. Gatica-Perez, Detecting abandoned luggage items in a public space, Proceedings of the 9th IEEE International Workshop on Performance Evaluation in Tracking and Surveillance (PETS'06), 2006, pp. 75–82.
[144] Y. Zhou, S. Yan, T. Huang, Pair-activity classification by bi-trajectories analysis, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[145] B. Ni, S. Yan, A. Kassim, Recognizing human group activities with localized causalities, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1470–1477.
[146] T. Yu, S. Lim, K. Patwardhan, N. Krahnstoever, Monitoring, recognizing and discovering social networks, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1462–1469.
[147] Q. Yu, G. Medioni, Motion pattern interpretation and detection for tracking moving vehicles in airborne video, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 2671–2678.
[148] E. Swears, A. Hoogs, Functional scene element recognition for video scene analysis, Workshop on Motion and Video Computing, 2009. WMVC'09, IEEE, 2009, pp. 1–8.
[149] I. Laptev, T. Lindeberg, Space–time interest points, Proceedings of the International Conference on Computer Vision, 2003, pp. 432–439.
[150] C. Harris, M. Stephens, A combined corner and edge detector, Proceedings of the Alvey Vision Conference, 1988, pp. 147–151.
[151] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS¡¯05), 2005, pp. 65–72.
[152] K. Rapantzikos, Y. Avrithis, S. Kollias, Dense saliency-based spatiotemporal feature points for action recognition, Proceedings of the Conference on Computer Vision and Pattern Recognition, 2009, pp. 1454–1461.
[153] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, Proceedings of International Conference on Pattern Recognition, 2004, pp. 32–36.
[154] J. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, Int. J. Comput. Vis. 79 (3) (2008) 299–318.
[155] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L.V. Gool, A comparison of affine region detectors, Int. J. Comput. Vis. 65 (2005) 43–72.
[156] N. Ikizler, P. Duygulu, Histogram of oriented rectangles: a new pose descriptor for human action recognition, Image Vis. Comput. 27 (10) (2009) 1515–1526.
[157] Z. Zhao, A. Elgammal, Human activity recognition from frame's spatiotemporal representation, Proceedings of the International Conference on Pattern Recognition, 2008, pp. 1–4.
[158] S. Nowozin, G. Bakir, K. Tsuda, Discriminative subsequence mining for action classification, Proceedings of the International Conference on Computer Vision, 2007, pp. 1–8.
[159] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, International Conference on Multimedia, 2007, pp. 357–360.
[160] J. Liu, S. Ali, M. Shah, Recognizing human actions using multiple features, Proceedings of the Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[161] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, Proceedings of the Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[162] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos 'in the wild', Proceedings of the Conference on Computer Vision and Pattern Recognition, 2009, pp. 1996–2003.
[163] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, R. Nevatia, Event detection and analysis from video streams, IEEE Trans. Pattern Anal. Mach. Intell. 23 (8) (2001) 873–889.
[164] A. Prati, S. Calderara, R. Cucchiara, Using circular statistics for trajectory shape analysis, Proceedings of the Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[165] E.L. Andrade, S. Blunsden, R.B. Fisher, Modelling crowd scenes for event detection, Proceedings of the International Conference on Pattern Recognition, 2006, pp. 175–178.
[166] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, J. Rehg, A scalable approach to activity recognition based on object use, Proceedings of the International Conference on Computer Vision, 2007, pp. 1–8.
[167] D. Helbing, P. Molnar, Social force model for pedestrian dynamics, Physical Review E.
[168] W. Yu, A. Johansson, Modeling crowd turbulence by many-particle simulations, Phys. Rev. E 76 (4) (2007) 046105.
[169] A. Treuille, S. Cooper, Z. Popovic, Continuum crowds, ACM Trans. Graph. 25 (3) (2006) 1160–1168.
[170] S. Ali, M. Shah, Floor fields for tracking in high density crowd scenes, Proceedings of the European Conference on Computer Vision, 2008, pp. 1–14.
[171] G. Antonini, S.V. Martinez, M. Bierlaire, J.P. Thiran, Behavioral priors for detection and tracking of pedestrians in video sequences, Int. J. Comput. Vis. (2006) 159–180.
[172] P. Scovanner, M.F. Tappen, Learning pedestrian dynamics from the real world, Proceedings of the International Conference on Computer Vision, 2009, pp. 381–388.
[173] S. Pellegrini, A. Ess, K. Schindler, L.V. Gool, You'll never walk alone: modeling social behavior for multi-target tracking, Proceedings of the International Conference on Computer Vision, 2009, pp. 261–268.
[174] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, Proceedings of the Conference on Computer Vision and Pattern Recognition, 2009, pp. 935–942.
[175] S. Wu, B. Moore, M. Shah, Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes, Proceedings of the Conference on Computer Vision and Pattern Recognition, 2010, pp. 2054–2060.
[176] X. Cui, Q. Liu, M. Gao, D. Metaxas, Abnormal detection using interaction energy potentials, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 3161–3167.
[177] S. Lu, D. Metaxas, D. Samaras, J. Oliensis, Using multiple cues for hand tracking and model refinement, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE, 2003, pp. 443–450.
[178] D. Metaxas, Physics-based Deformable Models: Applications to Computer Vision, Graphics, and Medical Imaging, vol. 389, Springer, 1997.
[179] L. Kakadiaris, D. Metaxas, Model-based estimation of 3D human motion, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1453–1459.
[180] D. Metaxas, B. Liu, F. Yang, P. Yang, N. Michael, C. Neidle, Recognition of nonmanual markers in American Sign Language (ASL) using non-parametric adaptive 2D–3D face tracking, The International Conference on Language Resources and Evaluation, 2012.
[181] J. Liu, B. Liu, S. Zhang, F. Yang, P. Yang, D.N. Metaxas, C. Neidle, Recognizing eyebrow and periodic head gestures using CRFs for non-manual grammatical marker detection in ASL, IEEE International Conference on Automatic Face and Gesture Recognition, 2013.
[182] J. Huang, S. Zhang, H. Li, D. Metaxas, Composite splitting algorithms for convex optimization, Comput. Vis. Image Underst. 115 (12) (2011) 1610–1622.
[183] J. Huang, S. Zhang, D. Metaxas, Fast optimization for mixture prior models, European Conference on Computer Vision, 2010, pp. 607–620.
[184] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc. B (Stat. Methodol.) 68 (1) (2005) 49–67.
[185] J. Huang, T. Zhang, The benefit of group sparsity, Ann. Stat. 38 (4) (2010) 1978–2004.
[186] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, D. Metaxas, Automatic image annotation using group sparsity, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 3312–3319.
[187] C. Neidle, C. Vogler, A new web interface to facilitate access to corpora: development of the ASLLRP data access interface, Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon, 2012.
[188] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, A. Thangali, The American Sign Language lexicon Video Dataset, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08, IEEE, 2008, pp. 1–8.