Journal of Machine Learning Research 17 (2016) 1-54 Submitted 11/14; Revised 1/16; Published 4/16

# Challenges in multimodal gesture recognition

**Sergio Escalera** SERGIO@MAIA.UB.ES
*Computer Vision Center UAB and University of Barcelona*
**Vassilis Athitsos** ATHITSOS@UTA.EDU
*University of Texas*
**Isabelle Guyon** GUYON@CHALEARN.ORG
*ChaLearn, Berkeley, California*

**Editors:** Zhuowen Tu

## Abstract

This paper surveys the state of the art on multimodal gesture recognition and introduces the JMLR special topic on gesture recognition 2011-2015. We began right at the start of the Kinect$^{TM}$ revolution when inexpensive infrared cameras providing image depth recordings became available. We published papers using this technology and other more conventional methods, including regular video cameras, to record data, thus providing a good overview of uses of machine learning and computer vision using multimodal data in this area of application. Notably, we organized a series of challenges and made available several datasets we recorded for that purpose, including tens of thousands of videos, which are available to conduct further research. We also overview recent state of the art works on gesture recognition based on a proposed taxonomy for gesture recognition, discussing challenges and future lines of research.

**Keywords:** Gesture Recognition, Time Series Analysis, Multimodal Data Analysis, Computer Vision, Pattern Recognition, Wearable sensors, Infrared Cameras, Kinect$^{TM}$.

## 1. Introduction

Gestures are naturally performed by humans. Gestures are produced as part of deliberate actions, signs or signals, or subconsciously revealing intentions or attitude. They may involve the motion of all parts of the body, but the arms and hands, which are essential for action and communication, are often the focus of studies. Facial expressions are also considered gestures and provide important cues in communication.

Gestures are present in most daily human actions or activities, and participate to human communication by either complementing speech or substituting themselves to spoken language in environments requiring silent communication (under water, noisy environments, secret communication, etc.) or for people with hearing disabilities. The importance of gestures in communication is rooted in primal behaviors: the gesture-first theory, supported by the analysis of mirror neurons in primates (Hewes, 1973), indicated that the first steps of language phylogenetically were not speech, nor speech with gesture, but were gestures alone (McNeil, 2012; Hewes, 1973). See examples of primate communication by means of gestures in Figure 1.

Given the indubitable importance of gestures in human activities, there has been huge interest by the Computer Vision and Machine Learning communities to analyze human gestures from visual data in order to offer new non-intrusive technological solutions. For completeness, in this paper we

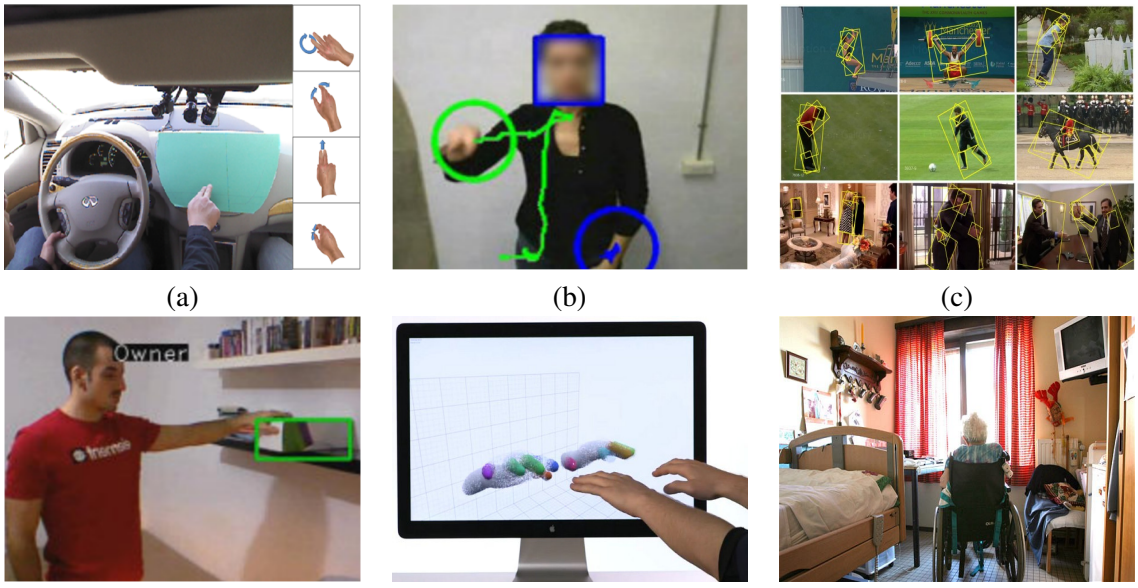

Figure 1: Example of possible bonobo iconic gestures. (a) Start of swing gesture (or shove); (b) End of swing gesture (or shove); (c) Start of iconic swing, other bonobo starts to move; (d) End of iconic swing, other moving. Image from (McNeil, 2012).

also review some gesture recognition systems with data acquired from wearable sensors, although the comprehensive review of papers focus on the analysis of different visual modalities.

Applications are countless, like Human Computer Interaction (HCI), Human Robot Interaction (HRI) (also named human machine interaction HMI), communication, entertainment, security, art, semiotics, commerce and sports, while having an important social impact in assistive technologies for the handicapped and the elderly. Some examples of applications are illustrated in Fig. 2.

In addition to the recent advances in human and gesture recognition from classical RGB visual data, the automatic analysis of human body from sensor data keeps making rapid progress with the constant improvement of (i) new published methods that constantly push the state-of-the-art and (ii) the recent availability of inexpensive 3D video sensors such as Kinect$^{TM}$, providing a complementary source of information, and thus allowing the computation of new discriminative feature vectors and improved recognition by means of fusion strategies. In section 2 we review the state of the art in gesture recognition.

In order to push research and analyze the gain of multimodal methods for gesture recognition, in 2011 and 2012, ChaLearn organized a challenge on single user one-shot-learning gesture recognition with data recorded with Kinect$^{TM}$in which 85 teams competed. Starting from baseline methods making over 50% error (measured in Leveinshtein distance, a metric counting the number of substitutions, insertions and deletions, analogous to an error rate), the winners brought the error rate below 10%. While there was still some margin of improvement on such tasks to reach human performance (which is below 2% error), we were encouraged to make the task harder to push the state of the art in computer vision. In our second ChaLearn challenge on Multimodal Gesture Recognition in 2013, we proposed a user-independent task with data recorded with Kinect$^{TM}$, with a larger vocabulary and continuously performed gestures. Of 60 participating teams, the winner attained an error rate of 10% on this data set, in terms of Leveinshtein distance. In 2014, we used the same Multimodal Gesture Recognition dataset with the objective of performing gesture spotting. The winner of the competition, with a deep learning architecture, obtained an overlapping near 0.9. Lastly, in 2014 and 2015 we ran an action spotting challenge with a new dataset consisting of RGB sequences of actors performing different isolated and collaborative actions in outdoor environments. Future challenges we are planning include the analysis of gestures taking into account face and contextual information, involving many modalities in the recognition process. In this paper we also review other existing international challenges related to gesture recognition.




(a) (b) (c)

(d) (e) (f)

Figure 2: Some applications of gesture recognition. (a) Gesture recognition for driver assistance, from (Ohn-Bar and Trivedi, 2014), (b) Sign Language Recognition, (c) action/gesture recognition for content retrieval and categorization, from (Ma et al., 2013), (d) surveillance, (e) Human Computer/Robot/Machine Interaction, and (f) Assistive technology for people with reduced autonomy.

Our first workshop at CVPR from our 2011 challenge emphasized mostly 2D video data meanwhile our second and third workshops at CVPR, ICPR, ICMI, ECCV conferences from our 2012, 2013, and 2014 challenges were focused on affordable 3D sensors for gesture recognition research, also including audio information. In ECCV 2014 and CVPR 2015 workshops we also promoted different aspects of looking at people, including pose recovery, activity recognition, and scene understanding where humans are present. In addition to best challenge results, many research papers devoted to gesture recognition were published and presented in our challenge workshops. We also invited keynote speakers in diverse areas of pose and gesture research, including sign language recognition, body posture analysis, action and activity recognition, and facial expression or emotion recognition.

In this special topic on gesture recognition, extension of best challenge and workshop papers from previous events have been published. In addition, new description and learning strategies papers related to gesture recognition have been published. All of them will be shortly reviewed in the following sections.

The rest of the paper is organized as follows: Section 2 reviews the state of the art on gesture recognition, defining a taxonomy to describe existing works as well as available databases for gesture and action recognition. Section 3 describes the series of gesture and action recognition challenges organized by ChaLearn, describing the data, objectives, schedule, and achieved results by the participants. For completeness we also review other existing gesture challenge organizations. In Section 4 we review the published papers in this gesture recognition topic which are

related to ChaLearn competitions. Section 5 describes special topic published papers related to gesture recognition which are not based on ChaLearn competitions. Finally, Section 6 discusses main observations about the published papers.

## 2. Related Work in Gesture Recognition

In this section we present a taxonomy for action/gesture recognition, we review most influential works in the field, and finally we review existing datasets for action/gesture recognition together with the performance obtained by state of the art methods.

### 2.1 Taxonomy for gesture recognition

Fig. 3 is an attempt to create a taxonomy of the various components involved in conducting research in action/gesture recognition. We include various aspects relating to the problem setting, the data acquisition, the tools, the solutions, and the applications.

First, regarding the problem setting, the interpretation of gestures critically depends on a number of factors, including the environment in which gestures are performed, their span in time and space, and the intentional meaning in terms of symbolic description and/or the subconscious meaning revealing affective/emotional states. The problem setting also involves different actors who may participate in the execution of gestures and actions: human(s) and/or machine(s) (robot, computer, etc.), performing with or without tools or interacting or not with objects. Additionally, independently of the considered modality, for some gestures/actions different parts of the body are involved. While many gesture recognition systems only focus on arms and hands, full body motion/configuration and facial expressions can also play a very important role. Another aspect of the problem setting involves whether recognized gestures are static or dynamic. For the first case, just considering features from an input frame or any other acquisition device describing spatial configuration of body limbs, a gesture can be recognized. In the second case, the trajectory and pose of body limbs provide the highest discriminative information for gesture recognition. In some settings, gestures are defined based not only on the pose and motion of the human, but also on the surrounding context, and more specifically on the objects that the human interacts with. For such settings, one approach for achieving context awareness is scene analysis, where information is extracted from the scene around the subject (e.g., Pieropan et al. (2014); Shapovalova et al. (2011)). Another approach is to have the subject interact with intelligent objects. Such objects use embedded hardware and software to facilitate object recognition/localization, and in some cases to also monitor interactions between such objects and their environment (e.g., Czabke et al. (2010))

Second, the data are, of course, of very central importance, as in every machine learning application. The data sources may vary: when recognizing gestures, input data can come from different modalities, visual (RGB, 3D, or thermal, among others), audio, or wearable sensors (magnetic field trackers, instrumented (data) gloves, or body suits, among others). In the case of gloves, they can be active or passive. Active ones make use of a variety of sensors on a glove to measure the flexing of joints or the acceleration and communicates data to the host device using wired or wireless technology. Passive ones consist only of markers or colored gloves for finger detection by an external device such as a camera. Although most gestures are recognized by means of ambient intelligent systems, looking at the person from outside, some gesture recognition approaches are based on egocentric computing, using wearable sensors or wearable cameras that analyze, for instance, hand behaviors. Additionally, it is well-known that context provides rich information that can be useful

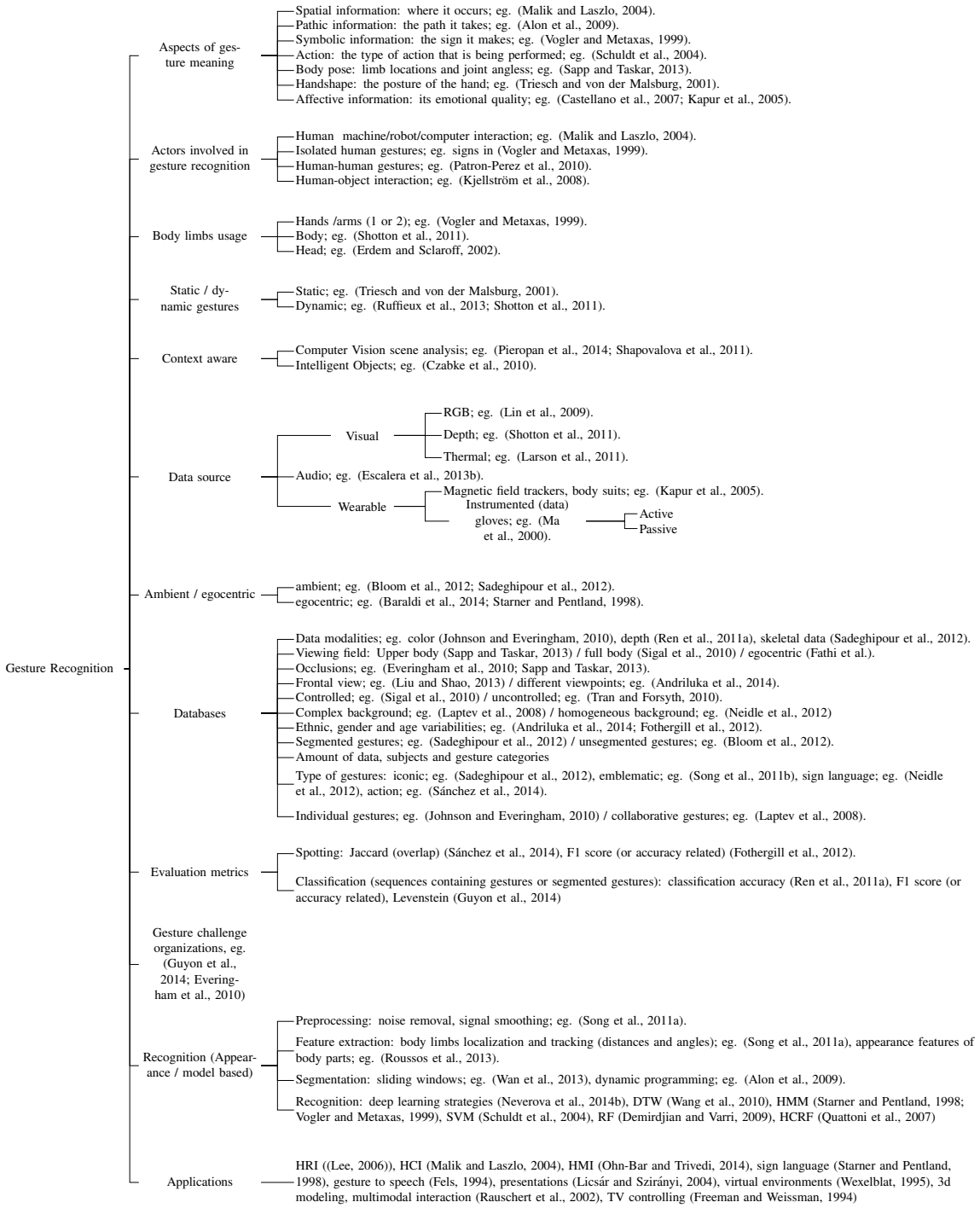
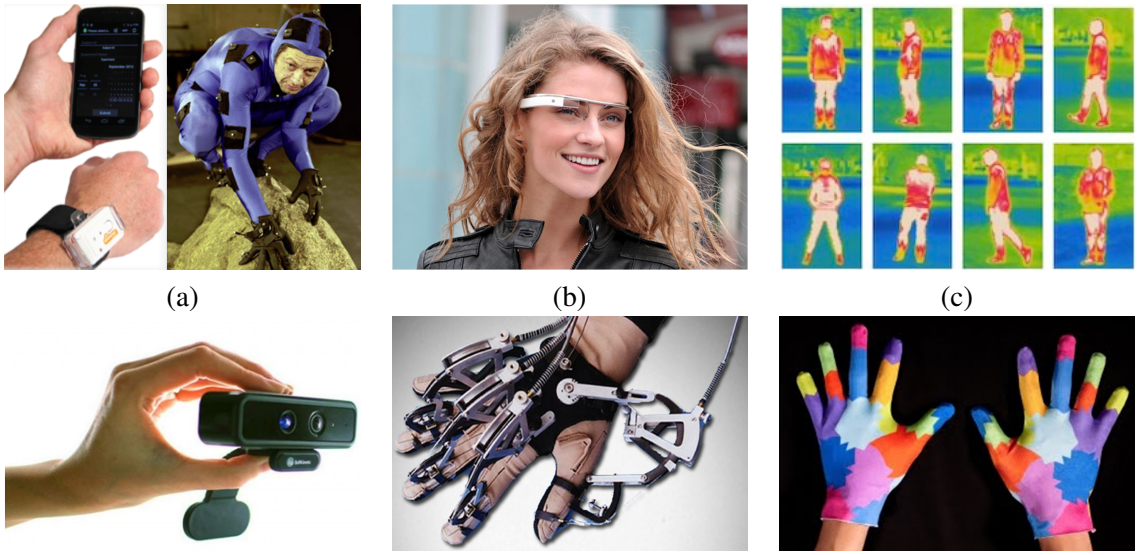



Figure 3: Taxonomy for gesture recognition.

to better infer the meaning of some gestures. Context information can be obtained by means of computer vision scene analysis, interaction with objects, but also via intelligent objects in the scene



(a) (b) (c)

(d) (e) (f)

Figure 4: Some examples of acquisition devices for gesture recognition. (a) left: mobile with GPS and accelerometer, right: inertial sensor with accelerometer and gyroscope, (b) Google Glasses for egocentric computing, (c) thermal imagery for action recognition, (d) audio-RGB-depth device, (e) active glove, and (f) passive glove.

(objects with sensors that emit signals related to proximity and interaction). Some examples of acquisition devices are shown in Figure 4.

Third, the field of gesture recognition has shaped up thanks to the adoption of standard methodology. In order to advance in the design of robust action/recognition approaches, several datasets with different complexity have been published, and several world challenges helped to push the research in the area. This required the definition of standard evaluation metrics to render methods comparable. Notably, when one wants to recognize actions/gestures from data, common steps involve pre-processing of the acquired data, feature extraction, segmentation of begin-end of gesture and its final gesture/action label classification. Many datasets include preprocessed and/or thoroughly annotated data.

Fourth, gesture recognition has offered many opportunities to algorithm developers to innovate. The approaches, which essentially can be categorized into appearance-based and model-based methods, are going to be reviewed in the next section. We will mention only the most influential works for action/gesture recognition illustrating various aspects of the problem setting, data acquisition, and methodology defined in our taxonomy. Note that although we defined a general taxonomy for gesture recognition, in this paper, we put special emphasis on computer vision and machine learning methods for action/gesture recognition.

Finally, our taxonomy would not be complete without the wide array of applications of gesture/action recognition, already mentioned in the introduction.

## 2.2 Overview of gesture recognition methods

Different surveys have been published so far reviewing gesture recognition systems (LaViola Jr., 1999; Mitra and Acharya, 2007; Chaudhary et al., 2011; Ibraheem and Khan, 2012; Avci et al., 2010; Khan and Ibraheem; Kausar and Javed, 2011). In this section, we present an up-to-date review of most influential works in the field.

### 2.2.1 RECOGNIZING STATIC GESTURES AND HAND POSE

In the case of static gestures, frequently hand shape is the important differentiating feature (Cui and Weng, 2000; Freeman and Roth, 1996; Kelly et al., 2010; Ren et al., 2011b; Triesch and von der Malsburg, 2002), although the pose of the rest of the body can also be important, e.g., (Yang et al., 2010; Van den Bergh et al., 2009). For static hand pose classification, some approaches rely on visual markers, such as a color glove with a specific color for each finger, e.g., (Wang and Popović, 2009). Other approaches can recognize the hand pose on unadorned hands. Appearance-based methods, like (Moghaddam and Pentland, 1995; Triesch and von der Malsburg, 2002; Freeman and Roth, 1996; Wu and Huang, 2000), can be used for recognizing static hand postures observed from specific viewpoints.

Model-based methods for hand pose estimation (Oikonomidis et al., 2011; de La Gorce et al., 2011; Oikonomidis et al., 2010; Rehg and Kanade, 1995) typically match visual observations to instances of a predefined hand model. Single frame pose estimation methods try to solve the hand pose estimation problem without relying on temporal information (Athitsos and Sclaroff, 2003). Most recently, due to the advent of commercially available depth sensors, there is an increased interest in methods relying on depth data (Keskin et al., 2012; Mo and Neumann, 2006; Oikonomidis et al., 2011; Pugeault and Bowden, 2011; Lopes et al., 2014).

### 2.2.2 FROM BODY PART DETECTION TO HOLISTIC PATTERN DETECTION

Dynamic gestures are characterized by both the pose and the motion of the relevant body parts. Much effort has traditionally be put into detecting first **body parts** and then tracking their motion. In color videos, detecting hands can be quite challenging, although better performance can be achieved by placing additional constraints on the scene and the relative position of the subject and the hands with respect to the camera (Cui and Weng, 2000; Isard and Blake, 1998; Kolsch and Turk, 2004; Ong and Bowden, 2004; Stefanov et al., 2005; Stenger et al., 2003; Sudderth et al., 2004). Commonly-used visual cues for hand detection such as skin color, edges, motion, and background subtraction (Chen et al., 2003; Martin et al., 1998) may also fail to unambiguously locate the hands when the face, or other "hand-like" objects are moving in the background.

In (Li and Kitani, 2013) the authors propose a hand segmentation approach from egocentric RGB data by the combination of color and texture features. In(Baraldi et al., 2014), dense features are extracted around regions selected by a new hand segmentation technique that integrates super-pixel classification, temporal and spatial coherence. Bag of visual words and linear SVM are used for final representation and classification.

Depth cameras have become widely available in recent years, and hand detection (in tandem with complete body pose estimation) using such cameras (and also in combination with other visual modalities) can be performed sufficiently reliably for many applications (Shotton et al., 2011; Hernandez-Vela et al., 2012). The authors of (Ren et al., 2013) propose a part-based hand gesture recognition system using Kinect$^{TM}$sensor. Finger-EarthMover's Distance (FEMD) metric is pro-

posed to measure the dissimilarity between hand shapes. It matches the finger parts while not the whole hand based on hand segmentation and contour analysis. The method is tested on their own 10-gesture dataset.

Instead of estimating hand position and/or body pose before recognizing the gesture, an alternative is to customize the recognition module so that it does not require the exact knowledge of hand positions, but rather accepts as input a list of several candidate hand locations (Alon et al., 2009; Sato and Kobayashi, 2002; Hernandez-Vela et al., 2013b).

Another approach is to use **global image/video features**. Such global features include motion energy images (Bobick and Davis, 2001), thresholded intensity images and difference images (Dreuw et al., 2006), 3D shapes extracted by identifying areas of motion in each video frame (Gorelick et al., 2007) and histograms of pairwise distances of edge pixels (Nayak et al., 2005). Gestures can also be modelled as rigid 3D patterns (Ke et al., 2005), from which features can be extracted using 3D extensions of rectangle filters (Viola and Jones, 2001). The work of (Kong et al., 2015) uses pixel-level attributes in a hierarchical architecture of 3D kernel descriptors, and efficient match kernel is used to recognize gestures from depth data.

Along similar lines, (Ali and Shah, 2010) propose a set of kinematic features that are derived from the optical flow for human action recognition in videos: divergence, vorticity, symmetric and antisymmetric flow fields, second and third principal invariants of flow gradient and rate of strain tensor, and third principal invariant of rate of rotation tensor, which define spatiotemporal patterns. These kinematic features are computed by Principal Component Analysis (PCA). Then multiple instance learning (MIL) is applied for recognition in which each action video is represented by a bag of kinematic modes. The proposal is evaluated on the RGB Weizmann and KTH action data sets, showing comparable result to state of the art performances.

Much effort has also been put into **spatiotemporal invariant features**. In (Yuan et al., 2011) the authors propose a RGB action recognition system based on a pattern matching approach, named naive Bayes mutual information maximization (NBMIM). Each action is characterized by a collection of spatiotemporal invariant features which are matched with an action class by measuring the mutual information between them. Based on this matching criterion, action detection is to localize a subvolume in the volumetric video space that has the maximum mutual information toward a specific action class. A novel spatiotemporal branch-and-bound (STBB) search algorithm is designed to efficiently find the optimal solution. Results show high recognition results on KTH, CMU, and MSR data sets, showing speed up inference in comparison with standard 3D branch-and-bound.

Another example is the paper of (Derpanis et al., 2013) in which a compact local descriptor of video dynamics is proposed for action recognition in RGB data sequences. The descriptor is based on visual spacetime oriented energy measurements. An associated similarity measure is introduced that admits efficient exhaustive search for an action template, derived from a single exemplar video, across candidate video sequences. The method is speeded up by means of a GPU implementation. Method is evaluated on UCF and KTH data sets, showing comparable results to state of the art methods.

The work of (Yang and Tian, 2014b) presents a coding scheme to aggregate low-level descriptors into the super descriptor vector (SDV). In order to incorporate the spatio-temporal information, the super location vector (SLV) models the space-time locations of local interest points in a compact way. SDV and SLV are combined as the super sparse coding vector (SSCV) which jointly models the motion, appearance, and location cues. The approach is tested on HMDB51 and Youtube with higher performance in comparison to state of the art approaches.

### 2.2.3 SEGMENTATION OF GESTURES AND GESTURE SPOTTING

Dynamic gesture recognition methods can be further categorized based on whether they make the assumption that gestures have already been segmented, so that the start frame and end frame of each gesture is known. Gesture spotting is the task of recognizing gestures in unsegmented video streams, that may contain an unknown number of gestures, as well as intervals were no gesture is being performed. Gesture spotting methods can be broadly classified into two general approaches: the direct approach, where temporal segmentation precedes recognition of the gesture class, and the indirect approach, where temporal segmentation is intertwined with recognition:

- **Direct methods** (also called heuristic segmentation) first compute low-level motion parameters such as velocity, acceleration, and trajectory curvature (Kang et al., 2004) or mid-level motion parameters such as human body activity (Kahol et al., 2004), and then look for abrupt changes (e.g., zero-crossings) in those parameters to identify candidate gesture boundaries.

- **Indirect methods** (also called recognition-based segmentation) detect gesture boundaries by finding, in the input sequence, intervals that give good recognition scores when matched with one of the gesture classes. Most indirect methods (Alon et al., 2009; Lee and Kim, 1999; Oka, 1998) are based on extensions of Dynamic Programming (DP) e.g., Dynamic Time Warping (DTW) (Darrell et al., 1996; Kruskal and Liberman, 1983), Continuous Dynamic Programming (CDP) (Oka, 1998), various forms of Hidden Markov Models (HMMs) (Brand et al., 1997; Chen et al., 2003; Stefanov et al., 2005; Lee and Kim, 1999; Starner and Pentland, 1998; Vogler and Metaxas, 1999; Wilson and Bobick, 1999), and most recently, Conditional Random Fields (Lafferty et al., 2001; Quattoni et al., 2007). Also hybrid probabilistic and dynamic programming approaches have been recently published (Hernandez-Vela et al., 2013a). In those methods, the gesture endpoint is detected by comparing the recognition likelihood score to a threshold. The threshold can be fixed or adaptively computed by a non-gesture garbage model (Lee and Kim, 1999; Yang et al., 2009), equivalent to silence models in speech.

When attempting to recognize unsegmented gestures, a frequently encountered problem is the *subgesture problem*: false detection of gestures that are similar to parts of other longer gestures. (Lee and Kim, 1999) address this issue using heuristics to infer the user's completion intentions, such as moving the hand out of camera range or freezing the hand for a while. An alternative is proposed in (Alon et al., 2009), where a learning algorithm explicitly identifies subgesture/supergesture relationships among gesture classes, from training data.

Another common approach for gesture spotting is to first extract features from each frame of the observed video, and then to provide a sliding window of those features to a recognition module, which performs the classification of the gesture (Corradini, 2001; Cutler and Turk, 1998; Darrell et al., 1996; Oka et al., 2002; Starner and Pentland, 1998; Yang et al., 2002)). Oftentimes, the extracted features describe the position and appearance of the gesturing hand or hands (Cutler and Turk, 1998; Darrell et al., 1996; Starner and Pentland, 1998; Yang et al., 2002)). This approach can be integrated with recognition-based segmentation methods.

### 2.2.4 ACTION AND ACTIVITY RECOGNITION

The work of (Li et al., 2010) presents an action graph to model explicitly the dynamics of 3D actions and a bag of 3D points to characterize a set of salient postures that correspond to the nodes

in the action graph. The authors propose a projection based sampling scheme to sample the bag of 3D points from the depth maps. In (Sminchisescu et al., 2006) it is proposed the first conditional/discriminative chain model for action recognition.

The work of (Zanfir et al., 2013) propose the non-parametric Moving Pose (MP) framework for low-latency human action and activity recognition. The moving pose descriptor considers both pose information as well as differential quantities (speed and acceleration) of the human body joints within a short time window around the current frame. The descriptor is used with a modified kNN classifier that considers both the temporal location of a particular frame within the action sequence as well as the discrimination power of its moving pose descriptor compared to other frames in the training set. The method shows comparable results to state of the art methods on MSR-Action3D and MSR-DailyActivities3D data sets.

In (Oreifej and Liu, 2013), it is proposed a new descriptor for activity recognition from videos acquired by a depth sensor. The depth sequence is described using a histogram capturing the distribution of the surface normal orientation in the 4D space of time, depth, and spatial coordinates. To build the histogram, 4D projectors are created, which quantize the 4D space and represent the possible directions for the 4D normal. Projectors are initialized using the vertices of a regular polychoron. Projectors are refined using a discriminative density measure, such that additional projectors are induced in the directions where the 4D normals are more dense and discriminative. The proposed descriptor is tested on MSR Actions 3D, MSR Gesture 3D, and MSR Daily Activity 3D, slightly improving state of the art results.

In (Wang et al., 2014), the authors propose to characterize the human actions with an "actionlet" ensemble model, which represents the interaction of a subset of human joints. Authors train an ensemble of SVM classifiers related to actionlet patterns, which includes 3D joint features, Local Occupancy Patterns, and Fourier Temporal Pyramid. Results on CMU MoCap, MSR-Action3D, MSR-DailyActivity3D, Cornell Activity, and Multiview 3D data sets show comparable and better performance than state of the art approaches.

The work of (Yang and Tian, 2014a) presents an approach for activity recognition in depth video sequences. Authors cluster hypersurface normals in a depth sequence to form the polynormal which is used to jointly characterize the local motion and shape information. In order to globally capture the spatial and temporal orders, an adaptive spatio-temporal pyramid is introduced to subdivide a depth video into a set of space-time grids. It is then proposed a scheme of aggregating the low-level polynormals into the super normal vector (SNV) which can be seen as a simplified version of the Fisher kernel representation. Authors validate the proposed approach on MSRAction3D, MSRDailyActivity3D, MSRGesture3D, and MSRActionPairs3D data sets slightly improving in all cases state of the art performances.

In (Yu et al., 2014) the authors propose the orderlets to capture discriminative information for gesture recognition from depth maps. Orderlet features are discovered looking for frequent sets of skeleton joints that provide discriminative information. Adaboost is used for orderlets selection. Results on the ORGBD data set shows a recognition rate of 71.4% mean class average accuracy, improving by near 5% state of the art results on this data set, and near 20% improvement regarding frame level classification. However the results showed on the MSR-DailyActivity3D data set are inferior to the ones reported in (Luo et al., 2014).

The work of (Liang et al., 2014) presents a depth-based method for hand detection and pose recognition by segmentation of different hand parts. Authors based on RF for initial multipart hand

segmentation. Then, a Superpixel-Markov Random Field (SMRF) parsing scheme is used to enforce the spatial smoothness and the label co-occurrence prior to remove the misclassified regions.

### 2.2.5 APPROACHES USING NON-VIDEO MODALITIES AND MULTIMODAL APPROACHES

In terms of multimodal approaches for gesture recognition, (Luo et al., 2014) propose a sparse coding-based temporal pyramid matching approach (ScTPM) for feature representation using depth maps. The authors also propose the Center-Symmetric Motion Local Ternary Pattern (CS-Mltp) descriptor to capture spatial-temporal features from RGB videos. By fusing both RGB and Depth descriptors, the authors improve state of the art results on MSR-Action3D and MSR-DailyActivity3D data sets, with a 6% and 7% of improvement, respectively.

In (Ionescu et al., 2014), it is presented the Human3.6M data set, consisting of 3.6Million accurate 3D Human poses, acquired by recording the performance of 5 female and 6 male subjects, under 4 different viewpoints, for training realistic human sensing systems and for evaluating the next generation of human pose estimation models and algorithms. Authors also provide a set of large scale statistical models and evaluation baselines for the dataset illustrating its diversity.

In (Xiao et al., 2014) a wearable Immersion CyberGlove II is used to capture the hand posture and the vision-based Microsoft Kinect$^{TM}$takes charge of capturing the head and arm posture. An effective and real-time human gesture recognition algorithm is also proposed.

In (Liang et al., 2013) it is proposed to detect and segment different body parts using RGB and Depth data sequences. The method uses both temporal constraints and spatial features, and performs hand parsing and 3D fingertip localization for hand pose estimation. The hand parsing algorithm incorporates a spatial-temporal feature into a Bayesian inference framework to assign the correct label to each image pixel. The 3D fingertip localization algorithm adapts is based on geodesic extrema extraction to fingertip detection. The detected 3D fingertip locations are finally used for hand pose estimation with an inverse kinematics solver. The work of (Joshi et al., 2015) use random forest for both segmenting and classifying gesture categories from data coming from different sensors.

Although many works base only on inertial data (Benbasat and Paradiso, 2001; Berlemont et al., 2015), multimodal approaches are often considered in order to combine trajectory information will pose analysis based on visual data. The works of (Liu et al., 2014; Pardo et al., 2013) present approaches for gesture recognition based on the combination of depth and inertial data. In (Liu et al., 2014) skeleton obtained from depth data and data from inertial sensors are train within HMM in order to perform hand gesture recognition. A similar approach is presented in (Pardo et al., 2013), but also recognizing objects present in the scene and using DTW for recognition with the objective of performing ambient intelligent analysis to support people with reduced autonomy. In (Gowing et al., 2014), it is presented a comparison of WIMU aWireless/Wearable Inertial Measurement Unit and Kinect$^{TM}$. However, comparison is performed independently, without considering a fusion strategy.

The work of (Appenrodt et al., 2009) is one of the few that compare the performance of different segmentation approaches for gesture recognition comparing RGB, depth, and thermal modalities. They propose a simple segmentation approach of faces and one hand for recognizing letters and numbers for HCI. They obtained higher performance by the use of depth maps. Unfortunately no mutimodal fusion approaches are tested in order to analyze when each modality can complement the information provided by the rest of modalities.

The work of (Escalera et al., 2013b) summarizes a 2013 challenge on multimodal gesture recognition, where in addition to RGB and depth data, audio can be used to identify the performed gestures.

Few works considered context information in order to improve gesture/action recognition systems. In (Wilhelm) it is proposed to adapt gesture recognition based on a dialogue manager as a partially observable Markov decision process (POMDP). In (Caon et al., 2011) two Kinect$^{TM}$devices and smart objects are used to estimate proximity and adapt the recognition prior of some gestures.

The recent emergence of deep learning systems in computer vision have also been applied to action/gesture recognition systems. In (Neverova et al., 2014a), it is presented a deep learning based approach for hand pose estimation, targeting gesture recognition. The method integrates local and global structural information into the training objective. In (Nagi et al., 2011), deep neural network (NN) combining convolution and max-pooling (MPCNN) is proposed for supervised feature learning and classification of RGB hand gestures given by humans to mobile robots using colored gloves. The hand contour is retrieved by color segmentation, then smoothed by morphological image processing which eliminates noisy edges. The system classifies 6 gesture classes with 96% accuracy, improving performance of several state of the art methods. The work of (Duffner et al., 2014) presents an approach that classifies 3D gestures using jointly accelerometer and gyroscope signals from a mobile device using convolutional neural network with a specific structure involving a combination of 1D convolution. In (Molchanov et al., 2015) convolutional deep neural networks are used to fuse data from multiple sensors (short-range radar, a color camera, and a depth camera) and to classify the gestures in a driver assistance scenario.

### 2.3 Sign language recognition

An important application of gesture recognition is sign language recognition. American Sign Language (ASL) is used by 500,000 to two million people in the U.S. (Lane et al., 1996; Schein, 1989). Overall, national and local sign languages are used all over the world as the natural means of communication in deaf communities.

Several methods exist for recognizing isolated signs, as well as continuous signing. Some researchers have reported results on continuous signing with vocabularies of thousands of signs, using input from digital gloves, e.g., (Yao et al., 2006). However, glove-based interfaces are typically expensive for adoption by the general public, as well as intrusive, since the user has to wear one or two gloves connected with wires to a computer.

Computer vision methods for sign language recognition offer hope for cheaper, non-intrusive interfaces compared to methods using digital gloves. Several such methods have been proposed (Bauer et al., 2000; Cui and Weng, 2000; Dreuw et al., 2006; Kadir et al., 2004; Starner and Pentland, 1998; Vogler and Metaxas, 1999; Wang et al., 2010; Zieren and Kraiss, 2005). However, computer vision methods typically report lower accuracies compared to methods using digital gloves, due to the difficulty of extracting accurate information about the articulated pose and motion of the signer.

An important constraint limiting the accuracy of computer vision methods is the availability of training data. Using more examples per sign typically improves accuracy (see, e.g., (Kadir et al., 2004; Zieren and Kraiss, 2005)). However, existing datasets covering large vocabularies have only a limited number of examples per sign. As an example, the ASLLVD dataset (Athitsos et al., 2008) includes about 3,000 signs, but only two examples are available for most of the signs. Some interest-

ing research has aimed at enabling automated construction of large datasets. For example, (Cooper and Bowden, 2009) aim at automatically generating large corpora by automatically segmenting signs from close-captioned sign language videos. As another example, (Farhadi et al., 2007) propose a method where sign models are learned using avatar-produced data, and then transfer learning is used to create models adapted for specific human signers.

The recent availability of depth cameras such as Kinect$^{TM}$has changed the methodology and improved performance. Depth cameras provide valuable 3D information about the position and trajectory of hands in signing. Furthermore, detection and tracking of articulated human motion is significantly more accurate in depth video than in color video. Several approaches have been published in recent years that use depth cameras to improve accuracy in sign language recognition Conly et al. (2015); Wang et al. (2015a); Zafrulla et al. (2011).

### 2.4 Data sets for gesture and action recognition

Tens of gesture recognition datasets have been made available to the research community over the last several years. A summary of available datasets is provided in Table 1. In that table, for each data set we mark some important attributes of the dataset, such as the type of gestures it contains, the data modalities it provides, the viewing field, background, amount of data, and so on. Regarding the "occlusions" attribute in that table, we should clarify that it only refers to occlusions of the subject by other objects (or subjects), and not to self occlusions. Self occlusions are quite common in gestures, and are observed in most datasets. We should also note that, regarding the complexity of the background, dynamic and/or cluttered backgrounds can make gesture recognition challenging in color images and video. At the same time, a complex background can be quite easy to segment if depth or skeletal information is available, as is the case in several datasets on Table 1.

In order to be able to fit Table 1 in a single page, we had to use abbreviations quite heavily. Table 2 defines the different acronyms and abbreviations used in Table 1.

The datasets we have created for our challenges have certain unique characteristics, that differentiate them from other existing datasets. The CDG2011 dataset (Guyon et al., 2014) has a quite diverse collection of gesture types, including static, pantomime, dance, signs, and activities. This is in contrast to other datasets, that typically focus on only one or maybe two gesture types. Furthermore, the CDG2011 dataset uses an evaluation protocol that emphasizes one-shot learning, whereas existing data sets typically have several training examples for each class. The CDG2013 dataset introduces audio data to the mix of color and depth that is available in some other data sets, such as (Sadeghipour et al., 2012; Bloom et al., 2012).

| Dataset | Actors/ Objects | Body parts | Static/ Dynamic | Modalities | Viewing field | Occl. | Viewp. Var. | Controlled | Background | Variab. | Seg. | Amount | Subjects | Classes | Type | Indiv./ collab. | Eval. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HUMANEVA (Sigal et al., 2010) | IH | F | SF | MC,ST | F | No | V | C | SF,C | E,G | SF | 40000 F | 5 | CBP | BP | I | MPJPE |
| Human3.6M (Ionescu et al., 2014) | IH,O | F | SF | MC,D,ST | F | No | V | C | SF,C | G,AU,EU | SF | 3.6M | 11 | CBP | BP | I | Multiple |
| LEEDS SPORTS (Johnson and Everingham, 2010) | IH,O | F | SF | C | F | No | V | U | C,SF | AY,E,G | SF | 2000 F | U1 | CBP | BP | I | Fer. |
| Pascal VOC people (Everingham et al., 2010) | IH,HH,O | F | SF | C | F | Some | V | U | C,SF | A,E,G | SF | 632 F | U1 | CBP | BP | B | Pascal |
| UIUC People (Tran and Forsyth, 2010) | IH,O | F | SF | C | F | No | V | U | C,SF | AY,E,G | SF | 593 F | U1 | CBP | BP | I | Fer. |
| Buffy (Ferrari et al., 2008) | IH,HH,O | U | SF | C | U | Some | V | U | C,SF | AY,G | SF | 748 F | UM | CBP | BP | B | Fer. |
| Parse (Ramanan, 2006) | IH,O | F | SF | C | F | No | V | U | C,SF | AY,E,G | SF | 305 F | U1 | CBP | BP | MI | Ram. |
| MPII Pose (Andriluka et al., 2014) | IH,O | F | SF | C | F | Yes | V | U | C,SF | A,E,G | SF | 25000 F | U1 | CBP | BP | MI | Fer. |
| FLIC Pose (Sapp and Taskar, 2013) | IH,HH,O | U | SF | C | U | Some | V | U | C,SF | AY,E,G | SF | 5003 F | UM | CBP | BP | MI | Sap. |
| H3D (Bourdev and Malik, 2009) | IH,HH,O | U | SF | C | U | Some | V | U | C,SF | A,E,G | SF | 520 F | US | CBP | BP | MI | Sap. |
| CDG2011 (Guyon et al., 2014) | IH | Mixed | Mixed | C,D | MU | Few | Fixed | C | ST,C | E,G | No | 50000 G | 20 | CDG11 | Mixed | I | L |
| CDG2013 (Escalera et al., 2013b) | IH | H | Dynamic | A,C,D,ST | U | No | F | C | ST,C | G | No | 13858 G | 27 | 20 | E | I | L |
| 3DIG (Sadeghipour et al., 2012) | IH | H | Dynamic | C,D,ST | U | No | F | C | ST,U | AY,E,G | Yes | 1739 G | 29 | 20 | I | I | F |
| HuPBA8K+ (Sánchez et al., 2014) | IH,HH | F | Dynamic | C | F | Yes | Fixed | C | ST,U | G, AY | No | 8000 F | 14 | 11 | A | B | L |
| HOHA (Laptev et al., 2008) | IH,HH,HO | F | Dynamic | C | F | Some | V | U | SD,C | AY,E,G | Yes | 475 V | UM | 8 | A | B | CA |
| KTH (Schuldt et al., 2004) | IH | F | Dynamic | G | F | No | SV | C | ST,U | G | Yes | 2391 A | 25 | 6 | A | I | CA |
| MSRC-12 (Fothergill et al., 2012) | IH | F | Dynamic | ST | F | No | F | C | N/A | A,E,G | No | 719359 F | 30 | 12 | E,I | I | F |
| G3D (Bloom et al., 2012) | IH | F | Dynamic | C,D,ST | F | No | Fixed | C | ST,C | U | No | 80000 F | 10 | 20 | A | I | F |
| ASLLVD (Neidle et al., 2012) | IH | H | Dynamic | MC | U | No | F | C | ST,U | G | Yes | 9794 G | 6 | 3314 | S | I | RRC |
| UTA ASL (Conly et al., 2013) | IH | H | Dynamic | C,D | U | No | F | C | ST,U | E,A | Yes | 1313 G | 2 | 1113 | S | I | RCC |
| ChAirGest (Ruffieux et al., 2013) | IH | H | Dynamic | ChAir | U | No | F | C | ST,C | U | No | 1200 G | 10 | 10 | I,E | I | F,ATSR |
| SKIG (Liu and Shao, 2013) | IH | H | Dynamic | C,D | A | No | F | C | ST,CU | U | Yes | 1080 G | 10 | 6 | I,E | I | CA |
| 6DMG (Chen et al., 2012) | IH | H | Dynamic | 6DMG | H | No | F | C | N/A | G,AU,EU | Yes | 5600 G | 28 | 20 | I | I | CA |
| MSRGesture3D (Kurakin et al., 2012) | IH | H | Dynamic | B | H | No | Fixed | C | N/A | U | Yes | 336 G | 10 | 12 | S | I | CA |
| NATOPS (Song et al., 2011b) | IH | U | Dynamic | S,D,B | U | No | Fixed | C | ST,U | U | Yes | 9600 G | 20 | 24 | E | I | CA |
| NTU Dataset (Ren et al., 2011a) | IH | H | Static | C,D | U | No | F | C | SF,C | G | SF | 1000 G | 10 | 10 | HS | I | CA |
| Keck Gesture (Lin et al., 2009) | IH | H | Dynamic | C | F | No | F | C | ST,U | No | Yes | 294 G | 3 | 14 | E | I | CA |
| Cambridge Gesture (Kim et al., 2007) | IH | H | Dynamic | C | H | No | F | C | ST,U | U | Yes | 900 G | 2 | 9 | E | I | CA |
| (Triesch and von der Malsburg, 2001) | IH | H | Static | G | H | No | Fixed | C | SF,CU | U | SF | 717 G | 24 | 10 | HS | I | CA |

Table 1: Datasets. “Occl.” stands for “occlusions”. “View. variab.” stands for “viewpoint variabilities”. “Variab.” stands for “variabilities in gender, age, ethnicity”. “Seg.” stands for “segmented”. “Indiv/Collab.” stands for “individual or collaborative gestures”. “Type” stands for “gesture type”. “Eval.” stands for “evaluation”.

| Taxonomy Attribute | Acronym/Abbreviation | Meaning |
|---|---|---|
| Actors/Objects | HH | human-human interactions |
| Actors/Objects | IH | isolated human |
| Actors/Objects | O | humans with objects |
| Body parts | F | full body |
| Body parts | H | hands |
| Static/Dynamic | SF | Subjects are in motion, but each frame is individually classified |
| Modalities | 6DMG | WorldViz PPT-X4 (position + 3D orientation) + Wii Remote Plus (acceleration and angular speeds) |
| Modalities | A | audio |
| Modalities | B | binary segmentation mask |
| Modalities | C | RGB (color) |
| Modalities | ChAir | RGB, depth, skeletal, four inertial motion units |
| Modalities | D | depth |
| Modalities | G | grayscale |
| Modalities | MC | multiple cameras |
| Modalities | S | stereo images |
| Modalities | ST | skeletal tracking |
| Viewing field | A | arm and hand |
| Viewing field | E | egocentric |
| Viewing field | F | full body |
| Viewing field | MU | upper body in most cases |
| Viewing field | U | upper body |
| Occlusions | | |
| Viewpoints | F | Frontal |
| Viewpoints | Fixed | Fixed viewpoint for each class |
| Viewpoints | SV | Fixed viewpoint for some classes, variable for other classes |
| Viewpoints | V | Variable viewpoint |
| Controlled/Uncontrolled | C | Controlled |
| Controlled/Uncontrolled | U | Uncontrolled |
| Background | CU | some cluttered, some uncluttered |
| Background | C | cluttered |
| Background | D | dynamic |
| Background | SF | each frame is individually classified, so background is seen only from a single frame |
| Background | SD | static in some cases, dynamic in some cases |
| Background | ST | static |
| Background | U | uncluttered |
| Variabilities in gender/age/ethnicity | A | variabilities in age |
| Variabilities in gender/age/ethnicity | AU | unspecified whether there are variabilities in age |
| Variabilities in gender/age/ethnicity | AY | mostly non-senior adults |
| Variabilities in gender/age/ethnicity | E | variabilities in ethnicity |
| Variabilities in gender/age/ethnicity | EU | unspecified whether there are variabilities in ethnicity |
| Variabilities in gender/age/ethnicity | G | variabilities in gender |
| Variabilities in gender/age/ethnicity | U | unspecified |
| Segmented/Unsegmented | SF | each frame is individually classified |
| Amount of data | A | action samples |
| Amount of data | F | frames |
| Amount of data | G | gesture samples |
| Amount of data | V | video clips |
| Number of subjects | U1 | Unspecified, but most subjects appear in only one sample |
| Number of subjects | UM | Unspecified, but most subjects appear in several samples |
| Number of subjects | US | Unspecified, but some subjects appear in more than one sample |
| Classes | CBP | continuous space of body |
| Classes | CDG11 | about 300, but broken into subsets of 8-12 classes |
| Gesture type | A | action |
| Gesture type | BP | body pose |
| Gesture type | D | deictic |
| Gesture type | E | emblematic |
| Gesture type | HS | handshape |
| Gesture type | I | iconic |
| Gesture type | S | sign |
| Individual or collaborative | B | both individual and collaborative |
| Individual or collaborative | C | collaborative |
| Individual or collaborative | I | individual |
| Individual or collaborative | MI | mostly individual |
| Evaluation criteria | ATSR | Defined in (Ruffieux et al., 2013), based on difference between detected and ground truth endpoints, normalized by duration of the gesture |
| Evaluation criteria | CA | Classification accuracy |
| Evaluation criteria | F | F-score |
| Evaluation criteria | Fer. | Defined in (Ferrari et al., 2008), checks if detected endpoints are within distance of half length (of the body part in question) from the ground truth position. |
| Evaluation criteria | L | Levenshtein distance |
| Evaluation criteria | MPJPE | Mean per-joint position error (measured as Euclidean distance). |
| Evaluation criteria | Pascal | At least 50% overlap of bounding boxes on all body parts. |
| Evaluation criteria | Ram. | Defined in (Ramanan, 2006), average negative log likelihood of correct pose. |
| Evaluation criteria | RCC | Defined in (Wang et al., 2010), based on rank of the correct class for each test sign. For any R, report percentage of test signs for which the correct class was in the top R classes. |
| Evaluation criteria | Sap. | Defined in (Sapp and Taskar, 2013), Accuracy is based on (variable) threshold pixel distance between joint location and ground truth, scaled so that the torso length in the ground truth is 100 pixels. |

Table 2: Acronyms and abbreviations used in the table of datasets.

## 3. Gesture recognition challenges

In this section we review the series of gesture and action recognition challenges organized by ChaLearn from 2011 to 2015, as well as other international challenges related to gesture recognition.

### 3.1 First ChaLearn Gesture Recognition Challenge (2011-2012): One shot learning

ChaLearn launched in 2012 a challenge with prizes donated by Microsoft using datasets described in (Guyon et al., 2014). We organized two rounds in conjunction with the CVPR conference (Providence, Rhode Island, USA, June 2012) and the ICPR conference (Tsukuba, Japan, November 2012). Details on the challenge setting and results are found in (Guyon et al., 2013). We briefly summarize the setting and results.

#### 3.1.1 2011-2012 CHALLENGE PROTOCOL AND EVALUATION

The task of the challenge was to built a learning system capable of learning a gesture classification problem from a **single training example** per class, from dynamic video data complemented by a depth map obtained with Kinect$^{TM}$. The rationale behind this setting is that, in many computer interface applications to gesture recognition, users want to customize the interface to use their own gestures. Therefore they should be able to retrain the interface using a small vocabulary of their own gestures. We have also experimented with other use cases in gaming and teaching gesture vocabularies. Additionally, the problem of one-shot-learning is of intrinsic interest in machine learning and the solutions deviced could carry over to other applications. It is in a certain way an extreme case of transfer learning.

To implement this setting in the challenge, we collected a large dataset consisting of batches, each batch corresponding to the video recording of short sequences of gestures performed by the same person. The gestures in one batch pertained to a small vocabulary of gestures taken from a variety of application domains (sign language for the deaf, traffic signals, pantomimes, dance postures, etc.). During the development phase, the participants had access to hundreds of batches of diverse gesture vocabularies. This played the role of "source domain data" in the transfer learning task. The goal of the participants was to get ready to receive new batches from different gesture performers and different gesture vocabularies, playing the role of "transfer domain data". Their system would then need to learn from a single example of gesture performed by the particular performer, before being capable of recognizing the rest of the gestures in that batch. The full dataset is available from `http://gesture.chalearn.org/data`.

More specifically, each batch was split into a training set (of one example for each gesture) and a test set of short sequences of one to 5 gestures. Each batch contained gestures from a different small vocabulary of 8 to 12 gestures, for instance diving signals, signs of American Sign Language representing small animals, Italian gestures, etc. The test data labels were provided for the development data only (source domain data), so the participants could self-evaluate their systems and pre-train parts of it as is expected from transfer learning methods. The data also included 20 validation batches and 20 final evaluation batches as transfer domain data used by the organizers to evaluate the participants. In those batches, only the labels of the training gestures (one example each) was provided, the rest of the gesture sequences were unlabelled and the goal of the participants

was to predict those labels. We used the Kaggle platform to manage submissions[1] The participants received immediate feed-back on validation data on a on-line leaderboard. The final evaluation was carried out on the final evaluation data, and those results were only revealed after the challenge was over. The participants had a few days to train their systems and upload their predictions. Prior to the end of the development phase, the participants were invited to submit executable software for their best learning system to a software vault. This allowed the competition organizers to check their results and ensure the fairness of the competition.

To compare prediction labels for gesture sequences to the truth values, we used the generalized Levenshtein distances (each gesture counting as one token). The final evaluation score was computed as the sum of such distances for all test sequences, divided by the total number of gestures in the test batch. This score is analogous to an error rate. However, it can exceed one. Specifically, for each video, the participants provided an ordered list of labels $R$ corresponding to the recognized gestures. We compared this list to the corresponding list of labels T in the prescribed list of gestures that the user had to play. These are the "true" gesture labels (provided that the users did not make mistakes). We computed the generalized Levenshtein distance $L(R,T)$, that is the minimum number of edit operations (substitution, insertion, or deletion) that one has to perform to go from $R$ to $T$ (or vice versa). The Levenhstein distance is also known as "edit distance". For example: $L([124],[32]) = 2; L([1],[2]) = 1; L([222],[2]) = 2$.

We provided code to browse though the data, a library of computer vision and machine learning techniques written in Matlab featuring examples drawn from the challenge datasets, and an end-to-end baseline system capable of processing challenge data and producing a sample submission. The competition pushed the state of the art considerably. The participants narrowed down the gap in performance between the baseline recognition system initially provided ($\simeq 60\%$ error) and human performance ($\simeq 2\%$ error) by reaching $\simeq 7\%$ error in the second round of the challenge. There remains still much room for improvement, particularly to recognize static postures and subtle finger positions.

### 3.1.2 2011-2012 CHALLENGE DATA

The datasets are described in details in a companion paper (Guyon et al., 2014). Briefly, the data are organized in batches: development batches devel01-480, validation batches valid01-20, and final evaluation batches final01-20 (for round 1) and final21-40 (for round 2). For the development batches, we provided all the labels. To evaluate the performances on "one-shot-learning" tasks, the valid and final batches were provided with labels only for **one example of each gesture class** in each batch (training examples). The goal was to automatically predict the gesture labels for the remaining unlabelled gesture sequences (test examples).

Each batch includes 100 recorded gestures grouped in sequences of 1 to 5 gestures performed by the same user. The gestures are drawn from a small vocabulary of 8 to 12 unique gestures, which we call a "lexicon". For instance a gesture vocabulary may consist of the signs to referee volleyball games or the signs to represent small animals in the sign language for the deaf. We selected lexicons from nine categories corresponding to various settings or application domains (Figure 5):

1. **Body language** gestures (like scratching your head, crossing your arms).

1. For round 1: `http://www.kaggle.com/c/GestureChallenge`. For round 2: `http://www.kaggle.com/c/GestureChallenge2`.

| Team | Public score on validation set | Private score on final set #1 | For comparison score on final set #2 |
|---|---|---|---|
| Alfnie | 0.1426 | 0.0996 | 0.0915 |
| Pennect | 0.1797 | 0.1652 | 0.1231 |
| OneMillionMonkeys | 0.2697 | 0.1685 | 0.1819 |
| Immortals | 0.2543 | 0.1846 | 0.1853 |
| Zonga | 0.2714 | 0.2303 | 0.2190 |
| Balazs Godeny | 0.2637 | 0.2314 | 0.2679 |
| SkyNet | 0.2825 | 0.2330 | 0.1841 |
| XiaoZhuwWudi | 0.2930 | 0.2564 | 0.2607 |
| Baseline method 1 | 0.5976 | 0.6251 | 0.5646 |

Table 3: Results of round 1. In round 1 the baseline method was a simple template matching method (see text). For comparison, we show the results on the final set number 2 not available in round 1.

2. **Gesticulations** performed to accompany speech.
3. **Illustrators** (like Italian gestures).
4. **Emblems** (like Indian Mudras).
5. **Signs** (from sign languages for the deaf).
6. **Signals** (like referee signals, diving signals, or Marshalling signals to guide machinery or vehicle).
7. **Actions** (like drinking or writing).
8. **Pantomimes** (gestures made to mimic actions).
9. **Dance postures.**

During the challenge, we did not disclose the identity of the lexicons and of the users.

#### 3.1.3 2011-2012 CHALLENGE RESULTS

The results of the top ranking participants were checked by the organizers who reproduced their results using the code provided by the participants before they had access to the final evaluation data. All of them passed successfully the verification process. These results are shown in Tables 3 and 4.

#### 3.1.4 2011-2012 CHALLENGE, SUMMARY OF THE WINNER METHODS

The results of the challenge are analyzed in details, based on papers published in this special topic and on descriptions provided by the top ranking participants in their fact sheets (Guyon et al., 2013). We briefly summarize notable methods below.

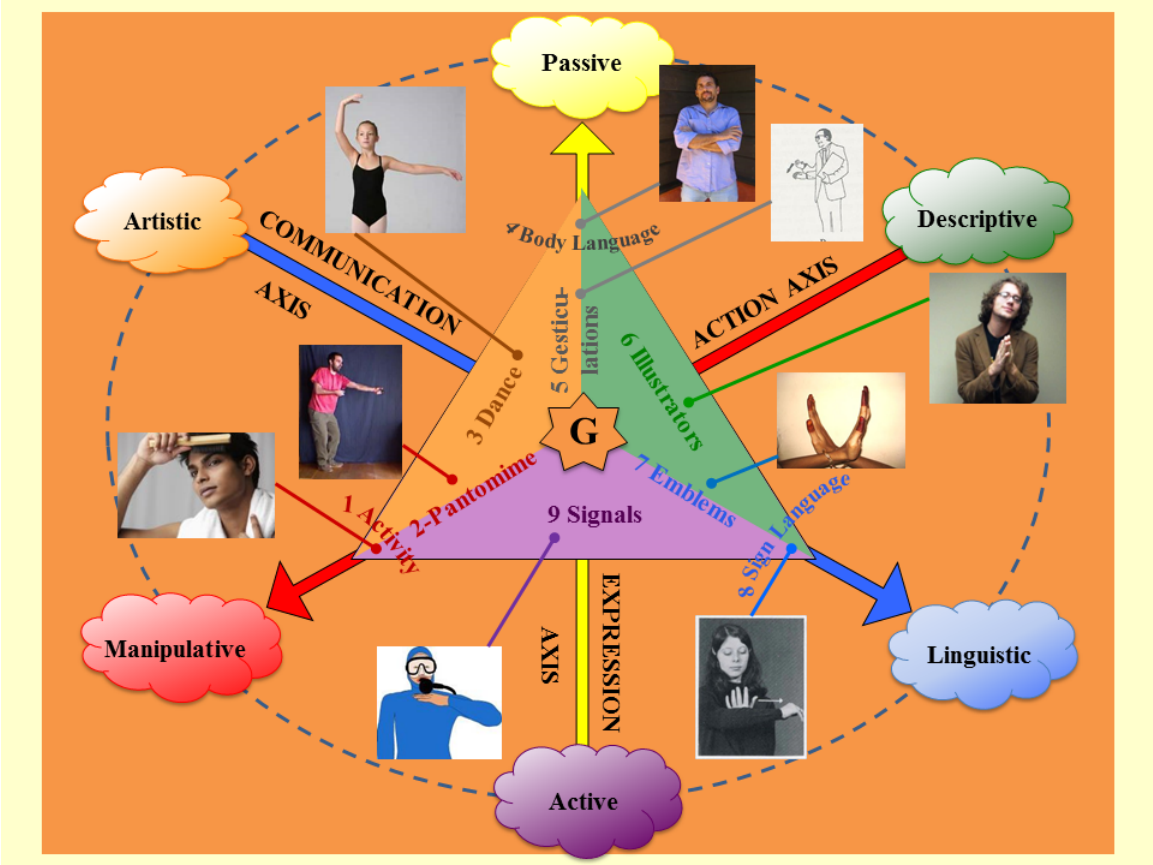



Figure 5: **Types of gestures.** We created a classification of gesture types according to purpose defined by three complementary axes: communication, expression and action. We selected 85 gesture vocabularies, including Italian gestures, Indian Mudras, Sign language for the deaf, diving signals, pantomimes, and body language.

**The winner of both rounds** (Alfonso Nieto Castañon of Spain, a.k.a. alfnie) used a novel technique called "Motion Signature analyses", inspired by the neural mechanisms underlying information processing in the visual system. This is an unpublished method using a sliding window to perform simultaneously recognition and temporal segmentation, based solely on depth images. The method, described by the authors as a "Bayesian network", is similar to a Hidden Markov Model (HMM). It performs simultaneous recognition and segmentation using the Viterbi algorithm. The preprocessing steps include Wavelet filtering replacement of missing values and outlier detection. Notably, this method is one of the fastest despite the fact that he implemented it in Matlab (close to real time on a regular laptop). The author claims that it has linear complexity in image size, number of frames, and number of training examples.

**The second best ranked participants** (team Pennect of Universit of Pennsylvania, USA, in round 1 and team Turtle Tamers of Slovakia, in round 2) used very similar methods and performed similarly. The second team published their results in this special topic (Konecny and Hagara, 2014).

| Team | Public score on validation set | For comparison score on final set #1 | Private score on final set #2 |
|---|---|---|---|
| Alfnie | 0.0951 | 0.0734 | 0.0710 |
| Turtle Tamers | 0.2001 | 0.1702 | 0.1098 |
| Joewan | 0.1669 | 0.1680 | 0.1448 |
| Wayne Zhang | 0.2814 | 0.2303 | 0.1846 |
| Manavender | 0.2310 | 0.2163 | 0.1608 |
| HIT CS | 0.1763 | 0.2825 | 0.2008 |
| Vigilant | 0.3090 | 0.2809 | 0.2235 |
| Baseline method 2 | 0.3814 | 0.2997 | 0.3172 |

Table 4: Results of round 2. In round 2, the baseline method was the "Principal Motion" method (see text).

Both methods are based on an HMM-style model using HOG/HOF features to represent movie frames. They differ in that Pennect used RGB images only while Turtle Tamers used both RGB and depth. Another difference is that Pennect used HOG/HOF features at 3 different scales while Turtle Tamers created a bag of features using K-means clustering from only 40x40 resolution and 16 orientation bins. Pennect trained a one-vs-all linear classifier for each frame in every model and used the discriminant value as a local state score for the HMM while Turtle Tamers used a quadratic-chi kernel metric for comparing pairs of frames in the training and test movie. As preprocessing, Pennect uses mean subtraction and compensates for body translations while Turtle Tamers replaces the missing values by the median of neighboring values. Both teams claim a linear complexity in number of frames, number of training examples, and image size. They both provided Matlab software that processes all the batches of the final test set on a regular laptop in a few hours.

The next best ranked participants (who won **third place in round 2**), the Joewan team, who published in this special topic (Wan et al., 2013), used a slightly different approach. They relied on the motion segmentation method provided by the organizers to pre-segment videos. They then represented each video as a bag of 3D MOSIFT features (integrating RGB and depth data), and then used a nearest neighbor classifier. Their algorithm is super-quadratic in image size, linear in number of frames per video, and linear in number of training examples. The method is rather slow and takes over a day to process all the batches of the final test set on a regular laptop.

**The third best ranked team** in round 1 (OneMillionMonkeys) also used an HMM in which a state is created for each frame of the gesture exemplars. This data representation is based on edge detection in each frame. Edges are associated with several attributes including the X/Y coordinates, their orientation, their sharpness, their depth and location in an area of change. To provide a local state score to the HMM for test frames, OneMillionMonkeys calculated the joint probability of all the nearest neighbors in training frames using a Gaussian model. The system works exclusively from the depth images. The system is one of the slowest proposed. Its processing speed is linear in number of training examples but quadratic in image size and number of frames per video. The method is rather slow and takes over a day to process all the batches of the final test set on a regular laptop.

Methods robust against translation include those of Joewan (Wan et al., 2013) and Immortals/Manavender (this is the same author under two different pseudonyms for round 1 and round 2). The team Immortals/Manavender published their method in this special topic (Malgireddy et al., 2013). Their representations are based on a bag of visual words, inspired by techniques used in action recognition (Laptev, 2005). Such representations are inherently shift invariant. The slight performance loss in translated data may be due to partial occlusions.

Although the team Zonga did not end up ranking among top ranking participants, the authors, who published their method in this special topic, proposed a very original method and ended up winning the best paper award. Notably, their outperformed all baseline methods early on in the challenge by applying their method without tuning it to the tasks of the challenge and remained at the top of the leaderboard for several weeks. The used a novel technique based on tensor geometry, which provides a data representation exhibiting desirable invariances and yields a very discriminating structure for action recognition.

ChaLearn also organized demonstration competitions of gesture recognition systems using Kinect$^{TM}$, in conjunction with those events. Novel data representations were proposed to tackle with success, in real time, the problem of hand and finger posture recognition. The demonstration competition winners showed systems capable of accurately tracking in real time hand postures in application for touch free exploration of 3D medical images for surgeons in the operating room, finger spelling (sign language for the deaf), virtual shopping, and game controlling. Combining the methods proposed in the demonstration competition tackling the problem of hand postures and those of the quantitative evaluation focusing on the dynamics of hand and arm movements is a promising direction of future research. For a long lasting impact, the challenge platform, the data and software repositories have been made available for further research[2].

### 3.2 ChaLearn Multimodal Gesture Recognition Challenge 2013

The focus of this second challenge was on *multiple instance, user independent learning of gestures from multimodal data*, which means learning to recognize gestures from several instances for each category performed by different users, drawn from a vocabulary of 20 gesture categories(Escalera et al., 2013b,a). A gesture vocabulary is a set of unique gestures, generally related to a particular task. In this challenge we focus on the recognition of a vocabulary of 20 Italian cultural/anthropological signs (Escalera et al., 2013b), see Figure 6 for one example of each Italian gesture.

#### 3.2.1 2013 CHALLENGE DATA

In all the sequences, a single user is recorded in front of a Kinect$^{TM}$, performing natural communicative gestures and speaking in fluent Italian. The main characteristics of the dataset of gestures are:

• 13.858 gesture samples recorded with the Kinect$^{TM}$camera, including audio, skeletal model, user mask, RGB, and depth images.

• RGB video stream, 8-bit VGA resolution (640×480) with a Bayer color filter, and depth sensing video stream in VGA resolution (640×480) with 11-bit. Both are acquired in 20 fps on average.

• Audio data is captured using Kinect$^{TM}$20 multiarray microphone.

• A total number of 27 users appear in the data set.

2. http://gesture.chalearn.org/

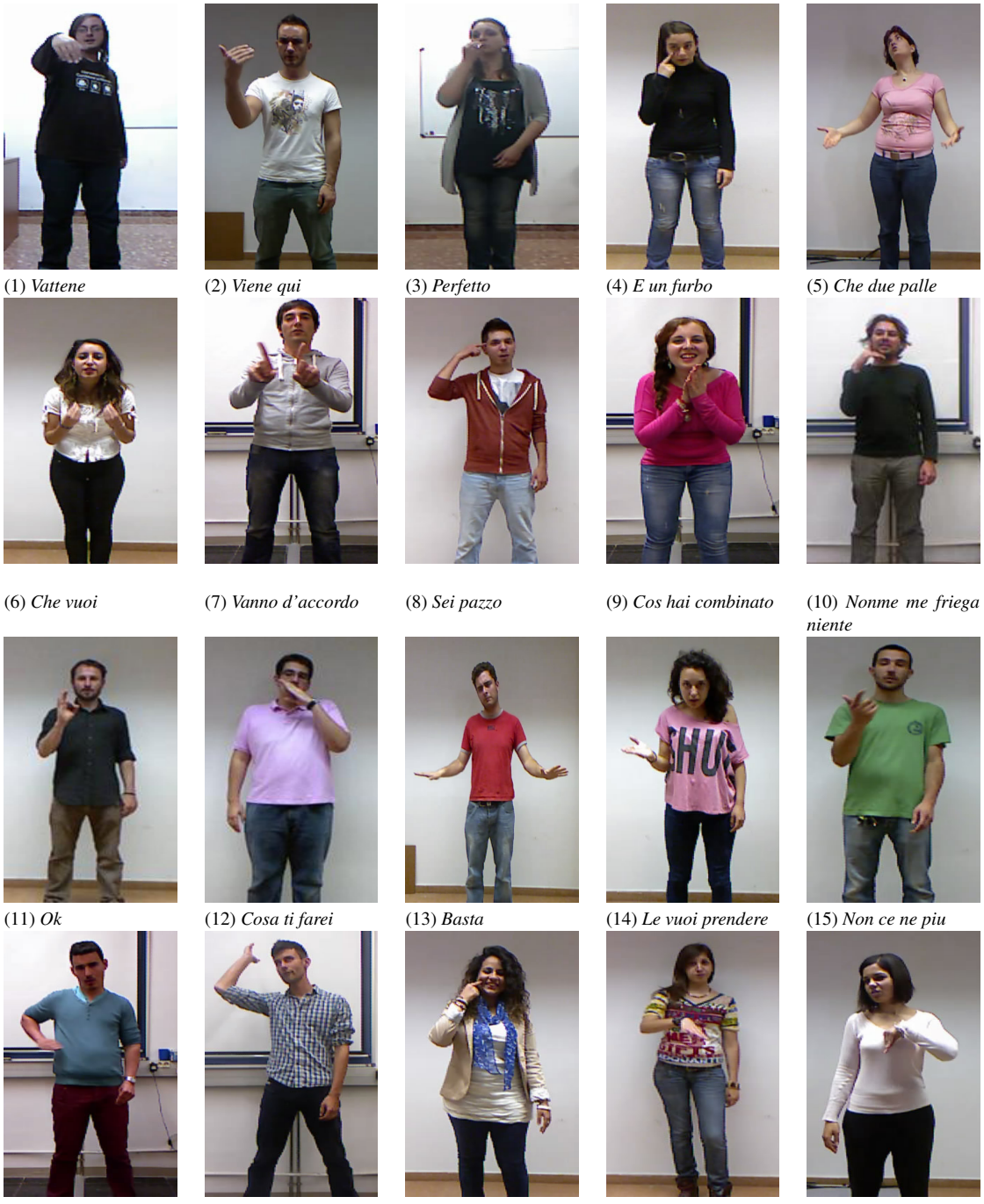



Figure 6: Data set gesture categories.

• The data set contains the following number of sequences, development: 393 (7.754 gestures), validation: 287 (3.362 gestures), and test: 276 (2.742 gestures), each sequence lasts between 1 and 2 minutes and contains between 8 and 20 gesture samples, around 1.800 frames. The total number

Table 5: Easy and challenging aspects of the data.

| **Easy** |
|---|
| Fixed camera |
| Near frontal view acquisition |
| Within a sequence the same user |
| Gestures performed mostly by arms and hands |
| Camera framing upper body |
| Several available modalities: audio, skeletal model, user mask, depth, and RGB |
| Several instances of each gesture for training |
| Single person present in the visual field |
| **Challenging** |
| *Within each sequence*: |
| Continuous gestures without a resting pose |
| Many gesture instances are present |
| Distracter gestures out of the vocabulary may be present in terms of both gesture and audio |
| *Between sequences*: |
| High inter and intra-class variabilities of gestures in terms of both gesture and audio |
| Variations in background, clothing, skin color, lighting, temperature, resolution |
| Some parts of the body may be occluded |
| Different Italian dialects |

of frames of the data set is 1.720.800.

• All the gesture samples belonging to 20 main gesture categories from an Italian gesture dictionary are annotated at frame level indicating the gesture label.

• 81% of the participants were Italian native speakers, while the remaining 19% of the users were not Italian, but Italian-speakers.

• All the audio that appears in the data is from the Italian dictionary. In addition, sequences may contain distractor words and gestures, which are not annotated since they do not belong to the main dictionary of 20 gestures.

This dataset, available at `http://sunai.uoc.edu/chalearn`, presents various features of interest as listed in Table 5. Examples of the provided visual modalities are shown in Figure 7.

RGB Depth User mask Skeletal model

Figure 7: Different data modalities of the provided data set.

### 3.2.2 2013 CHALLENGE PROTOCOL AND EVALUATION

As in our previous 2011-2012 challenge, it consisted of two main components: a development phase (April 30th to Aug 1st) and a final evaluation phase (Aug 2nd to Aug 15th). The submission and

evaluation of the challenge entries was via the *Kaggle* platform [3]. The official participation rules were provided on the website of the challenge. In addition, publicity and news on the ChaLearn Multimodal Gesture Recognition Challenge were published in well-known online platforms, such as LinkedIn, Facebook, Google Groups and the ChaLearn website.

During the development phase, the participants were asked to build a system capable of learning from several gesture samples a vocabulary of 20 Italian sign gesture categories. To that end, the teams received the development data to train and self-evaluate their systems. In order to monitor their progress they could use the validation data for which the labels were not provided. The prediction results on validation data could be submitted online to get immediate feed-back. A real-time leaderboard showed to the participants their current standing based on their validation set predictions.

During the final phase, labels for validation data were published and the participants performed similar tasks as those performed in previous phase, using the validation data and training data sets in order to train their system with more gesture instances. The participants had only few days to train their systems and upload them. The organizers used the final evaluation data in order to generate the predictions and obtain the final score and rank for each team. At the end, the final evaluation data was revealed, and authors submitted their own predictions and fact sheets to the platform.

As an evaluation metric we also used the Levenshtein distance described in previous section. A public score appeared on the leaderboard during the development period and was based on the validation data. Subsequently, a private score for each team was computed on the final evaluation data released at the end of the development period, which was not revealed until the challenge was over. The private score was used to rank the participants and determine the prizes.

#### 3.2.3 2013 CHALLENGE RESULTS

The challenge attracted high level of participation, with a total of 54 teams and near 300 total number of entries. This is a good level of participation for a computer vision challenge requiring very specialized skills. Finally, 17 teams successfully submitted their prediction in final test set, while providing also their code for verification and summarizing their method by means of a fact sheet questionnaire.

After verifying the codes and results of the participants, the final scores of the top rank participants on both validation and test sets were made public: these results are shown in Table 6, where winner results on the final test set are printed in bold. In the end, the final error rate on the test data set was around 12%.

#### 3.2.4 2013 CHALLENGE SUMMARY OF THE WINNER METHODS

Table 7 shows the summary of the strategies considered by each of the top ranked participants on the test set. Interestingly, the three top ranked participants agree in the modalities and segmentation strategy considered, although they differ in the final applied classifier. Next, we briefly describe in more detail the approach designed by the three winners of the challenge.

**The first ranked team** *IVAMM* on the test set used a feature vector based on audio and skeletal information, and applied late fusion to obtain final recognition results. A simple time-domain end-point detection algorithm based on joint coordinates is applied to segment continuous data sequences into candidate gesture intervals. A Gaussian Hidden Markov Model is trained with 39-

3. https://www.kaggle.com/c/multimodal-gesture-recognition

Table 6: Top rank results on validation and test sets.

| TEAM | Validation score | Test score |
|---|---|---|
| IVA MM | 0.20137 | **0.12756** |
| WWEIGHT | 0.46163 | **0.15387** |
| ET | 0.33611 | **0.16813** |
| MmM | 0.25996 | 0.17215 |
| PPTK | 0.15199 | 0.17325 |
| LRS | 0.18114 | 0.17727 |
| MMDL | 0.43992 | 0.24452 |
| TELEPOINTS | 0.48543 | 0.25841 |
| CSI MM | 0.32124 | 0.28911 |
| SUMO | 0.49137 | 0.31652 |
| GURU | 0.51844 | 0.37281 |
| AURINKO | 0.31529 | 0.63304 |
| STEVENWUDI | 1.43427 | 0.74415 |
| JACKSPARROW | 0.86050 | 0.79313 |
| JOEWAN | 0.13653 | 0.83772 |
| MILAN KOVAC | 0.87835 | 0.87463 |
| IAMKHADER | 0.93397 | 0.92069 |

Table 7: Team methods and results. Early and late refer to early and late fusion of features/classifier outputs. HMM: Hidden Markov Models. KNN: Nearest Neighbor. RF: Random Forest. Tree: Decision Trees. ADA: Adaboost variants. SVM: Support Vector Machines. Fisher: Fisher Linear Discriminant Analysis. GMM: Gaussian Mixture Models. NN: Neural Networks. DGM: Deep Boltzmann Machines. LR: Logistic Regression. DP: Dynamic Programming. ELM: Extreme Learning Machines.

| TEAM | Test score | Rank | Modalities | Segmentation | Fusion | Classifier |
|---|---|---|---|---|---|---|
| IVA MM | 0.12756 | 1 | Audio,Skeleton | Audio | None | HMM,DP,KNN |
| WWEIGHT | 0.15387 | 2 | Audio,Skeleton | Audio | Late | RF,KNN |
| ET | 0.16813 | 3 | Audio,Skeleton | Audio | Late | Tree,RF,ADA |
| MmM | 0.17215 | 4 | Audio,RGB+Depth | Audio | Late | SVM,Fisher,GMM,KNN |
| PPTK | 0.17325 | 5 | Skeleton,RGB,Depth | Sliding windows | Late | GMM,HMM |
| LRS | 0.17727 | 6 | Audio,Skeleton,Depth | Sliding windows | Early | NN |
| MMDL | 0.24452 | 7 | Audio,Skeleton | Sliding windows | Late | DGM+LR |
| TELEPOINTS | 0.25841 | 8 | Audio,Skeleton,RGB | Audio,Skeleton | Late | HMM,SVM |
| CSI MM | 0.28911 | 9 | Audio,Skeleton | Audio | Early | HMM |
| SUMO | 0.31652 | 10 | Skeleton | Sliding windows | None | RF |
| GURU | 0.37281 | 11 | Audio,Skeleton,Depth | DP | Late | DP,RF,HMM |
| AURINKO | 0.63304 | 12 | Skeleton,RGB | Skeleton | Late | ELM |
| STEVENWUDI | 0.74415 | 13 | Audio,Skeleton | Sliding windows | Early | DNN,HMM |
| JACKSPARROW | 0.79313 | 14 | Skeleton | Sliding windows | None | NN |
| JOEWAN | 0.83772 | 15 | Skeleton | Sliding windows | None | KNN |
| MILAN KOVAC | 0.87463 | 16 | Skeleton | Sliding windows | None | NN |
| IAMKHADER | 0.92069 | 17 | Depth | Sliding windows | None | RF |

dimension MFCC features and generates confidence scores for each gesture category. A Dynamic Time Warping based skeletal feature classifier is applied to provide complementary information. The confidence scores generated by the two classifiers are firstly normalized and then combined to produce a weighted sum. A single threshold approach is employed to classify meaningful gesture intervals from meaningless intervals caused by false detection of speech intervals.

**The second ranked team** *WWEIGHT* combined audio and skeletal information, using both joint spatial distribution and joint orientation. The method first searches for regions of time with high audio-energy to define 1.8-second-long windows of time that potentially contained a gesture. This had the effect that the development, validation, and test data were treated uniformly. Feature

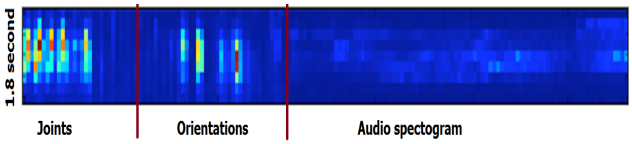



Figure 8: ExtraTreesClassifier Feature Importance.


Figure 9: Recognition of test sequence by the three challenge winners. Black bin means that the complete list of ordered gestures has been successfully recognized.

vectors are then defined using a log-spaced audio spectrogram and the joint positions and orientations above the hips. At each time sample the method subtracts the average 3D position of the left and right shoulders from each 3D joint position. Data is down-sampled onto a 5 Hz grid considering 1.8 seconds. There were 1593 features total (9 time samples $\times$ 177 features per time sample). Since some of the detected windows can contain distractor gestures, an extra 21st label is introduced, defining the 'not in the dictionary' gesture category. Python's scikit-learn was used to train two models: an ensemble of randomized decision trees (ExtraTreesClassifier, 100 trees, 40% of features) and a K-Nearest Neighbor model (7 neighbors, L1 distance). The posteriors from these models are averaged with equal weight. Finally, a heuristic is used (12 gestures maximum, no repeats) to convert posteriors to a prediction for the sequence of gestures.

Figure 8 shows the mean feature importance for the windows size of 1.8 seconds for the three sets of features: joint coordinates, joint orientations, and audio spectogram. One can note that features from the three sets are selected as discriminative by the classifier, although skeletal features becomes more useful for the ExtraTreesClassifier. Additionally, the most discriminative features are those in the middle of the windows size, since begin-end features are shared among different gestures (transitions) and thus are less discriminative for the classifier.

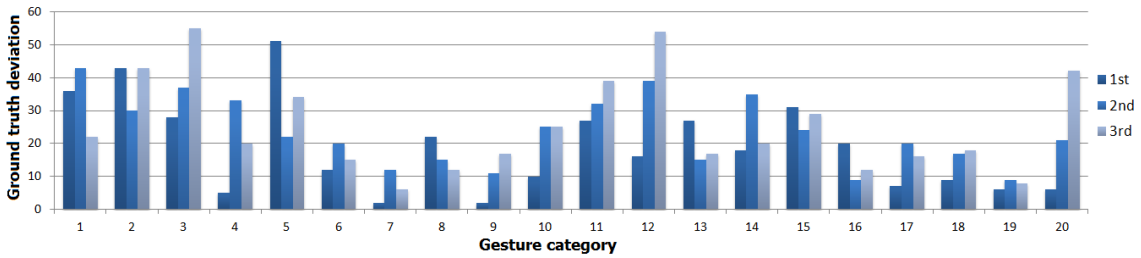



Figure 10: Deviation of the number of gesture samples for each category by the three winners in relation to the GT data.

**The third ranked team** *ET* combined the output decisions of two designed approaches. In the first approach, they look for gesture intervals (unsupervised) using the audio files and extract these features from intervals (MFCC). Using these features, authors train a random forest and gradient boosting classifier. The second approch uses simple statistics (median, var, min, max) on the first 40

frames for each gesture to build the training samples. The prediction phase uses a sliding window. The authors create a weighted average of the output of these two models. The features considered were skeleton information and audio signal.

Finally, we extracted some statistics from the results of the three challenge winners in order to analyze common points and difficult aspects of the challenge. Figure 9 shows the recognition of the 276 test sequences by the winners. Black bin means that the complete list of ordered gestures was successfully recognized for those sequences. Once can see that there exists some kind of correlation among methods. Taking into account that consecutive sequences belong to the same user performing gestures, it means that some some gestures are easier to recognize than others. Since different users appear in the training and test sequences, it is sometimes difficult for the models to generalize to the style of new users, based on the gesture instances used for training.

We also investigated the difficulty of the problem by gesture category, within each of the 20 Italian gesture categories. Figure 10 shows for each winner method the deviation between the number of gesture instances recognized and the total number of gestures, for each category. This was computed for each sequence independently, and adding the deviation for all the sequences. In that case, a zero value means that the participant method recognized the same number of gesture instances for a category that was recorded in the ground truth data. Although we cannot guarantee with this measure that the order of recognized gesture matches with the ground truth, it gives us an idea of how difficult the gesture sequences were to segment into individual gestures. Additionally, the sum of total deviation for all the gestures for all the teams was 378, 469, and 504, which correlates with the final rank of the winners. The figure suggests a correlation between the performance of the three winners. In particular, categories 6, 7, 8, 9, 16, 17, 18, and 19 were the ones that achieved most accuracy for all the participants, meanwhile 1, 2, 3, 5, and 12 were the ones that introduced the highest recognition error. Note that the public data set provides accurate label annotations of end-begin of gestures, and thus, a more detailed recognition analysis could be performed applying a different recognition measurement to Leveinstein, such as Jaccard overlapping or sensitivity score estimation, which will also allow for confusion matrix estimation based on both inter and intra user and gesture category variability. This is left to future work.

### 3.3 ChaLearn Multimodal Gesture Spotting Challenge 2014

In ChaLearn LAP 2014 (Escalera et al., 2014) we focused on the user-independent automatic spotting of a vocabulary of 20 Italian cultural/anthropological signs in image sequences, see Figure 6.

#### 3.3.1 2014 GESTURE CHALLENGE DATA

This challenge was based on an Italian gesture data set, called *Montalbano gesture dataset*, an enhanced version of the ChaLearn 2013 multimodal gesture recognition challenge (Escalera et al., 2013b,a) with more ground-truth annotations. In all the sequences, a single user is recorded in front of a Kinect$^{TM}$, performing natural communicative gestures and speaking in fluent Italian. Examples of the different visual modalities are shown in Figure 7.

The main characteristics of the data set are:

• Largest data set in the literature, with a large duration of each individual performance showing no resting poses and self-occlusions.

• There is no information about the number of gestures to spot within each sequence, and several distracter gestures (out of the vocabulary) are present.

| Training seq. | Validation seq. | Test seq. | Sequence duration | FPS |
|---|---|---|---|---|
| 393 (7,754 gestures) | 287 (3,362 gestures) | 276 (2,742 gestures) | 1-2 min | 20 |
| Modalities | Num. of users | Gesture categories | Labeled sequences | Labeled frames |
| RGB, Depth, User mask, Skeleton | 27 | 20 | 13,858 | 1,720,800 |

Table 8: Main characteristics of the *Montalbano* gesture dataset.

• High intra-class variability of gesture samples and low inter-class variability for some gesture categories.

A list of data attributes for data set used in track 3 is described in Table 8.

### 3.3.2 2014 GESTURE CHALLENGE PROTOCOL AND EVALUATION

The challenge was managed using the Microsoft Codalab platform[4]. We followed a development (February 9 to May 20 2014) and tests phases (May 20th - June 1st 2014)) as in our previous challenges.

To evaluate the accuracy of action/interaction recognition, we use the Jaccard Index, For the $n$ action, interaction, and gesture categories labelled for a RGB/RGBD sequence $s$, the Jaccard Index is defined as:

$$J_{s,n} = \frac{A_{s,n} \bigcap B_{s,n}}{A_{s,n} \bigcup B_{s,n}}, \tag{1}$$

where $A_{s,n}$ is the ground truth of action/interaction/gesture $n$ at sequence $s$, and $B_{s,n}$ is the prediction for such an action at sequence $s$. $A_{s,n}$ and $B_{s,n}$ are binary vectors where 1-values correspond to frames in which the $n-$th action is being performed. The participants were evaluated based on the mean Jaccard Index among all categories for all sequences, where motion categories are independent but not mutually exclusive (in a certain frame more than one action, interaction, gesture class can be active).

In the case of false positives (e.g. inferring an action, interaction or gesture not labelled in the ground truth), the Jaccard Index is 0 for that particular prediction, and it will not count in the mean Jaccard Index computation. In other words $n$ is equal to the intersection of action/interaction/gesture categories appearing in the ground truth and in the predictions.

An example of the calculation for two actions is shown in Figure 11. Note that in the case of recognition, the ground truth annotations of different categories can overlap (appear at the same time within the sequence). Also, although different actors appear within the sequence at the same time, actions/interactions/gestures are labelled in the corresponding periods of time (that may overlap), there is no need to identify the actors in the scene. The example in Figure 11 shows the mean Jaccard Index calculation for different instances of actions categories in a sequence (single red lines denote ground truth annotations and double red lines denote predictions). In the top part of the image one can see the ground truth annotations for actions walk and fight at sequence $s$. In the center part of the image a prediction is evaluated obtaining a Jaccard Index of 0.72. In the bottom part of the image the same procedure is performed with the action fight and the obtained Jaccard Index is 0.46. Finally, the mean Jaccard Index is computed obtaining a value of 0.59.

4. https://www.codalab.org/competitions/

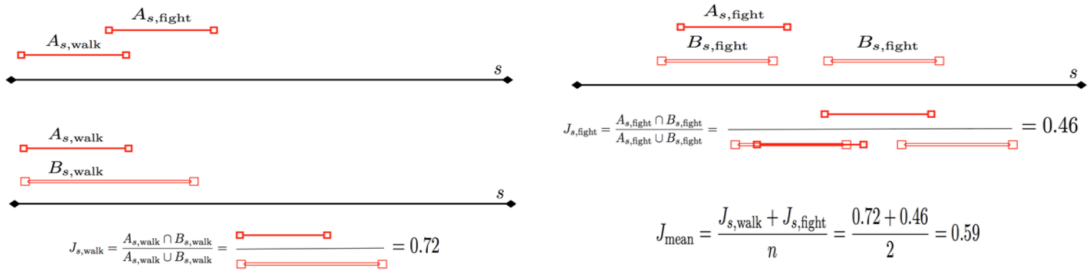



Figure 11: Example of mean Jacquard Index calculation for gesture and action/interaction spotting.

### 3.3.3 2014 GESTURE CHALLENGE RESULTS

Table 10 summarizes the methods of the 17 participants that contributed to the test set of track 3. Although DTW and HMM (and variants) were in the last edition of the ChaLearn Multimodal Gesture competition (Escalera et al., 2013b,a), random forest has been widely applied in this 2014 edition. Also, three participants used deep learning architectures.

| Team name | Accuracy | Rank | Features | Dimension reduction | Clustering | Classifier | Temporal coherence | Gesture representation |
|---|---|---|---|---|---|---|---|---|
| CUHK-SWJTU | **0.507173** | 1 | Improved trajectories (Wang and Schmid, 2013) | PCA | - | SVM | Sliding windows | Fisher Vector |
| ADSC | **0.501164** | 2 | Improved trajectories (Wang and Schmid, 2013) | - | - | SVM | Sliding windows | - |
| SBUVIS | **0.441405** | 3 | Improved trajectories (Wang and Schmid, 2013) | - | - | SVM | Sliding windows | - |
| DonkeyBurger | 0.342192 | 4 | MHI, STIP | - | Kmeans | Sparse code | Sliding windows | - |
| UC-T2 | 0.121565 | 5 | Improved trajectories (Wang and Schmid, 2013) | PCA | - | Kmeans | Sliding windows | Fisher Vector |
| MindLAB | 0.008383 | 6 | MBF | - | Kmeans | RF | Sliding windows | BoW |
| MMLAB | **0.5385** | 1 | IDT | PCA | - | SVM | - | Fisher Vector |
| FIKIE | **0.5239** | 2 | IDT | PCA | - | HMM | Appearance+Kalman filter | - |

Table 9: Top rows: Action/interaction 2014 recognition results. MHI: Motion History Image; STIP: Spatio-Temporal interest points; MBF: Multiscale Blob Features; BoW: Bag of Visual Words; RF: Random Forest. Bottom two rows: Action/interaction 2015 recognition results. IDT: Improved Dense Trajectories (Wang and Schmid, 2013).

| Team | Accuracy | Rank | Modalities | Features | Fusion | Temp. segmentation | Dimension reduction | Gesture representation | Classifier |
|---|---|---|---|---|---|---|---|---|---|
| LIRIS | **0.849987** | 1 | SK, Depth, RGB | RAW, SK joints | Early | Joints motion | - | - | DNN |
| CraSPN | **0.833904** | 2 | SK, Depth, RGB | HOG, SK | Early | Sliding windows | - | BoW | Adaboost |
| JY | **0.826799** | 3 | SK, RGB | SK, HOG | Late | MRF | PCA | - | MRF, KNN |
| CUHK-SWJTU | 0.791933 | 4 | RGB | Improved trajectories (Wang and Schmid, 2013) | - | Joints motion | PCA | Fisher Vector, VLAD | SVM |
| Lpigou | 0.788804 | 5 | Depth, RGB | RAW, SK joints | Early | Sliding windows | Max-pooling CNN | - | CNN |
| stevenwudi | 0.787310 | 6 | SK, depth | RAW | Late | Sliding windows | - | - | HMM, DNN |
| Ismar | 0.746632 | 7 | SK | SK | - | Sliding windows | - | - | RF |
| Quads | 0.745449 | 8 | SK | SK quads | - | Sliding windows | - | Fisher Vector | SVM |
| Telepoints | 0.688778 | 9 | SK, Depth, RGB | STIPS, SK | Late | Joints motion | - | - | SVM |
| TUM-fortiss | 0.648979 | 10 | SK, Depth, RGB | STIPS | Late | Joints motion | - | - | RF, SVM |
| CSU-SCM | 0.597177 | 11 | Skeleton, Depth, mask | HOG, Skeleton | Late | Sliding windows | - | 2DMTM | SVM, HMM |
| iva.mm | 0.556251 | 12 | Skeleton, RGB, depth | Skeleton, HOG | Late | Sliding windows | - | BoW | SVM, HMM |
| Terrier | 0.539025 | 13 | Skeleton | Skeleton | - | Sliding windows | - | - | RF |
| Team Netherlands | 0.430709 | 14 | Skeleton, Depth, RGB | MHI | Early | DTW | Preserving projections | - | SVM, RT |
| VecsRel | 0.408012 | 15 | Skeleton, Depth, RGB | RAW, skeleton joints | Late | DTW | - | - | DNN |
| Samgest | 0.391613 | 16 | Skeleton, Depth, RGB, mask | Skeleton, blobs, moments | Late | Sliding windows | - | - | HMM |
| YNL | 0.270600 | 17 | Skeleton | Skeleton | - | Sliding windows | - | Fisher Vector | HMM, SVM |

Table 10: Multimodal gesture recognition results. SK: Skeleton; DNN: Deep Neural Network; RF: Ranfom Forest; 2DMTM: 2D motion trail model; RT: Regression Tree.

#### 3.3.4 2014 GESTURE CHALLENGE SUMMARY OF THE WINNER METHODS

Next, we describe the main characteristics of the three winning methods.

**First place:** The proposed method was based on a deep learning architecture that iteratively learned and integrated discriminative data representations from individual channels, modelling cross-modality correlations and short- and long-term temporal dependencies. This framework combined three data modalities: depth information, grayscale video and skeleton stream ("articulated pose"). Articulated pose served as an efficient representation of large-scale body motion of the upper body and arms, while depth and video streams contained complementary information about more subtle hand articulation. The articulated pose was formulated as a set of joint angles and normalized distances between upper-body joints, augmented with additional information reflecting speed and acceleration of each joint. For the depth and video streams, the authors did not rely on hand-crafted descriptors, but on discriminatively learning joint depth-intensity data representations with a set of convolutional neural layers. Iterative fusion of data channels was performed at output layers of the neural architecture. The idea of learning at multiple scales was also applied to the temporal dimension, such that a gesture was considered as an ordered set of characteristic motion impulses, or dynamic poses. Additional skeleton-based binary classifier was applied for accurate gesture localization. Fusing multiple modalities at several spatial and temporal scales led to a significant increase in recognition rates, allowing the model to compensate for errors of the individual classifiers as well as noise in the separate channels.

**Second place:** The approach combined a sliding-window gesture detector with multimodal features drawn from skeleton data, color imagery, and depth data produced by a first-generation Kinect$^{TM}$sensor. The gesture detector consisted of a set of boosted classifiers, each tuned to a specific gesture or gesture mode. Each classifier was trained independently on labeled training data, employing bootstrapping to collect hard examples. At run-time, the gesture classifiers were evaluated in a one-vs-all manner across a sliding window. Features were extracted at multiple temporal scales to enable recognition of variable-length gestures. Extracted features included descriptive statistics of normalized skeleton joint positions, rotations, and velocities, as well as HOG descriptors of the hands. The full set of gesture detectors was trained in under two hours on a single machine, and was extremely efficient at runtime, operating at 1700 fps using skeletal data.

**Third place:** The proposed method was based on four features: skeletal joint position feature, skeletal joint distance feature, and histogram of oriented gradients (HOG) features corresponding to left and right hands. Under the naïve Bayes assumption, likelihood functions were independently defined for every feature. Such likelihood functions were non-parametrically constructed from the training data by using kernel density estimation (KDE). For computational efficiency, k-nearest neighbor (kNN) approximation to the exact density estimator was proposed. Constructed likelihood functions were combined to the multimodal likelihood and this serves as a unary term for our pairwise Markov random field (MRF) model. For enhancing temporal coherence, a pairwise term was additionally incorporated to the MRF model. Final gesture labels were obtained via 1D MRF inference efficiently achieved by dynamic programming.

### 3.4 ChaLearn Action and Interaction Spotting Challenge 2014

The goal of this challenge was to perform automatic action and interaction spotting of people appearing in RGB data sequences.

| Training actions | Validation actions | Test actions | Sequence duration | FPS |
|---|---|---|---|---|
| 150 | 90 | 95 | 9× 1-2 min | 15 |
| Modalities | Num. of users | Action categories | interaction categories | Labelled sequences |
| RGB | 14 | 7 | 4 | 235 |

Table 11: Action and interaction data characteristics.

### 3.4.1 2014 ACTION CHALLENGE DATA

We presented a novel fully limb labelled dataset, the Human Pose Recovery and Behavior Analysis *HuPBA* $8k+$ dataset (Sánchez et al., 2014). This dataset is formed by more than 8000 frames where 14 limbs are labelled at pixel precision, thus providing $124,761$ annotated human limbs. The characteristics of the data set are:

• The images are obtained from 9 videos (RGB sequences) and a total of 14 different actors appear in the sequences. The image sequences have been recorded using a stationary camera with the same static background.

• Each video (RGB sequence) was recorded at 15 fps rate, and each RGB image was stored with resolution $480 \times 360$ in BMP file format.

• For each actor present in an image 14 limbs (if not occluded) were manually tagged: Head, Torso, R-L Upper-arm, R-L Lower-arm, R-L Hand, R-L Upper-leg, R-L Lower-leg, and R-L Foot.

• Limbs are manually labelled using binary masks and the minimum bounding box containing each subject is defined.

• The actors appear in a wide range of different poses and performing different actions/gestures which vary the visual appearance of human limbs. So there is a large variability of human poses, self-occlusions and many variations in clothing and skin color.

• Several actions and interactions categories are labelled at frame level.

A key frame example for each gesture/action category is shown in Figure 12. The challenges the participants had to deal with for this new competition are:

• 235 action/interaction samples performed by 14 actors.

• Large difference in length about the performed actions and interactions. Several distracter actions out of the 11 categories are also present.

• 11 action categories, containing isolated and collaborative actions: Wave, Point, Clap, Crouch, Jump, Walk, Run, Shake Hands, Hug, Kiss, Fight. There is a high intra-class variability among action samples.

Table 11 summarizes the data set attributes for the case of action/interaction spotting.

### 3.4.2 2014 ACTION CHALLENGE PROTOCOL AND EVALUATION

To evaluate the accuracy of action/interaction recognition, we use the Jaccard Index as defined in Section 3.3.2.

### 3.4.3 2014 ACTION CHALLENGE RESULTS

In this section we summarize the methods proposed by the participants and the winning methods. Six teams submitted their code and predictions for the test sets. Top rows of Table 9 summarizes the approaches of the participants who uploaded their models. One can see that most methods are based on similar approaches. In particular, alternative representations to classical BoW were considered, as Fisher Vector and VLAD (Jegou et al., 2012). Most methods perform sliding windows and SVM


(a) Wave (b) Point (c) Clap

(d) Crouch (e) Jump (f) Walk

(g) Run (h) Shake hands (i) Hug

(j) Kiss (k) Fight (l) Idle

Figure 12: Key frames of the *HuPBA* $8K+$ dataset used in the tracks 1 and 2, showing actions ((a) to (g)), interactions ((h) to (k)) and the idle pose (l).

classification. In addition, to refine the tracking of interest points, 4 participants used improved trajectories (Wang and Schmid, 2013).

### 3.4.4 2014 ACTION CHALLENGE SUMMARY OF THE WINNER METHODS

Next, we describe the main characteristics of the three winning methods.

**First place:** The method was composed of two parts: video representation and temporal segmentation. For the representation of video clip, the authors first extracted improved dense trajectories with HOG, HOF, MBHx, and MBHy descriptors. Then, for each kind of descriptor, the participants trained a GMM and used Fisher vector to transform these descriptors into a high dimensional super vector space. Finally, sum pooling was used to aggregate these codes in the whole video clip and normalize them with power L2 norm. For the temporal recognition, the authors resorted to a temporal sliding method along the time dimension. To speed up the processing of detection, the authors designed a temporal integration histogram of Fisher Vector, with which the pooled Fisher Vector was efficiently evaluated at any temporal window. For each sliding window, the authors used the pooled Fisher Vector as representation and fed it into the SVM classifier for action recognition.

**Second place:** a human action detection framework called "mixture of heterogeneous attribute analyzer" was proposed. This framework integrated heterogeneous attributes learned from various types of video features including static and dynamic, local and global features, to boost the action detection accuracy. The authors first detected a human from the input video by SVM-HOG detector and performed forward-backward tracking. Multiple local human tracks are linked into long trajectories by spatial-temporal graph based matching. Human key poses and local dense motion trajectories were then extracted within the tracked human bounding box sequences. Second, the authors proposed a mining method that learned discriminative attributes from three feature modalities: human trajectory, key pose and local motion trajectory features. The mining framework was based on the exemplar-SVM discriminative middle level feature detection approach. The learned discriminative attributes from the three types of visual features were then mixed in a max-margin learning algorithm which also explores the combined discriminative capability of heterogeneous feature modalities. The learned mixed analyzer was then applied to the input video sequence for action detection.

**Third place:** The framework for detecting actions in video is based on improved dense trajectories applied on a sliding windows fashion. Authors independently trained 11 one-versus-all kernel SVMs on the labelled training set for 11 different actions. The feature and feature descriptions used are improved dense trajectories, HOG, HOF, MBHx and MBHy. During training, for each action, a temporal sliding window is applied without overlapping. For every action, a segment was labelled 0 (negative) for a certain action only if there is no frame in this segment labelled 1. The feature coding method was bag-of-features. For a certain action, the features associated with those frames which are labelled 0 (negative) are not counted when we code the features of the action for the positive segments with bag-of-features. On the basis of the labelled segments and their features, a kernel SVM was trained for each action. During testing, non-overlap sliding window was applied for feature coding of the video. Every frame in a segment was consistently labelled as the output of SVM for each action. The kernel type, sliding window size and penalty of SVMs were selected during validation. When building the bag-of-features, the clustering method was $K$-means and the vocabulary size is 4000. For one trajectory feature in one frame, all the descriptors were connected to form one description vector. The bag-of-features were built upon this vector.

### 3.5 ChaLearn Action and Interaction Spotting Challenge 2015

The goal of this challenge was to perform automatic action and interaction spotting of people appearing in RGB data sequences. This corresponds to the second round of 2014 Action/Interaction challenge (Baró et al., 2015). Data, protocol, and evaluation were defined as explained in Section 3.4.

#### 3.5.1 2015 ACTION CHALLENGE RESULTS

Results of the two top ranked participants are shown in bottom rows of Table 9. One can see that the methods of the participants are similar to the ones applied in the 2014 challenge for the same dataset (Top rows of Table 9). Results of this second competition round improved by 2% the results obtained in the first round of the challenge.

#### 3.5.2 2015 ACTION CHALLENGE SUMMARY OF THE WINNER METHODS

**First winner:** This method is an improvement of the system proposed in (Peng et al., 2015), which is composed of two parts: video representation and temporal segmentation. For the representation of video clip, the authors first extracted improved dense trajectories with HOG, HOF, MBHx, and MBHy descriptors. Then, for each kind of descriptor, the participants trained a GMM and used Fisher vector to transform these descriptors into a high dimensional super vector space. Finally, sum pooling was used to aggregate these codes in the whole video clip and normalize them with power L2 norm. For the temporal recognition, the authors resorted to a temporal sliding method along the time dimension. To speed up the processing of detection, the authors designed a temporal integration histogram of Fisher Vector, with which the pooled Fisher Vector was efficiently evaluated at any temporal window. For each sliding window, the authors used the pooled Fisher Vector as representation and fed it into the SVM classifier for action recognition. A summary of this method is shown in Figure 13.

**Second winner:** The method implements an end-to-end generative approach from feature modelling to activity recognition. The system combines dense trajectories and Fisher Vectors with a temporally structured model for action recognition based on a simple grammar over action units. The authors modify the original dense trajectory implementation of (Wang and Schmid, 2013) to avoid the omission of neighborhood interest points once a trajectory is used (the improvement is shown in Figure 14). They use an open source speech recognition engine for the parsing and segmentation of video sequences. Because a large data corpus is typically needed for training such systems, images were mirrored to artificially generate more training data. The final result is achieved by voting over the output of various parameter and grammar configurations.

### 3.6 Other international competitions for gesture and action recognition

In addition to the series of ChaLearn Looking at People challenges, different international challenges have also been performed in the field of action/gesture recognition. Some of them are reviewed below.

The ChAirGest challenge (Ruffieux et al., 2013) is a research oriented competition designed to compare multimodal gesture recognizers. The provided data came from one Kinect camera and 4 Inertial Motion Units (IMU) attached to the right arm and neck of the subject. The dataset contains 10 different gestures, started from 3 different resting postures and recorded in two different lighting

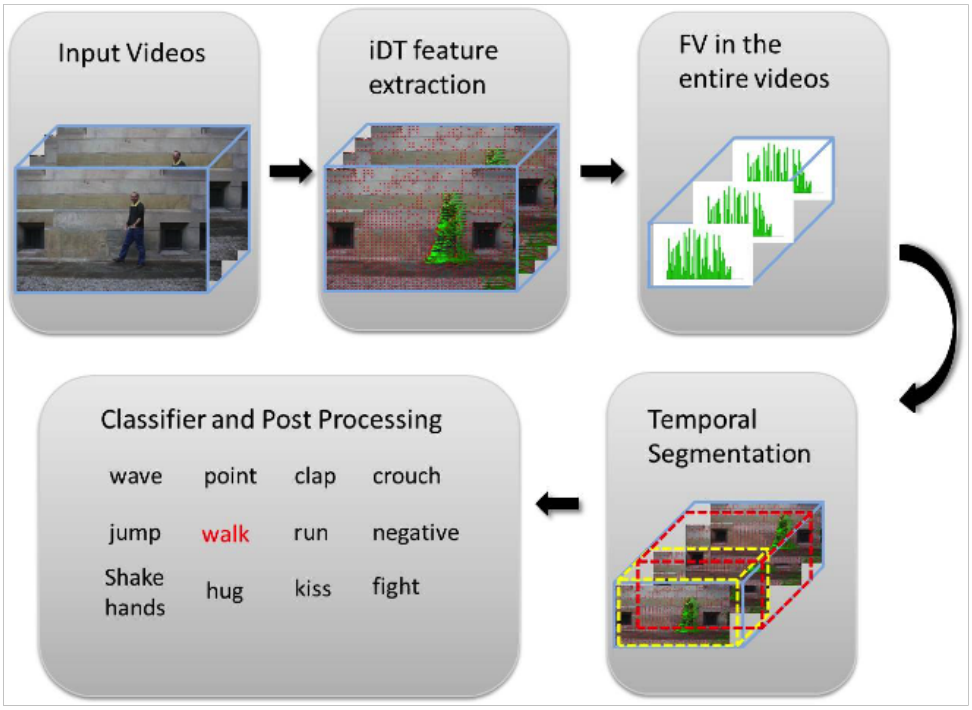
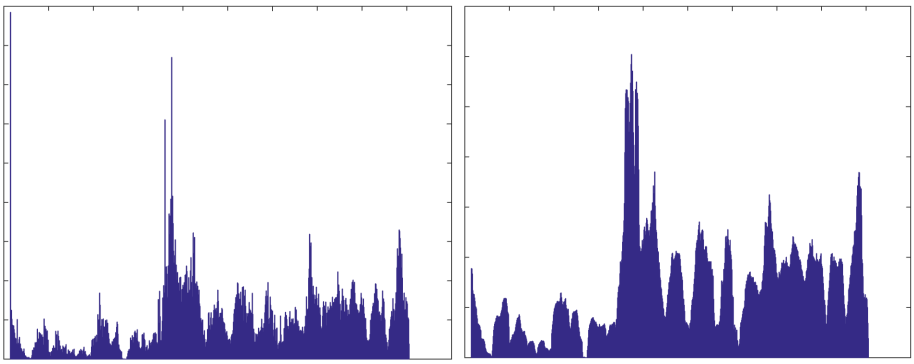



Figure 13: Method summary for MMLAB team (Wang et al., 2015b).



(a) (b)

Figure 14: Example of DT feature distribution for the first 200 frames of Seq01 for FKIE team, (a) shows the distribution of the original implementation, (b) shows the distribution of the modified version.

conditions by 10 different subjects. Thus, the total dataset contains 1200 annotated gestures split in continuous video sequences containing a variable number of gestures. The goal of the challenge was to promote research on methods using multimodal data to spot and recognize gestures in the context of close human-computer interaction.

In 2015, OpenCV ran at CVPR different challenges[5], one of them focusing on gesture recognition, using the ChaLearn gesture recognition data. The winner of the competition was also the work that won the ChaLearn challenge at ECCV 2014 (Neverova et al., 2014b).

Another recent action recognition competition is the THUMOS challenge (Gorban et al., 2015). Last round of the competition was ran at CVPR 2015. The last round of the challenge contained a forward-looking data set with over 430 hours of video data and 45 million frames (70% larger than THUMOS'14). All videos were collected from YouTube. Two tracks were performed: 1) Action Classification, for whole-clip action classification on 101 action classes, and 2) Temporal Action Localization, for action recognition and temporal localization on a subset of 20 action classes.

## 4. Summary of Special Topic Papers Not Related to the Challenges

In this special topic, in addition to the papers that described systems that participated in the ChaLearn gesture challenges, there are also several papers relating to broader aspects of gesture recognition, including topics such as sign language recognition, facial expression analysis, and facilitating development of real-world systems.

Three of these papers propose new methods for gesture and action recognition (Malgireddy et al., 2013; Fanello et al., 2013; Lui, 2012), that were also evaluated on parts of the CDG2011 dataset (Guyon et al., 2014), used in the ChaLearn challenges held in 2011 and 2012. More specifically, (Malgireddy et al., 2013) present an approach for detecting and recognizing activities and gestures. Hierarchical models are built to describe each activity as a combination of other, more simple activities. Each video is recursively divided into segments consisting of a fixed number of frames. The relationship between observed features and latent variables is modelled using a generative model that combines aspects of dynamic Bayesian networks and hierarchical Bayesian models. (Fanello et al., 2013) describe a system for real-time action recognition using depth video. The paper proposes specific features that are well adapted to real-time constraints, based on histograms of oriented gradients and histograms of optical flow. Support vector machines trained on top of such features perform action segmentation and recognition. (Lui, 2012) propose a method for representing gestures as tensors. These tensors are treated as points on a product manifold. In particular, the product manifold is factorized into three manifolds, one capturing horizontal motion, one capturing vertical motion, and one capturing 2D appearance. The difference between gestures is measured using a geodesic distance on the product manifold.

One paper (Wang et al., 2012) studied an action recognition problem outside the scope of the ChaLearn contests, namely recognizing poses and actions from a single image. Hierarchical models are used for modelling body pose. Each model in this hierarchy covers a part of the human body that can range from the entire body to a specific rigid part. Different levels in this hierarchy correspond to different degrees of coarseness vs. detail in the models at each level.

Three papers proposed novel methods within the area of sign language recognition (Cooper et al., 2012; Nayak et al., 2012; Roussos et al., 2013). (Cooper et al., 2012) describe a method for sign language recognition using linguistic subunits that are learned automatically by the system. Different types of such subunits are considered, including subunits based on appearance and local motion of the hand, subunits based on combining tracked 2D hand trajectories and hand shape, and subunits based on tracked 3D hand trajectories. (Nayak et al., 2012) address the problem of learning a model for a sign that occurs multiple times in a set of sentences. One benefit from such an approach

5. http://code.opencv.org/projects/opencv/wiki/VisionChallenge

is that it does not require the start and end frame of each sign as training data. Another benefit is that the method identifies the aspects of a sign that are least affected by movement epenthesis, i.e., by signs immediately preceding or following the sign in question. (Roussos et al., 2013) present a method for classifying handshapes for the purpose of sign language recognition. Cropped hand images are converted to a normalized representation called "shape-appearance images", based on a PCA decomposition of skin pixel colors. Then, active appearance models are used to model the variation in shape and appearance of the hand.

One paper focused on the topic of facial expression analysis (Martinez and Du, 2012). The paper proposes a model for describing how humans perceive facial expressions of emotion. The proposed model consists of multiple distinct continuous spaces. Emotions can be represented using linear combinations of these separate spaces. The paper also discusses how the proposed model can be used to design algorithms for facial expression recognition.

Another set of papers contributed methods that address different practical problems, that are important for building real-world gesture interfaces (Nguyen-Dinh et al., 2014; Gillian and Paradiso, 2014; Kohlsdorf and Starner, 2013). One such problem is obtaining manual annotations and ground truth for large amounts of training data. Obtaining such manual annotations can be time consuming, and can be an important bottleneck in building a real system. Crowdsourcing is a potential solution, but crowdsourced annotations often suffer from noise, in the form of discrepancies in how different humans annotate the same data. In (Nguyen-Dinh et al., 2014), two template-matching methods are proposed, called SegmentedLCSS and WarpingLCSS, that explicitly deal with the noise present in crowdsourced annotations of gestures. These methods are designed for spotting gestures using wearable motion sensors. The methods quantize signals into strings of characters and then apply variations of the longest common subsequence algorithm (LCSS) to spot gestures.

In designing a real-world system, another important problem is rapid development. (Gillian and Paradiso, 2014) present a gesture recognition toolkit, a cross-platform open-source C++ library. The toolkit features a broad range of classification and regression algorithms and has extensive support for building real-time systems. This includes algorithms for signal processing, feature extraction and automatic gesture spotting.

Finally, choosing the gesture vocabulary can be an important implementation parameter. (Kohlsdorf and Starner, 2013) propose a method for choosing a vocabulary of gestures for a human-computer interface, so that gestures in that vocabulary have a low probability of being confused with each other. Candidate gestures for the interface can be suggested both by humans and by the system itself. The system compares examples of each gesture with a large repository of unlabelled sensor/motion data, to check how often such examples resemble typical session/motion patterns encountered in that specific application domain.

## 5. Summary of Special Topic Papers Related to 2011-2012 Challenges

Next we briefly review the contributions of the accepted papers for the special topic whose methods are applied on the data provided by 2011-2012 ChaLearn gesture recognition challenges. Interestingly, none of the methods proposed in these papers uses skeletal tracking. Instead, these methods use different features based on appearance and/or motion. Where the papers differ is in their choice of specific features, and also in the choice of gesture models that are built on top of the selected features.

(Konecny and Hagara, 2014) combine appearance (Histograms of Oriented Gradients) and motion descriptors (Histogram of Optical Flow) from RGB and depth images for parallel temporal segmentation and recognition. The Quadratic-Chi distance family is used to measure differences between histograms to capture cross-bin relationships. Authors also propose trimming videos by removing unimportant frames based on motion information. Finally, proposed descriptors with different Dynamic Time Warping variants are applied for final recognition.

In contrast to (Konecny and Hagara, 2014), which employes commonly used features, (Wan et al., 2013) proposes a new multimodal descriptor, as well as a new sparse coding method. The multimodal descriptor is called 3D EMoSIFT, is invariant to scale and rotation, and has more compact and richer visual representations than other state-of-the-art descriptors. The proposed sparse coding method is named simulation orthogonal matching pursuit (SOMP), and is a variant of BoW. Using SOMP, each feature can be represented by some linear combination of a small number of codewords.

A different approach is taken in (Jiang et al., 2015), which combines three different methods for classifying gestures. The first method uses an improved principal motion representation. In the second method, a particle-based descriptor and a weighted dynamic time warping are proposed for the location component classification. In the third method, the shape of the human subject is used, extracted from the frame in the gesture that exhibits the least motion. The explicit use of shape in this paper is in contrast to (Konecny and Hagara, 2014; Wan et al., 2013), where shape information is implicitly coded in the extracted features.

In (Goussies et al., 2014), the focus is on transfer learning. The proposed method did not do as well as the previous methods (Konecny and Hagara, 2014; Wan et al., 2013; Jiang et al., 2015) on the CDG 2011 dataset, but nonetheless it contributes novel ideas for transfer learning, that can be useful when the number of training examples per class is limited. This is in contrast to the other papers related to the 2011-2012 challenges, that did not address transfer learning. The paper introduces two mechanisms into the decision forest framework, in order to transfer knowledge from the source tasks to a given target task. The first one is mixed information gain, which is a data-based regularizer. The second one is label propagation, which infers the manifold structure of the feature space. The proposed approach show improvements over traditional decision forests in the ChaLearn Gesture Challenge and on the MNIST dataset.

## 6. Summary of Special Issue Papers Related to 2013 Challenge

Next we briefly review the contributions of the accepted papers for the special issue whose methods are applied on the data provided by 2013 ChaLearn multimodal gesture recognition challenge. An important difference between this challenge and the previous ChaLearn challenges is the multimodal nature of the data. Thus, a key focus area for methods applied on this data is the problem of fusing information from multiple modalities.

(Wu and Cheng, 2014) propose a Bayesian Co-Boosting framework for multimodal gesture recognition. Inspired by boosting learning and co-training method, the system combines multiple collaboratively trained weak classifiers, Hidden Markov Models in this case, to construct the final strong classifier. During each iteration round, randomly a number of feature subsets are samples and weak classifiers parameters for each subset are estimated. The optimal weak classifier and its corresponding feature subset are retained for strong classifier construction. Authors also define an upper bound of training error and derive the update rule of instance's weight, which guarantees the

error upper bound to be minimized through iterations. This methodology won the ChaLearn 2013 ICMI competition.

(Pitsikalis et al., 2014) present a framework for multimodal gesture recognition that is based on a multiple hypotheses rescoring fusion scheme. Authors employ multiple modalities, i.e., visual cues, such as skeleton data, color and depth images, as well as audio, and extract feature descriptors of the hands movement, handshape, and audio spectral properties. Using a common hidden Markov model framework authors build single-stream gesture models based on which they can generate multiple single stream-based hypotheses for an unknown gesture sequence. By multimodally rescoring these hypotheses via constrained decoding and a weighted combination scheme, authors end up with a multimodally-selected best hypothesis. This is further refined by means of parallel fusion of the monomodal gesture models applied at a segmental level. The proposed methodology is tested on the ChaLearn 2013 ICMI competition data.

## 7. Discussion

We reviewed the gesture recognition topic, defining a taxonomy to characterize state of the art works on gesture recognition. We also reviewed the gesture and action recognition challenges organized by ChaLearn from 2011 to 2015, as well as other international competitions related to gesture recognition. Finally, we reviewed the papers submitted to the Special Topic on Gesture Recognition 2011-2014 we organized at Journal of Machine Learning Research.

Regarding the ChaLearn gesture recognition challenges, we began right at the start of the Kinect$^{TM}$revolution when inexpensive infrared cameras providing image depth recordings became available. We published papers using this technology and other more conventional methods, including regular video cameras, to record data, thus providing a good overview of uses of machine learning and computer vision using multimodal data in this area of application. Notably, we organized a series of challenges and made available several datasets we recorded for that purpose, including tens of thousands of videos, which are available to conduct further research [6].

Regarding the papers published in the gesture recognition special topic related to 2011-2012 challenges with the objective of performing one-shot learning, most of the authors proposed new multimodal descriptors taking benefit from both RGB and Depth cues in order to describe human body features, both static and dynamic ones. As the recognition strategies, common techniques used were variants of classical well-known Dynamic Time Warping and Hidden Markov Models. In particular, the most efficient techniques so far have used sequences of features processed by graphical models of the HMM/CRF family, similar to techniques used in speech recognition. Authors also considered a gesture candidate sliding window and motion-based video-cutting approaches. Last approach was frequently used since sign language videos included a resting pose. Also interesting novel classification strategies were proposed, such as multilayered decomposition, where different length gesture units are recognize at different levels ((Jiang et al., 2015)).

Regarding the papers published in the gesture recognition special topic related to 2013 and 2014 challenges with the objective of performing user independent multiple gesture recognition from large volumes of multimodal data (RGB, Depth and audio), different classifiers for gesture recognition were used by the participants. In 2013, the preferred one was Hidden Markov Models (used by the first ranked team of the challenge), followed by Random Forest (used by the second and third winners). Although several state of the art learning and testing gesture techniques were

6. http://gesture.chalearn.org/

applied at the last stage of the methods of the participants, still the feature vector descriptions are mainly based on MFCC audio features and skeleton joint information. The published paper of the winner to the special topic presents a novel coboosting strategy, where a set of HMM classifiers and collaboratively included in a boosting strategy considering random sets of features ((Wu and Cheng, 2014)). In 2014, similar descriptors and classifiers were used, and in particular, three deep learning architectures were considered, including the method of the winner team (Neverova et al., 2014b).

In the case of the ChaLearn action/interaction challenges organized in 2014 and 2015 most methods were based on similar approaches. In particular, alternative representations to classical BoW were considered, as Fisher Vector and VLAD (Jegou et al., 2012). Most methods performed sliding windows and SVM classification. In addition, to refine the tracking of interest points, several participants used improved trajectories (Wang and Schmid, 2013).

From the review of the gesture recognition topic, the achieved results in the performed challenges and the rest of papers published in the gesture recognition Special Topic, one can observe that still it is possible that progress will also be made in feature extraction by making better use of the multimodal development data for better transfer learning. For instance, we think that structural hand information around hand joint could be useful to discriminate among gesture categories that may share similar trajectories of hand/arms. Also recent approaches have shown that Random Forest and Deep Learning, such as considering Convolutional Neural Networks, are powerful alternatives to classical gesture recognition approaches, which still open the door for future the design of new gesture recognition classifiers.

In the case of action detection or spotting, most of the methods are still based on sliding-windows approaches, which makes recognition a time-consuming task. Thus, the research on methods that can generate gesture/action candidates from data in a different fashion are still an open issue.

## Acknowledgments

This work has been partially supported by ChaLearn Challenges in Machine Learning `http://chalearn.org`, the Human Pose Recovery and Behavior Analysis Group[7], the Pascal2 network of excellence, NSF grants 1128296, 1059235, 1055062, 1338118, 1035913, 0923494, and Spanish project TIN2013-43478-P. Our sponsors include Microsoft and Texas Instrument who donated prizes and provided technical support. The challenges were hosted by Kaggle.com and Coralab.org who are gratefully acknowledged. We thank our co-organizers of ChaLearn gesture and action recognition challenges: Miguel Reyes, Jordi Gonzalez, Xavier Baro, Jamie Shotton, Victor Ponce, Miguel Angel Bautista, and Hugo Jair Escalante.

## References

S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE TPAMI*, 32, 2010.

J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff. A unified framework for gesture recognition and spatiotemporal gesture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intel-*

7. HuPBA research group: http://www.maia.ub.es/∼sergio/

*ligence (PAMI)*, 31(9):1685–1699, 2009.

M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. Human pose estimation: New benchmark and state of the art analysis. In *CCVPR*. IEEE, 2014.

J. Appenrodt, A. Al-Hamadi, M. Elmezain, and B. Michaelis. Data gathering for gesture recognition systems based on mono color-, stereo color- and thermal cameras. In *Proceedings of the 1st International Conference on Future Generation Information Technology*, FGIT '09, pages 78–86, 2009. ISBN 978-3-642-10508-1.

V. Athitsos and S. Sclaroff. Estimating hand pose from a cluttered image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 432–439, 2003.

V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali. The American Sign Language lexicon video dataset. In *IEEE Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis (CVPR4HB)*, 2008.

A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. J. M. Havinga. Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In Michael Beigl and Francisco J. Cazorla-Almeida, editors, *ARCS Workshops*, pages 167–176, 2010. ISBN 978-3-8007-3222-7.

L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara. Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In *Proc. of 10th IEEE Embedded Vision Workshop (EVW)*, Columbus, Ohio, June 2014.

X. Baró, J. Gonzàlez, J. Fabian, M. A. Bautista, M. Oliu, H. J. Escalante, I. Guyon, and S. Escalera. ChaLearn looking at people 2015 challenges: action spotting and cultural event recognition. *ChaLearn Looking at People, Computer Vision and Pattern Recognition*, 2015.

B. Bauer, H. Hienz, and K.-F. Kraiss. Video-based continuous sign language recognition using statistical methods. In *International Conference on Pattern Recognition*, pages 2463–2466, 2000.

A. Y. Benbasat and J. A. Paradiso. Compact, configurable inertial gesture recognition. In *CHI '01: CHI '01 extended abstracts on Human factors in computing systems*, pages 183–184. ACM Press, 2001. ISBN 1581133405.

S. Berlemont, G. Lefebvre, S. Duffner, and C. Garcia. Siamese neural network based similarity metric for inertial gesture classification and rejection. *Automatic Face and Gesture Recognition*, 2015.

V. Bloom, D. Makris, and V. Argyriou. G3D: A gaming action dataset and real time action recognition evaluation framework. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 7–12, 2012.

A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(3):257–267, 2001.

L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, pages 1365–1372. IEEE, 2009.

M. Brand, N. Oliver, and A.P. Pentland. Coupled Hidden Markov Models for complex action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 994–999, 1997.

M. Caon, Y. Yong, J. Tscherrig, E. Mugellini, and O. Abou Khaled. Context-aware 3D gesture interaction based on multiple Kinects. In *The First International Conference on Ambient Computing, Applications, Services and Technologies*, page 712, 2011. ISBN 978-1-61208-170-0.

G. Castellano, S. D. Villalba, and A. Camurri. Recognising human emotions from body movement and gesture dynamics. In *Affective computing and intelligent interaction*, pages 71–82. Springer, 2007.

A. Chaudhary, J. L. Raheja, K. Das, and S. Raheja. A survey on hand gesture recognition in context of soft computing. 133:46–55, 2011.

F.S. Chen, C.M. Fu, and C.L. Huang. Hand gesture recognition using a real-time tracking method and Hidden Markov Models. *Image and Video Computing*, 21(8):745–758, August 2003.

M. Chen, G. AlRegib, and B.-H. Juang. 6DMG: A new 6D motion gesture database. In *Multimedia Systems Conference*, pages 83–88, 2012.

C. Conly, P. Doliotis, P. Jangyodsuk, R. Alonzo, and V. Athitsos. Toward a 3D body part detection video dataset and hand tracking benchmark. In *Pervasive Technologies Related to Assistive Environments (PETRA)*, 2013.

C. Conly, Z. Zhang, and V. Athitsos. An integrated RGB-D system for looking up the meaning of signs. In *Pervasive Technologies Related to Assistive Environments (PETRA)*, 2015.

H. Cooper and R. Bowden. Learning signs from subtitles: A weakly supervised approach to sign language recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2568–2574, 2009.

H. Cooper, E.-J. Ong, N. Pugeault, and R. Bowden. Sign language recognition using sub-units. *Journal of Machine Learning Research (JMLR)*, 13(7):2205–2231, 2012.

A. Corradini. Dynamic time warping for off-line recognition of a small gesture vocabulary. In *Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems (RATFG-RTS)*, pages 82–89, 2001.

Y. Cui and J. Weng. Appearance-based hand sign recognition from intensity image sequences. *Computer Vision and Image Understanding*, 78(2):157–176, 2000.

R. Cutler and M. Turk. View-based interpretation of real-time optical flow for gesture recognition. In *Automatic Face and Gesture Recognition*, pages 416–421, 1998.

A. Czabke, J. Neuhauser, and T. C. Lueth. Recognition of interactions with objects based on radio modules. In *International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, 2010.

T.J. Darrell, I.A. Essa, and A.P. Pentland. Task-specific gesture analysis in real-time using interpolated views. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 18(12): 1236–1242, 1996.

M. de La Gorce, D. J. Fleet, and N. Paragios. Model-based 3D hand pose estimation from monocular video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(9):1793–1805, 2011.

D. Demirdjian and C. Varri. Recognizing events with temporal random forests. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, pages 293–296, 2009.

KG Derpanis, M Sizintsev, KJ Cannons, and RP Wildes. Action spotting and recognition based on a spatiotemporal orientation analysis. *IEEE TPAMI*, 35(3):527 – 540, 2013.

P. Dreuw, T. Deselaers, D. Keysers, and H. Ney. Modeling image variability in appearance-based gesture recognition. In *ECCV Workshop on Statistical Methods in Multi-Image and Video Processing*, pages 7–18, 2006.

S. Duffner, S. Berlemont, G. Lefebvre, and C. Garcia. 3D gesture classification with convolutional neural networks. In *The 39th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

U. M. Erdem and S. Sclaroff. Automatic detection of relevant head gestures in american sign language communication. In *International Conference on Pattern Recognition*, pages 460–463, 2002.

S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, I. Guyon, V. Athitsos, H. J. Escalante, L. Sigal, A. Argyros, C. Sminchisescu, R. Bowden, and S. Sclaroff. Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. *15th ACM International Conference on Multimodal Interaction*, pages 365–368, 2013a.

S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopés, I. Guyon, V. Athitsos, and H. J. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *ChaLearn Multi-Modal Gesture Recognition Grand Challenge and Workshop, 15th ACM International Conference on Multimodal Interaction*, 2013b.

S. Escalera, X. Baro, J. Gonzalez, M. Bautista, M. Madadi, M. Reyes, V. Ponce, H. J. Escalante, J. Shotton, and I. Guyon. ChaLearn looking at people challenge 2014: Dataset and results. *ChaLearn Looking at People, European Conference on Computer Vision*, 2014.

M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

S. R. Fanello, I. Gori, G. Metta, and F. Odone. Keep it simple and sparse: Real-time action recognition. *Journal of Machine Learning Research (JMLR)*, 14(9):2617–2640, 2013.

A. Farhadi, D. A. Forsyth, and R. White. Transfer learning in sign language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

A. Fathi, X. Ren, and J.M. Rehg. Learning to recognize objects in egocentric activities. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3281–3288.

S. S. Fels. *Glove-TalkII: Mapping hand gestures to speech using neural networks*. PhD thesis, University of Toronto, 1994.

V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 1737–1746, 2012.

W. T. Freeman and M. Roth. Computer vision for computer games. In *Automatic Face and Gesture Recognition*, pages 100–105, 1996.

W. T. Freeman and C. D. Weissman. Television control by hand gestures, 1994.

N. Gillian and J.A. Paradiso. The gesture recognition toolkit. *Journal of Machine Learning Research (JMLR)*, 14, 2014.

A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://www.thumos.info/`, 2015.

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.

N. Goussies, S. Ubalde, and M. Mejail. Transfer learning decision forests for gesture recognition. *Journal of Machine Learning Research (JMLR)*, 2014.

M. Gowing, A. Ahmadi, F. Destelle, D. S. Monaghan, N. E. O'Connor, and K. Moran. Kinect vs. low-cost inertial sensing for gesture recognition. *Lecture Notes in Computer Science, Springer*, 8325:484–495, 2014.

I. Guyon, V. Athitsos, P. Jangyodsuk, H.J. Escalante, and B. Hamner. Results and analysis of the ChaLearn gesture challenge 2012. In Xiaoyi Jiang, Olga Regina Pereira Bellon, Dmitry Goldgof, and Takeshi Oishi, editors, *Advances in Depth Image Analysis and Applications*, volume 7854 of *Lecture Notes in Computer Science*, pages 186–204. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-40302-6. doi: 10.1007/978-3-642-40303-3_19. URL `http://dx.doi.org/10.1007/978-3-642-40303-3_19`.

I. Guyon, V. Athitsos, P. Jangyodsuk, and H. J. Escalante. The ChaLearn gesture dataset (CGD 2011). *Machine Vision and Applications*, 25:1929–1951, 2014.

A. Hernandez-Vela, N. Zlateva, A. Marinov, M. Reyes, P. Radeva, D. Dimov, and S. Escalera. Graph cuts optimization for multi-limb human segmentation in depth maps. *IEEE Computer Vision and Pattern Recognition conference*, 2012.

A. Hernandez-Vela, M. A. Bautista, X. Perez-Sala, V. Ponce, S. Escalera, X. Baro, O. Pujol, and C. Angulo. Probability-based dynamic time warping and bag-of-visual-and-depth-words for human gesture recognition in RGB-D. *Pattern Recognition Letters*, http://dx.doi.org/10.1016/j.patrec.2013.09.009, 2013a.

A. Hernandez-Vela, M. Reyes, V. Ponce, and S. Escalera. Grabcut-based human segmentation in video sequences. *Sensors*, 12(1):15376–15393, 2013b.

G. Hewes. Primate communication and the gestural origins of language. *Current Antropology*, 14: 5–24, 1973.

N. A. Ibraheem and R. Z. Khan. Survey on various gesture recognition technologies and techniques. *International Journal of Computer Applications*, 50(7):38–44, 2012.

C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

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

H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE TPAMI*, 34(9):1704–1716, 2012.

F. Jiang, S. Zhang, S. Wu, Y. Gao, and D. Zhao. Multi-layered gesture recognition with Kinect. *Journal of Machine Learning Research (JMLR)*, 2015.

S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. doi:10.5244/C.24.12.

A. Joshi, S. Sclaroff, M. Betke, and C. Monnier. A random forest approach to segmenting and classifying gestures. *Automatic Face and Gesture Recognition*, 2015.

T. Kadir, R. Bowden, E. Ong, and A. Zisserman. Minimal training, large lexicon, unconstrained sign language recognition. In *British Machine Vision Conference (BMVC)*, volume 2, pages 939–948, 2004.

K. Kahol, P. Tripathi, and S. Panchanathan. Automated gesture segmentation from dance sequences. In *Automatic Face and Gesture Recognition*, pages 883–888, 2004.

H. Kang, C.W. Lee, and K. Jung. Recognition-based gesture spotting in video games. *Pattern Recognition Letters*, 25(15):1701–1714, November 2004.

A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. F. Driessen. Gesture-based affective computing on motion capture data. In Jianhua Tao, Tieniu Tan, and Rosalind W. Picard, editors, *Affective Computing and Intelligent Interaction*, Lecture Notes in Computer Science, pages 1–7. Springer Berlin Heidelberg, 2005.

S. Kausar and M.Y. Javed. A survey on sign language recognition. *Frontiers of Information Technology*, pages 95–98, 2011.

Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 166–173, 2005.

D. Kelly, J. McDonald, and C. Markham. A person independent system for recognition of hand postures used in sign language. *Pattern Recognition Letters*, 31(11):1359–1368, 2010.

C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *European Conference on Computer Vision (ECCV)*, pages 852–863, 2012.

R. Z. Khan and N. A. Ibraheem. Survey on gesture recognition for hand image postures. *Computer and Information Science*.

T.-K. Kim, S.-F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

H. Kjellström, J. Romero, D. Martínez, and D. Kragić. Simultaneous visual recognition of manipulation actions and manipulated objects. In *European Conference on Computer Vision*, volume 2, pages 336–349, 2008.

D. K. H. Kohlsdorf and T. E. Starner. MAGIC summoning: Towards automatic suggesting and testing of gestures with low probability of false positives during use. *Journal of Machine Learning Research (JMLR)*, 14(1):209–242, 2013.

M. Kolsch and M. Turk. Fast 2D hand tracking with flocks of features and multi-cue integration. In *IEEE Workshop on Real-Time Vision for Human-Computer Interaction*, pages 158–165, 2004.

J. Konecny and M. Hagara. One-shot-learning gesture recognition using hog-hof features. *Journal of Machine Learning Research*, 15:2513–2532, 2014. URL `http://jmlr.org/papers/v15/konecny14a.html`.

Y. Kong, B. Satarboroujeni, and Y. Fu. Hierarchical 3D kernel descriptors for action recognition using depth sequences. *Automatic Face and Gesture Recognition*, 2015.

J. B. Kruskal and M. Liberman. The symmetric time warping algorithm: From continuous to discrete. In *Time Warps*. Addison-Wesley, 1983.

A. Kurakin, Z. Zhang, and Z. Liu. A real time system for dynamic hand gesture recognition with a depth sensor. In *European Signal Processing Conference, EUSIPCO*, pages 1975–1979, 2012.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*, pages 282–289, 2001.

H. Lane, R. J. Hoffmeister, and B. Bahan. *A Journey into the Deaf-World*. DawnSign Press, San Diego, CA, 1996.

I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, September 2005. ISSN 0920-5691. doi: 10.1007/s11263-005-1838-7. URL `http://dx.doi.org/10.1007/s11263-005-1838-7`.

I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, pages 1–8, 2008.

E. Larson, G. Cohn, S. Gupta, X. Ren, B. Harrison, D. Fox, and S. Patel. HeatWave: Thermal imaging for surface user interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2565–2574, 2011.

J. J. LaViola Jr. A survey of hand posture and gesture recognition techniques and technology. Technical report, Providence, RI, USA, 1999.

H.K. Lee and J.H. Kim. An HMM-based threshold model approach for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):961–973, October 1999.

S.-W. Lee. Automatic gesture recognition for intelligent human-robot interaction. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 645–650, 2006.

C. Li and K. M. Kitani. Pixel-level hand detection for ego-centric videos. *CVPR*, 2013.

W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. *CVPR workshops*, pages 9 – 14, 2010.

H. Liang, J. Yuan, D. Thalmann, and Z. Zhang. Model-based hand pose estimation via spatial-temporal hand parsing and 3D fingertip localization. *The Visual Computer*, 29(6-8):837–848, 2013.

H. Liang, J. Yuan, and D. Thalmann. Parsing the hand in depth images. *IEEE Transactions on Multimedia*, 16(5):1241–1253, 2014.

A. Licsár and T. Szirányi. Hand gesture recognition in camera-projector system*. In *Computer Vision in Human-Computer Interaction*, pages 83–93. Springer, 2004.

Z. Lin, Z. Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In *IEEE International Conference on Computer Vision, ICCV*, pages 444–451, 2009.

K. Liu, C. Chen, R. Jafari, and N. Kehtarnavaz. Fusion of inertial and depth sensor data for robust hand gesture recognition. *IEEE Sensors Journal*, 14(6):1898–1903, 2014.

L. Liu and L. Shao. Learning discriminative representations from RGB-D video data. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1493–1500, 2013.

O. Lopes, M. Reyes, S. Escalera, and J. Gonzàlez. Spherical blurred shape model for 3-D object and pose recognition: Quantitative analysis and HCI applications in smart environments. *IEEE T. Cybernetics*, 44(12):2379–2390, 2014.

Y. M. Lui. Human gesture recognition on product manifolds. *Journal of Machine Learning Research (JMLR)*, 13(11):3297–3321, 2012.

J. Luo, W. Wang, and H. Qi. Spatio-temporal feature extraction and representation for RGB-D human action recognition. *PRL*, 2014.

J. Ma, W. Gao, J. Wu, and C. Wang. A continuous Chinese Sign Language recognition system. In *Automatic Face and Gesture Recognition*, pages 428–433, 2000.

S. Ma, J. Zhang, N. Ikizler-Cinbis, and S. Sclaroff. Action recognition and localization by hierarchical space-time segments. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

M. R. Malgireddy, I. Nwogu, and V. Govindaraju. Language-motivated approaches to action recognition. *Journal of Machine Learning Research*, 14:2189–2212, 2013. URL `http://jmlr.org/papers/v14/malgireddy13a.html`.

S. Malik and J. Laszlo. Visual touchpad: A two-handed gestural input device. In *International Conference on Multimodal Interfaces*, pages 289–296, 2004.

J. Martin, V. Devin, and J. L. Crowley. Active hand tracking. In *Automatic Face and Gesture Recognition*, pages 573–578, 1998.

A. Martinez and S. Du. A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. *Journal of Machine Learning Research (JMLR)*, 13(5): 1589–1608, 2012.

D. McNeil. How language began, gesture and speech in human evolution. *Cambridge editorial*, 2012.

S. Mitra and T. Acharya. Gesture recognition: A survey. *Trans. Sys. Man Cyber Part C*, 37(3): 311–324, May 2007. ISSN 1094-6977.

Z. Mo and U. Neumann. Real-time hand pose recognition using low-resolution depth images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1499–1505, 2006.

B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. Technical Report 326, MIT, June 1995.

P. Molchanov, S. Gupta, K. Kim, and K. Pulli. Multi-sensor system for drivers hand-gesture recognition. *Automatic Face and Gesture Recognition*, 2015.

J. Nagi, F. Ducatelle, G. A. Di Caro, D. C. Ciresan, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L. M. Gambardella. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *ICSIPA*, pages 342–347. IEEE, 2011. ISBN 978-1-4577-0243-3.

S. Nayak, S. Sarkar, and B. Loeding. Unsupervised modeling of signs embedded in continuous sentences. In *IEEE Workshop on Vision for Human-Computer Interaction*, 2005.

S. Nayak, K. Duncan, S. Sarkar, and B. Loeding. Finding recurrent patterns from continuous sign language sentences for automated extraction of signs. *Journal of Machine Learning Research (JMLR)*, 13(9):2589–2615, 2012.

C. Neidle, A. Thangali, and S. Sclaroff. Challenges in development of the American Sign Language lexicon video dataset (ASLLVD) corpus. In *Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, 2012.

N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. Hand segmentation with structured convolutional learning. *ACCV*, 2014a.

N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. Multi-scale deep learning for gesture detection and localization. *ChaLearn Looking at People, European Conference on Computer Vision*, 2014b.

L. Nguyen-Dinh, A. Calatroni, and G. Troster. Robust online gesture recognition with crowdsourced annotations. *Journal of Machine Learning Research (JMLR)*, 15, 2014.

E. Ohn-Bar and M. M. Trivedi. Hand gesture recognition in real-time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Transactions on Intelligent Transportation Systems*, 2014.

I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Markerless and efficient 26-DOF hand pose recovery. In *Asian Conference on Computer Vision (ACCV)*, 2010.

I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2088–2095, 2011.

K. Oka, Y. Sato, and H. Koike. Real-time fingertip tracking and gesture recognition. *IEEE Computer Graphics and Applications*, 22(6):64–71, 2002.

R. Oka. Spotting method for classification of real world data. *The Computer Journal*, 41(8):559–565, July 1998.

E. J. Ong and R. Bowden. A boosted classifier tree for hand shape detection. In *Face and Gesture Recognition*, pages 889–894, 2004.

O. Oreifej and Z. Liu. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. *CVPR*, pages 716 – 723, 2013.

A. Pardo, A. Clapes, S. Escalera, and O. Pujol. Actions in context: System for people with dementia. *2nd International Workshop on Citizen Sensor Networks (Citisen2013) at the European Conference on Complex Systems (ECCS'13)*, 2013.

A. Patron-Perez, M. Marszalek, A. Zisserman, and I. D. Reid. High five: Recognising human interactions in TV shows. In *British Machine Vision Conference BMVC*, 2010.

X. Peng, L. Wang, Z. Cai, and Y. Qiao. Action and gesture temporal spotting with super vector representation. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, editors, *Computer Vision - ECCV 2014 Workshops*, volume 8925 of *Lecture Notes in Computer Science*, pages 518–527. Springer International Publishing, 2015. ISBN 978-3-319-16177-8. doi: 10.1007/978-3-319-16178-5_36. URL http://dx.doi.org/10.1007/978-3-319-16178-5_36.

A. Pieropan, G. Salvi, K.Pauwels, and H. Kjellstrm. Audio-visual classification and detection of human manipulation actions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014.

V. Pitsikalis, A. Katsamanis, S. Theodorakis, and P. Maragos. Multimodal gesture recognition via multiple hypotheses rescoring. *Journal of Machine Learning Research (JMLR)*, 2014.

N. Pugeault and R. Bowden. Spelling it out: Real-time ASL fingerspelling recognition. In *ICCV Workshops*, pages 1114–1119, 2011.

A. Quattoni, S. B. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(10):1848–1852, 2007.

D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, pages 1129–1136, 2006.

I. Rauschert, P. Agrawal, R. Sharma, S. Fuhrmann, I. Brewer, and A. MacEachren. Designing a human-centered, multimodal GIS interface to support emergency management. In *ACM International Symposium on Advances in Geographic Information Systems*, pages 119–124, 2002.

J.M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *IEEE International Conference onComputer Vision (ICCV)*, volume 0, pages 612–617, 1995.

Z. Ren, J. Meng, J. Yuan, and Z. Zhang. Robust hand gesture recognition with Kinect sensor. In *ACM International Conference on Multimedia*, pages 759–760, 2011a.

Z. Ren, J. Yuan, and Z. Zhang. Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In *ACM International Conference on Multimedia*, pages 1093–1096, 2011b.

Z. Ren, J. Yuan, J. Meng, and Z. Zhang. Robust part-based hand gesture recognition using Kinect sensor. *IEEE Transactions on Multimedia*, 15(5):1110–1120, 2013.

A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos. Dynamic affine-invariant shape-appearance handshape features and classification in sign language videos. *Journal of Machine Learning Research (JMLR)*, 14(6):1627–1663, 2013.

S. Ruffieux, D. Lalanne, and E. Mugellini. ChAirGest: A challenge for multimodal mid-air gesture recognition for close HCI. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pages 483–488, 2013.

A. Sadeghipour, L.-P. Morency, and S. Kopp. Gesture-based object recognition using histograms of guiding strokes. In *British Machine Vision Conference*, pages 44.1–44.11, 2012.

D. Sánchez, M. A. Bautista, and S. Escalera. HuPBA 8k+: Dataset and ECOC-graphcut based segmentation of human limbs. *Neurocomputing*, 2014.

B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *CVPR*. IEEE, 2013.

Y. Sato and T. Kobayashi. Extension of Hidden Markov Models to deal with multiple candidates of observations and its application to mobile-robot-oriented gesture recognition. In *International Conference on Pattern Recognition (ICPR)*, pages II: 515–519, 2002.

J.D Schein. *At home among strangers*. Gallaudet U. Press, Washington, DC, 1989.

C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, volume 3, pages 32–36, 2004.

N. Shapovalova, W. Gong., M. Pedersoli, F. X. Roca, and J. Gonzalez. On importance of interactions and context in human action recognition. In *Pattern Recognition and Image Analysis*, pages 58–66, 2011.

J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1297–1304, 2011.

L. Sigal, A. O. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, 87(1-2):4–27, 2010.

C. Sminchisescu, A. Kanaujia, and D. Metaxas. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding*, 104:210–220, 2006.

Y. Song, D. Demirdjian, and R. Davis. Multi-signal gesture recognition using temporal smoothing hidden conditional random fields. In *Automatic Face and Gesture Recognition*, pages 388–393, 2011a.

Y. Song, D. Demirdjian, and R. Davis. Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database. In *Automatic Face and Gesture Recognition*, pages 500–506, 2011b.

T. Starner and A. Pentland. Real-time American Sign Language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (12):1371–1375, 1998.

N. Stefanov, A. Galata, and R. Hubbold. Real-time hand tracking with variable-length Markov Models of behaviour. In *Real Time Vision for Human-Computer Interaction*, 2005.

B. Stenger, A. Thayananthan, P.H.S. Torr, and R. Cipolla. Filtering using a tree-based estimator. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1063–1070, 2003.

E. Sudderth, M. Mandel, W. Freeman, and A. Willsky. Distributed occlusion reasoning for tracking with nonparametric belief propagation. In *Neural Information Processing Systems (NIPS)*, 2004.

D. Tran and D. Forsyth. Improved human parsing with a full relational model. In *ECCV*, pages 227–240. IEEE, 2010.

J. Triesch and C. von der Malsburg. A system for person-independent hand posture recognition against complex backgrounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1449–1453, 2001.

J. Triesch and C. von der Malsburg. Classification of hand postures against complex backgrounds using elastic graph matching. *Image and Vision Computing*, 20(13-14):937–943, 2002.

M. Van den Bergh, E. Koller-Meier, and L. Van Gool. Real-time body pose recognition using 2D or 3D haarlets. *International Journal of Computer Vision (IJCV)*, 83(1):72–84, 2009.

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 511–518, 2001.

C. Vogler and D Metaxas. Parallel Hidden Markov Models for American Sign Language recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 116–122, 1999.

J. Wan, Q. Ruan, W. Li, and S. Deng. One-shot learning gesture recognition from RGB-D data using bag of features. *Journal of Machine Learning Research*, 14:2549–2582, 2013. URL `http://jmlr.org/papers/v14/wan13a.html`.

H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*, 2013.

H. Wang, A. Stefan, S. Moradi, V. Athitsos, C. Neidle, and F. Kamangar. A system for large vocabulary sign search. In *Workshop on Sign, Gesture and Activity (SGA)*, 2010.

H. Wang, X. Chai, Y. Zhou, and X. Chen. Fast sign language recognition benefited from low rank approximation. In *Automatic Face and Gesture Recognition*, 2015a.

J. Wang, Z. Liu, Y. Wu, and J. Yuan. Learning actionlet ensemble for 3D human action recognition. *IEEE TPAMI*, 36(5):914 – 927, 2014.

R. Y. Wang and J. Popović. Real-time hand-tracking with a color glove. *ACM Transactions on Graphics*, 28(3):63:1–63:8, July 2009.

Y. Wang, D. Tran, Z. Liao, and D. Forsyth. Discriminative hierarchical part-based models for human parsing and action recognition. *Journal of Machine Learning Research (JMLR)*, 13(10): 3075–3102, 2012.

Z. Wang, L. Wang, W. Du, and Q. Yu. Action spotting system using Fisher vector. In *In CVPR ChaLearn Looking at People Workshop 2015*, 2015b.

A. Wexelblat. An approach to natural gesture in virtual environments. *ACM Transactions on Computer-Human Interactions*, 2(3):179–200, 1995.

M. Wilhelm. A generic context aware gesture recognition framework for smart environments. *PerCom Workshops*.

A. D. Wilson and A. F. Bobick. Parametric Hidden Markov Models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 21(9), 1999.

J. Wu and J. Cheng. Bayesian co-boosting for multi-modal gesture recognition. *Journal of Machine Learning Research (JMLR)*, 2014.

Y. Wu and T. S. Huang. View-independent recognition of hand postures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 88–94, 2000.

Y. Xiao, Z. Zhang, A. Beck, J. Yuan, and D. Thalmann. Human-robot interaction by understanding upper body gestures. *Presence*, 23(2):133 – 154, 2014.

H. D. Yang, S. Sclaroff, and S. W. Lee. Sign language spotting with a threshold model based on conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(7):1264–1277, July 2009.

M. H. Yang, N. Ahuja, and M. Tabb. Extraction of 2D motion trajectories and its application to hand gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24(8):1061–1074, 2002.

W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2030–2037, 2010.

X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. *CVPR*, 2014a.

X. Yang and Y. Tian. Action recognition using super sparse coding vector with spatio-temporal awareness. *ECCV*, 2014b.

G. Yao, H. Yao, X. Liu, and F. Jiang. Real time large vocabulary continuous sign language recognition based on OP/Viterbi algorithm. In *International Conference on Pattern Recognition*, volume 3, pages 312–315, 2006.

G. Yu, Z. Liu, and J. Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. *ACCV*, 2014.

J. Yuan, Z. Liu, and Y. Wu. Discriminative video pattern search for efficient action detection. *IEEE TPAMI*, 33(9):1728–1743, 2011.

Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti. American Sign Language recognition with the Kinect. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, pages 279–286, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0641-6. doi: 10.1145/2070481.2070532. URL http://doi.acm.org/10.1145/2070481.2070532.

M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. *ICCV*, 2013.

J. Zieren and K.-F. Kraiss. Robust person-independent visual sign language recognition. In *Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, volume 1, pages 520–528, 2005.